

UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

The Bystander Public for Online Incivility: A Development of the Bystander Counter-Intervention Model

Permalink

<https://escholarship.org/uc/item/6bk4v856>

Author

Kim, Yeweon

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Santa Barbara

The Bystander Public for Online Incivility:

A Development of the Bystander Counter-Intervention Model

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Communication

by

Yeweon Kim

Committee in charge:

Professor Amy Gonzales, Chair

Professor Joe Walther

Professor Miriam Metzger

December 2023

The dissertation of Yeweon Kim is approved.

Joe Walther

Miriam Metzger

Amy Gonzales, Committee Chair

July 2023

The Bystander Public for Online Incivility:
A Development of the Bystander Counter-Intervention Model

Copyright © 2023

by

Yeweon Kim

ACKNOWLEDGEMENTS

Thank God for his gift that words can't describe.

VITA OF YEWEON KIM
December 2023

EDUCATION

Bachelor of Arts in Visual & Performing Arts (major) and Mass Communication (minor),
Handong Global University, February 2009 (summa cum laude)
Master of Arts in Communication, Seoul National University, February 2014
Doctor of Philosophy in Media Arts and Sciences (major) and Information Science (minor),
Indiana University Bloomington, May 2018 (Transfer)
Doctor of Philosophy in Communication, University of California Santa Barbara, December
2023

PROFESSIONAL EMPLOYMENT

2023-2024: Research Assistant, University of California Santa Barbara Blum Center
2018-2023: Teaching Assistant, Department of Communication, University of California
Santa Barbara
2015-2018: Teaching Assistant, The Media School, Indiana University Bloomington
2014: Research Assistant, ICT Social Policy Center, Seoul National University
2011-2014: Teaching Assistant, Department of Communication, Seoul National University
2009-2010: Copywriter, 9 Fruits Media Inc.

PUBLICATIONS

- Kim, Y. W. (2022). Potentials and limitations of computer-mediated communication theories for online incivility research: A focus on bystander dynamics. *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 6280-6289). <http://hdl.handle.net/10125/80101>
- Gonzales, A. L., Kim, Y. W., & Wang, L. (2022). Multisolving innovations: How digital equity, eWaste, and right-to-repair policies can increase the supply of affordable computers. *Policy & Internet*. <https://doi.org/10.1002/poi3.331>
- Kim, Y. W., & Gonzales, A. L. (2022). When we tolerate online incivility: Dual-process effects of argument strength and heuristic cues in uncivil user comments. *Computers in Human Behavior*, 131, 107235. <https://doi.org/10.1016/j.chb.2022.107235>
- Kim, Y. W. (2021). Understanding the bystander audience in online incivility encounters: Conceptual issues and future research questions. *Proceedings of the 54th Hawaii International Conference on System Sciences* (pp. 2934-2943). <https://doi.org/10.24251/HICSS.2021.357>
- Carr, C. T., Kim, Y. W., Valov, J. J., Rosenbaum, J. E., Johnson, B. K., Hancock, J. T., & Gonzales, A. L. (2021). Getting our shift together: An explication of identity shift theory. *Journal of Media Psychology*, 33(4), 202-214. <https://doi.org/10.1027/1864-1105/a000314>
- Read, G. L., Yan, H. Y., Anderson, P. B., Partain, L. P. B., Vaughn, Z., Semivolos, A., Kim, Y. W., & Gonzales, A. L. (2021). Making stability dependable: Stable cellphone access leads to better health outcomes for those experiencing poverty. *Information,*

- Communication & Society*, 25(14), 2122-2139.
<https://doi.org/10.1080/1369118X.2021.1928263>
- Kim, Y. W., & Gonzales, A. (2018). You don't know me: Negative self-views interact with publicness and feedback to shape interpersonal impressions online. *Computers in Human Behavior*, 88, 195-204. <https://doi.org/10.1016/j.chb.2018.06.025>
- Kim, Y. W. & Herring, S. C. (2018). Is politeness catalytic and contagious? Effects on participation in online news discussions. *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 1955-1964).
<https://doi.org/10.24251/HICSS.2018.247>
- Lee, E.-J., & Kim, Y. W. (2016). Effects of infographics on news elaboration, acquisition, and evaluation: Prior knowledge and issue involvement as moderators. *New Media & Society*, 18, 1579-1598. <https://doi.org/10.1177/1461444814567982>
- Lee, E.-J., & Kim, Y. W. (2014). How social is Twitter use? Affiliative tendency and communication competence as predictors. *Computers in Human Behavior*, 39, 296-305.
<https://doi.org/10.1016/j.chb.2014.07.034>

AWARDS

The Society for the Psychological Study of Social Issues (SPSSI) Grant-in-Aid, 2022
 George D. McCune Dissertation Fellowship, University of California Santa Barbara, 2021-2022
 Graduate Division Dissertation Fellowship, University of California Santa Barbara, 2022
 Graduation with Honors, Handong Global University, 2009
 Highest Distinction Award, Handong Global University, 2004-2008

FIELDS OF STUDY

Computer-mediated Communication
 Information Communication Technology
 Digital Civility
 Online Public Sphere

ABSTRACT

The Bystander Public for Online Incivility:

A Development of the Bystander Counter-Intervention Model

Yeweon Kim

The dissertation proposes a theoretical framework that elucidates the decision-making process behind uncivil bystander behavior (*counter-intervention*). As an adaptation of the classic bystander intervention model (Latané & Darley, 1970), the *Bystander Counter-intervention Model* is developed to inform a multi-step process that outlines the psychological antecedents that contribute to online counter-intervention. Using mixed research methods, the dissertation explores facilitators of counter-intervention and weaves the findings into the new predictive model. The investigation also tackles other behavioral paths for bystanders, such as confronting (*intervention*) and ignoring (*non-intervention*), to validate the original bystander intervention model in the context of online incivility.

Two studies are presented that used a non-representative sample of U.S. adults who have experiences with online incivility as bystanders. First, a preliminary interview study ($N = 26$) was conducted to obtain qualitative insights into potential predictors of counter-intervention. The interview findings were assessed to inform the survey design. Three different surveys were conducted with those who have experiences of counter-intervening, intervening, and/or non-intervening in uncivil online discussions. Participants joined in one to three surveys depending on which type(s) of bystander behavior they had ever engaged in (Total $N = 372$; counter-intervention survey: $N = 206$; intervention survey: $N = 311$; non-

intervention survey: $N = 300$). The data from each survey were analyzed separately to illustrate different procedural steps that lead bystanders to enact each behavior.

Though the survey findings on counter-intervention did not fully support the initial hypotheses, post-hoc analyses led to a reformulated model. Findings from the alternative model reveal that bystanders who justify incivility perceive greater benefits than costs of supporting it as they become more habitually exposed to uncivil online discussions. Such perception of the benefits (vs. costs) of incivility support moves bystanders to feel a need to show that support personally (*counter-intervention responsibility*) but only when they have sufficient confidence to take such actions (*counter-intervention efficacy*). If all these steps are completed successfully, the likelihood of counter-intervention increases. Meanwhile, the survey findings on intervention and non-intervention did not sufficiently substantiate Latané and Darley's (1970) model. However, the reduced model tests corroborated that intervention is more likely, whereas non-intervention is less likely, when bystanders perceive incivility to be more severe and, in turn, feel greater responsibility and efficacy to confront it.

The dissertation demonstrates the merits of traditional bystander scholarship as well as advances it by documenting the hitherto unknown social psychological mechanisms of negative bystander behavior. The dissertation also illuminates online bystanders as dynamic and multi-faceted actors with shifting responses—prosocial, antisocial, or passive—as result of different step-by-step processes that occur depending on internal motivational states and situational demands. Findings are also of great practical significance to professionals and policymakers invested in bystander-focused approaches to online incivility prevention. The dissertation highlights the importance of multilayered behavior-modification guidelines to address this complex problem.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION	1
PHENOMENON OBSERVATION.....	1
THEORETICAL TARGET: THE BYSTANDER PUBLIC	3
CONTEXTUALIZING ONLINE INCIVILITY TODAY	6
DISSERTATION OUTLINE	11
FINDINGS AND CONTRIBUTIONS	15
CHAPTER II: BYSTANDER COUNTER-INTERVENTION IN ONLINE INCIVILITY	17
CONCEPTUALIZATION OF ANTISOCIAL BYSTANDER BEHAVIOR	17
POTENTIAL FACILITATORS OF BYSTANDER COUNTER-INTERVENTION	20
<i>Desensitization and Justification of Online Incivility</i>	21
<i>Senses of Responsibility and Efficacy for Counter-intervention</i>	23
<i>Perceived Benefits and Costs of Counter-intervention</i>	27
<i>Pro-perpetrator Attitudes and Pro-incivility Social Norms</i>	29
CHAPTER III: INTERVIEW METHOD	32
PARTICIPANTS	32
PROCEDURE.....	35
MEASURES	37
<i>Pre-Interview Screening Survey</i>	37
<i>Interview</i>	39
<i>Post-interview Survey</i>	41
ANALYTIC APPROACH.....	42
CHAPTER IV: INTERVIEW RESULTS.....	44
DESCRIPTIVE INFORMATION OF INTERVIEW PARTICIPANTS.....	44
RECALL OF BYSTANDER COUNTER-INTERVENTION EXPERIENCE	46
QUALITATIVE ANALYSES OF BYSTANDER COUNTER-INTERVENTION FACILITATORS	47
<i>Desensitization to Online Incivility</i>	47
<i>Justification of Online Incivility</i>	49
<i>Responsibility for Counter-intervention</i>	53
<i>Efficacy for Counter-intervention</i>	55
<i>Perceived Benefits of Counter-intervention</i>	57
<i>Perceived Costs of Counter-intervention</i>	59
<i>Attitudes toward Perpetrators</i>	61
<i>Social Norms Regarding Incivility</i>	63
CHAPTER V: INTERVIEW STUDY DISCUSSION.....	65
INTERPRETATIONS.....	65
LIMITATIONS	69
CHAPTER VI: TESTS OF THE BYSTANDER BEHAVIOR MODELS.....	71
MECHANISMS OF BYSTANDER COUNTER-INTERVENTION	71
<i>Online Incivility Exposure and Pro-perpetrator Attitudes</i>	73
<i>Justification of Incivility and Benefit-Cost Perception of Counter-intervention</i>	74
<i>Responsibility for Counter-intervention and Pro-incivility Social Norms</i>	76
<i>Self-Efficacy for Counter-intervention</i>	78
<i>A Multi-Step Path to Bystander Counter-intervention</i>	79

MECHANISMS OF BYSTANDER INTERVENTION AND NON-INTERVENTION	83
SUMMARY OF HYPOTHESES	86
CHAPTER VII: SURVEY METHOD	87
PARTICIPANTS AND PROCEDURE	88
<i>Screening Survey</i>	88
<i>Main Surveys</i>	90
MEASURES	94
<i>Online Bystander Behavior</i>	94
<i>Predictors of Bystander Counter-intervention</i>	97
<i>Predictors of Bystander Intervention and Non-Intervention</i>	101
<i>Screening and Control Measures</i>	103
ANALYTIC APPROACH.....	105
CHAPTER VIII: SURVEY RESULTS	107
DESCRIPTIVE INFORMATION OF SURVEY PARTICIPANTS.....	107
DATA ANALYSES.....	108
TEST OF THE INITIAL BYSTANDER COUNTER-INTERVENTION MODEL	110
<i>Direct Paths to Counter-intervention</i>	111
<i>Indirect Paths to Counter-intervention</i>	112
<i>Summary</i>	113
TEST OF THE ALTERNATIVE BYSTANDER COUNTER-INTERVENTION MODEL	114
<i>Joint Effects of Exposure and Justification</i>	115
<i>Joint Effects of BCI Benefit–Cost and BCI Efficacy</i>	116
<i>A Multi-Step Path to Bystander Counter-intervention</i>	117
<i>Direct Paths to Counter-intervention</i>	119
<i>Indirect Paths to Counter-intervention</i>	119
<i>Summary</i>	122
TESTS OF THE BYSTANDER INTERVENTION AND NON-INTERVENTION MODELS	123
<i>Direct Paths to Intervention and Non-intervention (H6a-d and H7a-d)</i>	124
<i>Indirect Paths to Intervention and Non-intervention (H6e and H7e)</i>	125
<i>Reduced Model Tests</i>	126
<i>Summary</i>	128
CHAPTER IX: SURVEY STUDY DISCUSSION	129
SUMMARIES AND INTERPRETATIONS	129
<i>The Initial Bystander Counter-intervention Model</i>	129
<i>The Alternative Bystander Counter-intervention Model</i>	132
<i>The Bystander Intervention and Non-intervention Models</i>	135
LIMITATIONS AND FUTURE RESEARCH DIRECTIONS	138
CHAPTER X: GENERAL DISCUSSION	140
THEORETICAL CONTRIBUTIONS	141
PRACTICAL SUGGESTIONS	145
FUTURE RESEARCH	148
CONCLUSION	150
REFERENCES	152
APPENDIX A	192
INTERVIEW PARTICIPANT RECRUITMENT POST	192

PRE-INTERVIEW SCREENING SURVEY QUESTIONNAIRE	193
POST-INTERVIEW SURVEY QUESTIONNAIRE.....	194
APPENDIX B	195
SURVEY PARTICIPANT RECRUITMENT POST.....	195
SCREENING SURVEY QUESTIONNAIRE.....	196
BYSTANDER COUNTER-INTERVENTION (BCI) SURVEY QUESTIONNAIRE	199
BYSTANDER INTERVENTION (BI) SURVEY QUESTIONNAIRE.....	201
BYSTANDER NON-INTERVENTION (BNI) SURVEY QUESTIONNAIRE	202

LIST OF TABLES

Table 1. Descriptive Statistics of Interview Participants	45
Table 2. Number of Survey Participants.....	91
Table 3. Descriptive Statistics of Survey Participants.....	92
Table 4. Intercorrelations among the Key Variables	109

LIST OF FIGURES

Figure 1. A Hypothesized Model of Bystander Counter-intervention.....	83
Figure 2. Hypothesized Models of Bystander Intervention and Non-Intervention	86
Figure 3. Test Results of the Initial Bystander Counter-intervention Model	114
Figure 4. A Proposed Alternative Bystander Counter-intervention Model	118
Figure 5. Exposure x Justification and Benefit–Cost x Efficacy Interactions	121
Figure 6. Test Results of the Alternative Bystander Counter-intervention Model	123
Figure 7. Test Results of the Full Bystander Intervention/Non-Intervention Models	128
Figure 8. Test Results of the Reduced Bystander Intervention/Non-Intervention Models .	129

Chapter I: Introduction

Phenomenon Observation

Although policymakers have long labeled *electronic aggression*—any kind of bullying, harassment, and hateful acts (e.g., making fun of someone, making rude or mean comments, making threatening or hurtful comments, spreading rumors or lies) perpetrated via computer-mediated communication (CMC)—as a critical public health concern (Centers for Disease Control and Prevention, 2009), it has increasingly intensified and even been considered a “new normal” in today’s online world (Antoci et al., 2018; Sabatini & Reggiani, 2019). A substantial majority of the U.S. public (78%) perceive that the uncivil tone and nature of discourse around the country has escalated to the point of posing dramatic social risks and crises (Pew Research Center, 2019). Among U.S. adults who responded to a nationwide survey on “Civility in America,” 9 in 10 said that the decrease in polite and respectful conduct and expression could lead to an atmosphere of intimidation and threats (89%), harassment (89%), discrimination (88%), violence (88%) and cyberbullying (87%), and 1 in 2 attributed blame for such severe civility deficit to the Internet and social media (69%)—where a significant share of the world’s public communication takes place in the 21st century (Weber Shandwick, 2017). At a time when 91.8% of the total population of the U.S. has access to the Internet (Petrosyan, 2022), and 97% of teens use it daily (Pew Research Center, 2022b), this is a disturbing trend.

A 2022 survey by the Anti-Defamation League, with a nationally representative base sample, depicted the daily realities of online incivility (Anti-Defamation League, 2022). It revealed that uncivil dialogue and hate speech are undoubtedly a common aspect of online lives among U.S. adults, taking place disproportionately on social media (e.g., Facebook

68%; Instagram 26%; Twitter 23%; YouTube 20%). 40% of this survey's respondents reported personally experiencing some type of online harassment (e.g., name-calling, purposeful embarrassment), with 27% of respondents experiencing severe online harassment (e.g., physical threats, sexual harassment). A large portion of respondents reported that they were harassed due to an aspect of their identity (58%; e.g., race/ethnicity, gender, religion, sexual orientation, physical appearance, disability) and political views (44%). Accordingly, people in marginalized identity groups face high rates of hate-based online harassment (65% compared to 38% of non-marginalized groups), especially for LGBTQ+ people (66%), Hispanics/Latinos (40%), Asian Americans (39%), women (37%), Jews (37%), Black/African Americans (34%), and Muslims (33%). Especially, race and ethnicity-based online harassment has increased against Asian Americans (21% in 2021 to 39% in 2022; paralleling the rise in anti-Asian hate during the pandemic), Hispanics/Latinos (31% in 2021 to 40% in 2022; linked to the national debate over immigration), and Black/African Americans (33% in 2021 to 34% in 2022; as the backlash against the Black Lives Matter protests), whereas harassment targeting White people has decreased (46% in 2021 to 41% in 2022). This is consistent with a report that higher shares of Asian Americans (36%), Black/African Americans (28%), and Hispanics/Latinos (22%) have seen their daily lives disrupted and unsafe due to fears of race and ethnicity-motivated violent crimes, compared with just 12% of White people (Pew Research Center, 2022a).

Particularly, 64% of U.S. adults (74% of Republicans; 57% of Democrats) were found to be more critical of social media's negative impact on their national democracy, compared to people in other countries, believing that it makes people less civil in the way they talk about politics and less likely to accept those from different social groups (Pew

Research Center, 2022c). More complicatedly, the algorithmic mechanisms of social media platforms such as Twitter and Facebook fuel the fire behind online incivility because they are built in a way that increases visibility for posts that provoke extreme audience reactions (Jedicke, 2020). This concern is illustrated in the 2019 case of four black, indigenous, and people of color (BIPOC) congresswomen (“The Squad”) being the targets of a large amount of incivility on Twitter due to their racial, gender, and religious minority identities in the U.S. A study analyzing this incident (Borah et al., 2022) reported that approximately two-thirds of 102,815 replies to the congresswomen’s tweets contained at least one type of incivility that expressed discrimination and hate sentiments towards the minority groups they represented (e.g., immigrants, Muslims, Black/African Americans), including: name-calling, stereotypes, vulgarity, aspirations, and pejorative wording. The problem is more than overwhelming negativity in the tone of public discourse. These instances can also pose threats to individual rights and threats to democracy (Borah et al., 2022). A particularly challenging problem in this context is that the number of uncivil replies against the minority groups almost doubled after the former president’s tweet attack on these women. This fact further reveals an aspect of online incivility that permeates contemporary U.S. politics, feeds a polarized society, and threatens democratic values.

Theoretical Target: The Bystander Public

As reflected in this incident, what makes online incivility a “new normal” is the occurrence of supportive reactions following an initial incident of aggressive or hateful conduct. The current dissertation starts from the perspective that a portion of a third-party crowd contributes to this undesired phenomenon. Borrowing the notion of *bystander public* (Snow et al., 1981), this dissertation focuses on a diffuse collective of mass-mediated

spectators who are not directly targeted by online incivility and may choose to support or even promote it. This motivates the dissertation to theorize such problematic responses to incivility by introducing a new concept of bystander behavior, namely *bystander counter-intervention*, as an adaptation of the conventional notion of *bystander intervention* in social psychology (Latané & Darley, 1970). The dissertation aims to create a new theoretical framework that elucidates the decision-making process underlying bystander behavior that is reinforcing or assisting in online incivility. Namely, the *Bystander Counter-intervention Model* (hereafter ‘BCI model’) is developed to inform a multi-step process that contributes to such negative bystander behavior. Findings reveal that bystanders who justify incivility in online discussions tend to perceive the benefits rather than the costs of supporting it as they become more exposed to such uncivil situations. Their perception of the benefits (vs. costs) of incivility support moves them to feel a need to personally show that support (i.e., *counter-intervention responsibility*). However, this process is found only when they have sufficient confidence to take such actions (i.e., *counter-intervention efficacy*). If all these steps are completed successfully, the likelihood of counter-intervention increases.

The motivation for this investigation is provided by the need to test the validity of Latané and Darley’s (1970) *Bystander Intervention Model* (hereafter ‘BI model’), a progenitor of the bystander research paradigm in social psychology, in the context of online incivility. To do this, this dissertation also examines other behavioral options for bystanders, such as confronting (i.e., *intervention*) or ignoring (i.e., *non-intervention*) of incivility. This classic model informs a multi-step process that bystanders go through to intervene (or not) in a critical situation—an incident of a “perpetrator” committing an antisocial or dangerous act against a “target (i.e., victim)” and, in many cases, involving “bystanders” who witness

the incident. Although the current investigation of the BI model did not fully validate the same process in the context of online incivility, the reduced model tests supported most parts of the original predictions of intervention and non-intervention. These findings are elaborated on in the discussion section of this dissertation.

With the aim of unveiling the dynamics and theoretical underpinnings of bystander counter-intervention, the current dissertation situates the topic of bystander behavior in the context of uncivil online discussions. A specific focus is on a portion of the bystander public who might amplify the spread of incivility, as opposed to those who perform corrective actions or commit inactions. To understand this phenomenon, the dissertation starts with observing the impacts of online incivility on public life and the range of bystander responses to it. The bulk of the dissertation is then devoted to exploring plausible explanations for the psychological mechanism of counter-intervention. Given the paucity of empirical evidence in this area of research, some foundational research questions are derived from a comprehensive literature review on the potential facilitators of counter-intervention and examined with a preliminarily qualitative study using interviews. The interview findings are incorporated into the development of specific hypotheses that guide the design of the BCI model. This new model, and the original BI model, are tested using quantitative data collected from a series of surveys. Findings are then interpreted as evidence that some different decision-making processes are set in play for counter-interveners, interveners, and non-interveners. The resultant knowledge will be of relevance not only to researchers of bystander scholarship but also practitioners invested in bystander-focused approaches to online incivility prevention.

Contextualizing Online Incivility Today

The most worrisome aspect of online incivility is its detrimental effect on targets, according to empirical reports (e.g., Anti-Defamation League, 2021; John et al., 2018; Pew Research Center, 2017b), in terms of psychological (e.g., anxiety, distress, depression), physical (e.g., insomnia, self-harm, suicide), interpersonal (e.g., distrust, relational conflict), and professional or even economic impacts (e.g., workplace bullying, unemployment, financial loss). These reports show that hate-based online harassment undermines potentials of digital democracy by driving people to leave online discussion platforms that increasingly become central to public discourse. For example, almost a third of respondents in the Anti-Defamation League's survey (29%) reported that they stopped or reduced social media use (e.g., Facebook, Twitter, Instagram) due to the actual or potential experience of being attacked, by posting less often, changing the privacy settings, and deleting these applications from their devices (Anti-Defamation League, 2022). Even worse, such consequences are not only limited to targets because online incivility is often performed publicly. According to the National College Student Bystander Intervention Study, 54% of respondents reported witnessing acts of bias, exclusion, discrimination, or violence directed at a particular and often marginalized social group (WITH US Center for Bystander Intervention at Cal Poly, 2020). Simply observing such acts causes similar mental health concerns for non-targets as targets of the acts (Magellan Health, 2012). This effect might be more pronounced because a single incident of online incivility can be observed by a larger share of the population beyond targets—as evidenced that 66% of adults observed online harassment compared to 41% who personally experienced it (Pew Research Center, 2017b) and that 87% of teens observed cyberbullying compared to 26% who personally experienced it (Siciliano, 2014).

The impacts on those non-targets include feeling anxious about possibly being targeted by similar treatment (34%) and, as a result, adjusting one's own online presence (47%; e.g., refraining from posting online, changing the privacy settings of online profiles, stopping the use of online services; Pew Research Center, 2017a). These results may cause a chilling effect on the bystander public by deterring them from joining online discussion activities (Hwang et al., 2014; Papacharissi, 2004).

However, the percentage of people who say online incivility is a major issue in American society has decreased in the past years (62% in 2017 to 55% in 2020), supporting a renewed concern that the public has become increasingly desensitized to and tolerant of this problem (Pew Research Center, 2021). This is especially manifested among bystanders who have not personally experienced incivility (60%) compared to those who have (52%). There is a growing body of evidence that complicates this concern by showing the positive effects of online incivility, such that the uncivil nature of an online comment does not impair audience perception of the comment in terms of its favorability (Graf et al., 2017) or informativeness and persuasiveness (Ng & Detenber, 2005). The public does not always hate seeing impoliteness and in some cases they show greater engagement in uncivil (vs. civil) online discussions (Pew Research Center, 2017a). Research has found that uncivil comments in online news discussions are likely to gain more attention (Kwon & Cho, 2016), receive more recommendations from an audience (Muddiman & Stroud, 2017), and elicit more willingness for political participation by arousing aggressive intentions (Chen & Lu, 2017)—contrary to the chilling effect hypothesis. This reveals the controversial role of incivility as it may rather facilitate sociopolitical conversation among a portion of the bystander public. Such mixed reactions to incivility are markers of the fact that the

bystander public is not a homogeneous group but a collective of diverse actors with different and potentially competing interests pooling together while an incivility scenario evolves.

Previous research has documented different types of bystander behavior observed in cyberbullying incidents (see the systematic review by Polanco-Levicán & Salvo-Garrido, 2021), generally summarized into three types: (a) prosocial behavior of defending the target (i.e., intervention), (b) passive behavior of ignoring the incident (i.e., non-intervention), and (c) antisocial behavior of reinforcing or assisting the perpetrator (i.e., counter-intervention).

Indeed, it is well-established that when bystanders stand up the target of online incivility (i.e., intervention) they can make a tremendous and positive difference. Research shows that intervention by a small number of users against online incivility can influence a larger audience (Schieb & Preuss, 2016) and such users in a moderator role can break the aggressive cycle (Ouvrein et al., 2018). Thus, there have been efforts in the real world to encourage bystanders to recognize their decisive role in resolving the problem of incivility. For example, some academic institutions or non-governmental organizations have developed intervention programs and initiatives to train bystanders as potential allies in efforts to rid digital environments of uncivil discourse (e.g., “Initiative to Revive Civility”; NICD, n.d.), cyberbullying (e.g., “Social Media TestDrive”; DiFranzo et al., 2019), and hate speech (e.g., “Civic.net Initiative”; Amadeu Antonio Foundation, n.d.). As a more accessible alternative, we often see educational guidelines or toolkits distributed online, which provide practical tips for enacting intervention against incivility incidents (e.g., Terry, 2020; Wright, 2020). Bystanders themselves have also led social movements (e.g., “#ichbinhier [#iamhere]”; I Am Here International.com, n.d.) and national campaigns (e.g., “No Hate Speech Movement”; Council of Europe, n.d.) to collectively combat hate speech online. In one

cyberbullying survey, the majority of respondents said that they would choose to intervene when they saw mean posts, comments, or pictures online, if “what was happening was unfair or morally wrong,” “was extremely hurtful behavior,” “intervention could be done anonymously,” and “others respected me for intervening” (Joyce et al., 2015).

However, efforts to promote intervention might be ineffective if bystanders chose to be invisible and remain silent (i.e., non-intervention). This is the most common outcome in reality, as we have long seen from offline bystanders in emergency situations since the infamous 1964 Kitty Genovese murder that was used as an exemplar of *bystander apathy* in early bystander research (Latané & Darley, 1968). This still happens today when bystanders witness bias, exclusion, and discrimination in their daily life and choose not to react in most of such cases (e.g., 67% of bystanders reported “I didn’t say or do anything”; WITH US Center for Bystander Intervention at Cal Poly, 2020). What may be as valuable as knowing the reasons why bystanders would intervene is understanding the reasons when they would not. Respondents of the aforementioned cyberbullying survey (Joyce et al., 2015) were actually found to take *no* action against the observed mean content online, because they felt “nothing can be done about it,” “it is not about me,” “it is not that serious,” and “the target deserves it.” Or, bystanders may not decide to act, even though they want to do so, due to “fear of retaliation and being bullied themselves,” “fear of losing their social status,” “lack of knowledge about the individuals involved, the incident, and whether they perceive someone to be right or wrong in the situation,” and “not knowing what to do” (stopbullying.gov, n.d.). This refusal to act is often the bystander public’s response online. About three-in-ten U.S. adults say that they never responded after noticing harassing behavior directed at someone online (Pew Research Center, 2017b). Today’s bystanders

with smartphones often join in the act of capturing and recording problematic incidents rather than becoming involved. Recently, we have seen tragic episodes where bystanders of assault uploaded or live-streamed video footage of the incident on social media (see Bates, 2014; McPhate, 2016; Narula, 2017 for some example episodes), instead of stepping in to help the target or at least calling the police. This act focuses on creating a digital record to draw mass attention to the incident and incite strong emotional responses (e.g., outrage, frustration), while avoiding taking analog action to affect an outcome of the incident at that moment (Han & Bowen, 2021; Xu, 2015). This new phenomenon is increasingly common—for example, it is emerging as a popular term, *wangluo weiguan*, that means “online surrounding gaze” (Bandurski, 2011) in China’s online activism. Such “unethical witnessing” shows a modern manifestation of the *bystander effect* (Badalge, 2017).

The problem becomes even more intractable when bystanders choose to come to the aid of the perpetrator (i.e., counter-intervention). A more tragic aspect of the “online surrounding gaze” phenomenon is that such disturbing digital records have often elicited favorable responses (e.g., liking, replying, sharing) from an online audience, adding to the suffering of the target (e.g., Bates, 2014; McPhate, 2016). Indeed, the studies by Panumaporn et al. (2020) and Erreygers et al. (2016) noted that at least 4.6% to 26.3% of bystanders counter-intervene in cyberbullying situations (vs. 34.6% to 42.3% intervene; 28% to 53.6% non-intervene). Based upon these observations, there is no question that it is possible for some portion of the bystander public to choose to encourage the perpetrator and revictimize the target by approving and forwarding the problematic online content. Online bystander research has focused heavily on comparisons between intervention and non-intervention (e.g., Naab et al., 2018; Polder-Verikie, 2012; Wang & Kim, 2021), but

relatively less attention has been paid to another option of bystander behavior that is supportive of the perpetrator. However, Panumaporn et al.'s (2020) study finds that nearly the same number of bystanders join in cyberbullying (26.3%) as ignore it (28%), suggesting that a considerable portion of the population may be involved in this social problem. A decision-making process for bystanders to endorse the perpetrator could be driven by personal motivations to fulfill their own needs for safety, connection, and power as well as by attempts to comply with the implicit and explicit rules of the culture from peers (Ferrás & Selman, 2014). Mirroring the perpetrator's uncivil act, especially under a social climate that encourages aggression, might be a way to confirm group inclusion, avoid victimization, and gain social capital from deviant sources (Evans & Smokowski, 2017). It is also worth noting that such negative bystander behavior is amplified online by the affordances of digital environments that allow anonymity and are subject to reduced regulation (Runions & Bak, 2015). In short, core to the culture of online incivility is a large share of the bystander public who serves as an "enabler" to spread it.

Dissertation Outline

Given this critical aspect of antisocial bystander behavior, the current dissertation rests on the belief that online incivility should be addressed as an act in two parts: the original incident by an incivility perpetrator, and behaviors of bystanders who reinforce or assist the perpetrator through liking, replying, sharing, or any other supportive action available online. The main focus of the dissertation is on this significant yet understudied portion of the bystander public that serves to amplify online incivility. There is surprisingly little data available on the specific psychological dynamics of this part of the online population. These gaps point to an important question that has yet to be answered regarding:

what factors facilitate bystanders' decision to tolerate and even support online incivility. The current dissertation attempts to address these questions using both qualitative and quantitative methods and to weave the resultant findings into a new theoretical framework focusing on counter-intervention. Additionally, it is a further motivation of this dissertation to elucidate the blurred and permeable boundaries between bystander roles. Traditionally, bystander behavior has been dichotomized as "intervene or not" and considered by and large fixed. However, some research has suggested that online bystanders are likely to behave on a continuous and fluctuating spectrum rather than within strict categories (e.g., DeSmet et al., 2014). This is reflected in the findings that people facing an online incivility incident switched between different roles that ranged from bystanders to perpetrators to targets as the incident progressed (e.g., #GorillaLivesMatter case referenced by Chen, 2017; Sun & Fichman, 2020). Evidence like this suggests there might be different motivating factors that trigger different bystander roles. Thus, a central goal of this dissertation is to explore the decision-making processes that lead bystanders to different behavioral outcomes. To that end, the scope of the investigation includes tests of not only the newly proposed BCI model but also the original BI model. Findings will provide insights into the mechanisms underlying online bystander dynamics through which they transform into counter-interveners, interveners, or non-interveners.

The current dissertation is composed of four main sections. The first section (Chapter 2) provides an overview of the theoretical underpinnings of this research, focusing on the psychological "why" of bystander counter-intervention. Given the relative paucity of existing knowledge in this area, the dissertation intends to openly explore what might drive people to engage in negative bystander behavior, by comprehensively observing their

intrinsic and extrinsic motivations or concerns that manifest in uncivil online discussions. To this end, the scope of an initial literature review goes beyond the traditional bystander research to include some well-established behavior prediction theories in social psychology, as well as recent CMC studies of online harassment and cyberbullying. This yields a list of potential predictors of counter-intervention that could be considered to constitute the proposed BCI model.

The next three sections of the dissertation are devoted to assessing whether and how those predictors of counter-intervention should be incorporated into the new model to be tested. Two empirical studies are presented that used samples of U.S. adults who have experiences with online incivility as bystanders. The second section (Chapter 3) presents a preliminary interview study that was conducted to obtain qualitative insights into potential predictors of counter-intervention. A series of one-on-one interviews were held via email with a small non-representative sample of U.S. adults ($N = 26$) who have experiences of counter-intervening in uncivil discussions during their online activities on social media, news websites, or other CMC platforms. The interview questions were designed to elicit information about counter-interveners' desensitization to and justification of online incivility, feelings of responsibility and efficacy for counter-intervention, attitudes and perceived social norms toward incivility perpetrators, and benefit–cost perceptions regarding counter-intervention. The collected data were examined using a *thematic analysis approach* (Braun & Clarke, 2006) to identify the common factors contributing to bystanders' decision to support incivility. The interview findings were assessed to inform the design of the main survey study in terms of what factors should be incorporated in the prediction of counter-

intervention, and how. Methodologically, some of the interview excerpts were used to generate the wording of the survey items.

From the third section (Chapter 4), the dissertation focuses entirely on the development of the BCI model. An additional literature review is provided, along with an interpretation of the interview findings, to formulate a series of hypotheses (H1-5) that guide the model design. Adapting the framework of Latané and Darley's (1970) BI model, the new model delineates a multi-step process leading to bystander counter-intervention, which includes direct and indirect paths of predictors to the behavioral outcome, as well as incorporates possible moderating factors that might strengthen or weaken those paths. In addition, two more sets of hypotheses (H6-7) are postulated to examine the applicability of Latané and Darley's (1970) model within the context of online incivility. The focus is on testing whether the original BI model's factors are useful to explain how likely bystanders engage in intervention and non-intervention in uncivil online discussions.

The final section (Chapter 5) presents a main survey study that was conducted to test the hypotheses. Three different surveys were administered with a non-representative sample of U.S. adults who have experiences of counter-intervening, intervening, and/or non-intervening in the same context of online incivility as that of the interview study. The focus of each survey is on investigating whether they go through particular procedural steps before deciding to enact each behavior. Participants joined in one to three surveys depending on which type(s) of bystander behavior they have ever engaged in (Total $N = 372$; bystander counter-intervention survey: $N = 206$; bystander intervention survey: $N = 311$; bystander non-intervention survey: $N = 300$). The data from each survey were analyzed separately using a *conditional process modeling* method (Hayes & Preacher, 2013) with goals of

elucidating a hitherto unknown mechanism of counter-intervention and validating the traditional predictions of intervention and non-intervention (Latané & Darley, 1970).

Findings and Contributions

Findings of the interview study revealed the common beliefs and values held among counter-interveners, including how they perceive uncivil online discussions and reflect on their behavioral decisions. The key takeaways from this study were compiled to design the survey study to construct a new model. By replicating and extending the original BI model, the current research presents an adapted BCI model that describes an alternative multi-step process by which the opposite behavioral outcome takes place. The results of the initial analysis did not fully support the hypothesized model but provided a sense of which variables are important in predicting counter-intervention and how they interrelate with each other, revealing that they do not constitute the same sequence as that of the original BI model. After revisiting the literature to gain further clarity about those key variables, post-hoc exploratory research questions were posed to reformulate the model for a more robust prediction. The resultant model was considerably improved in terms of model fit in the post-hoc test and demonstrated several factors that individually and collectively contribute to counter-intervention. On the one hand, two significant predictors were found for counter-intervention: frequency of online incivility exposure and feeling of counter-intervention responsibility. This suggests that counter-intervention is more likely when bystanders observe online incivility more often and when they feel greater need to support it. On the other hand, these two factors formed part of a decision-making process for counter-intervention that is likely to occur among bystanders who deem incivility justifiable. For this type of bystanders, more frequent exposure to incivility increases the perceived benefits (vs.

cost) of supporting it. This perception leads them to feel a need to reinforce or assist the perpetrator and, in turn, engage in counter-intervention. However, this effect occurs only with high self-efficacy to perform the intended behavior. Otherwise, counter-intervention is less likely due to decreased motivation to support the perpetrator.

Meanwhile, the survey findings on intervention and non-intervention did not sufficiently substantiate Latané and Darley's (1970) model. However, it did find that the exposure and responsibility factors were also important in this original model, as they were identified influential in the BCI model. First, frequency of incivility exposure was positively related to that of intervention but not non-intervention. This finding—along with the former finding of the positive association between exposure and counter-intervention—suggests that bystanders who are more often faced with uncivil situations become more likely to “act” in some way. Moreover, assuming responsibility for confronting the perpetrator was found to increase intervention but decrease non-intervention. Again, this finding—along with the former finding of counter-intervention responsibility—highlights the responsibility factor as a crucial boundary condition for bystander behavior. Finally, the current research corroborated that the intermediate steps in the original BI model worked sequentially as predicted, such that perceiving severity of incivility leads to a greater sense of responsibility to confront it and, in turn, a greater sense of efficacy to do so. These three steps further explain both intervention and non-intervention, according to the reduced model tests that only included the direct and indirect effects of incivility exposure on each behavioral outcome. The related analyses showed that intervention is more likely, whereas non-intervention is less likely, when bystanders perceive incivility more severe and, in turn, feel greater responsibility and efficacy to confront the perpetrator.

The current dissertation demonstrates the benefits of traditional bystander scholarship that informs the underlying aspects of human behavior in a critical social situation. The newly gained knowledge from the BCI model advances the bystander literature by revealing social psychological mechanisms of negative bystander behavior that has yet to be well documented. Mixed methods are used to ensure that comprehensive empirical evidence is made available to theorize the dynamics of counter-interveners. This research also illuminates online bystanders as dynamic and multi-faceted actors with shifting responses to incivility. Depending on internal motivational states and situational demands, they undergo different step-by-step processes yielding either prosocial, antisocial, or passive behavior. These findings are of great practical significance for coping with online incivility, offering tips for designing bystander training programs or digital platforms in a way to interrupt the process of counter-intervention but stimulate that of intervention.

Chapter II: Bystander Counter-Intervention in Online Incivility

Conceptualization of Antisocial Bystander Behavior

Bystanders who act as reinforcers and assistants of online incivility are quite common in virtual public spheres and may be motivated by a variety of factors with a range of outcomes, both positive and negative. Despite the often-stated concern about the chilling effect of incivility on public discourse (Hwang et al., 2014; Papacharissi, 2004), evidence shows that social media posts containing political disagreement and indignant rhetoric, compared to other types of posts, are often likely to elicit more audience engagement (e.g., liking, replying, sharing; Pew Research Center, 2017a). Furthermore, incivility in online discussion is found to help better recall of the discussion (Hwang et al., 2016), as well as elicit greater intention to participate politically (Borah, 2014) or engage in requested

behavior (e.g., donation; Beckert & Ziegele, 2020). Antagonistic bystanders could further amplify online incivility and aggravate negative impacts (e.g., distress, harm) for targets (Jones et al., 2015). Thus, this type of bystanders should be closely investigated as a key socio-contextual factor underlying the perpetuation of antisocial behavior (Troop-Gordon et al., 2019). Research has also warned that people become more aggressive when they engage in pro-bullying bystander behavior cumulatively (Evans et al., 2018). Relatedly, playing a reinforcer or assistant role in cyberbullying increases the likelihood that someone will be a perpetrator (Orue et al., 2023). As such, an increasing body of evidence suggests that certain types of bystanders can play a consequential role in developing online incivility phenomena.

Consistent with the traditional Participant Role Scale, which classifies the specific behaviors of third-party participants in a bullying situation (Salmivalli et al., 1998), research has often identified four major roles that bystanders assume in offline or online bullying situations: defender, outsider, reinforcer, and assistant (e.g., Hong & Lee, 2022; O'Connell et al., 1999; Sarmiento et al., 2019; Quirk & Campbell, 2015). Reinforcer/ assistant bystanders do not initiate bullying but join in it by providing positive feedback to the perpetrator and negative feedback to the target (e.g., laughing, questioning), unlike outsiders taking no action despite being aware of the situation (Salmivalli et al., 2011; Williams et al., 2018). Evidence suggests that a portion of cyberbullying bystanders identify themselves as reinforcers or assistants rather than outsiders or defenders (e.g., Brody, 2021; Pepler et al., 2021; Song & Oh, 2017, 2018), consistent with traditional bullying bystanders. Some studies suggest that negative bystander behavior is more prevalent online than in face-to-face contact (Barlińska et al., 2013) due to perceived anonymity and resultant deindividuation effects in CMC settings (Chan et al., 2022).

Yet not much is known about the portion of the population who empowers online incivility. Although a wealth of knowledge on prosocial bystander behavior has accumulated since Latané and Darley (1968) first pioneered the bystander intervention paradigm in social psychology, less progress has been made to date in conceptualizing and operationalizing the opposite behavioral pattern. Previous research has explored the characteristics of those showing negative bystander behavior to bullying in terms of their personalities (e.g., moral sensitivity, empathy; DeSmet et al., 2016; Shen et al., 2022), demographics (e.g., gender, age; Cao & Lin, 2015; Van Cleemput et al., 2014), evaluations of the incident (e.g., perceived severity or expectancy violation of bullying; Brody, 2021; Gutzwiller-Helfenfinger & Perren, 2021), relationships with the perpetrator or the target (Bastiaensens et al., 2014; Williams et al., 2018), perceptions of the social climate (e.g., descriptive and injunctive social norms; Bastiaensens et al., 2016; Dang & Liu, 2020), and personal experiences of being a perpetrator or a target (e.g., Panumaporn et al., 2020; Zhao et al., 2023). However, we know relatively little about the cognitive mechanisms that inform the decision to engage in antisocial bystander behavior. Such types of bystanders might experience a complex and multifaceted decision-making process before acting to support the perpetrator, just as prosocial bystanders face a variety of difficulties and dilemmas before determining whether and how to defend the target (Pepler et al., 2021). Identifying the underlying mental operations of negative bystander behavior can advance the understanding of people who become increasingly antipathetic towards their political or ideological opposites by joining in uncivil and hateful dialogues. Such knowledge is essential to better strategize how to mitigate online incivility spread by deterring its enablers.

Potential Facilitators of Bystander Counter-intervention

The current dissertation seeks to obtain a holistic understanding of the potential facilitators of bystander counter-intervention behavior (hereafter “BCI behavior”). The primary focus is on the traditionally studied variables in the bystander literature that have grounded Latané and Darley’s BI model. This model posits that bystanders tend to intervene in an incident when they perceive it as a critical emergency (e.g., a threatening situation to the target) and feel personal responsibility and efficacy to intervene to solve the incident. The dissertation examines whether those explanations can be adapted to predict the opposite behavioral outcome. In doing so, some behavior prediction theories are also borrowed to identify internal and external contributors to BCI behavior. For example, theories of desensitization (e.g., Funk et al., 2004) and moral disengagement (e.g., Bandura, 1999) provide insights on emotional and cognitive states that might affect perceptions of online incivility. The dissertation also adapts the conventional notions of bystander responsibility and efficacy to the context of counter-intervention. This uncovers whether BCI behavior might depend on bystanders’ feeling of social accountability toward supporting the perpetrator (not the target) and their self-confidence in acting to demonstrate that support. Relatedly, attention is further given to possible socio-psychological benefits and costs associated with BCI behavior, given that bystanders weigh the positive and negative consequences of alternative behavioral options before acting (Fischer et al., 2011; Latané & Nida, 1981). This inquiry is also guided by the cost-reward model (Dovidio et al., 1991) that advocates an economic perspective (i.e., minimizing costs and maximizing rewards) in predicting human behavior. Finally, the dissertation makes use of social categorization theory (Turner et al., 1987) and social cognitive theory (Bandura, 1986) to capture the

contextually dependent nature of bystander behavior. These theories suggest that bystanders could exploit social contextual cues about those involved in the incident (perpetrator, target, other bystanders) to determine their response. The resultant insights are used to assess how bystanders' own attitudes and perceived social norms might affect their behavioral decision.

Based on the theories and models cited above, the current dissertation explores different theoretical determinants of bystander behavior to seek plausible explanations for the drivers of BCI behavior. This leads to the wide scope of the literature review of the following variables: desensitization and justification of online incivility, responsibility and efficacy for counter-intervention, benefits and costs of counter-intervention, and pro-perpetrator attitudes and pro-incivility social norms. All these variables are preliminarily examined in the interview study conducted in the first phase of this research. The following literature review section presents four research questions on counter-interveners, which are broadly framed to fit the exploratory nature of the interview study. More specific and focused hypotheses of the dissertation, which guide the development of the BCI model, are generated later based on the interview findings.

Desensitization and Justification of Online Incivility

Given that bystanders tend to intervene in an incident that they interpret as a critical emergency (Latané & Darley, 1970), the extant literature documents the importance of the perceived severity of the incident in offline (Fischer et al., 2011; Macaulay et al., 2019) as well as online contexts (e.g., Kazerooni et al., 2018; Leonhard et al., 2018). For example, bystanders who perceived cyberbullying (Koehler & Weber, 2018) and racist online comments (Dickter & Newton, 2013) as more offensive were more likely found to defend the target. Given such evidence, it is likely that bystanders may engage in the opposite

behavior if they perceive online incivility as *not* serious enough to be confronted. In assessing this assumption, the current research explores whether BCI behavior could be considered a marker of *desensitization*—reduction in cognitive, physiological, emotional, or behavioral responsiveness resulting from extensive exposure to aggressive behavior (Funk et al., 2004). Research has shown that people who habitually observed bullying became desensitized to it, showing decreases in empathy and concern toward targets over time (e.g., Gini et al., 2008; Pabian et al., 2016; Rudy & Linz, 2008). Given that bystanders who were frequently exposed to hate speech became less sensitive to it (Bilewicz & Soral, 2020) and perceived it as being normalized online (SELMA, 2019), emotional desensitization to incivility might contribute to BCI behavior.

However, this affective factor may not fully explain support for the perpetrator. Often people consider online harassment as a justifiable or deserved response when the perpetrator uses a framework of *retributive justice*—a belief that individuals who commit a morally wrong act receive a proportional punishment for their misdeeds (Blackwell et al., 2018; Carlsmith & Darley, 2008). In such cases bystanders might accommodate their cognitive schemas in accordance with the perpetrator's intent to legitimize their incivility. Justification is known to function as a palliative strategy that helps sustain aggression by rendering it acceptable or admissible (Fujiwara et al., 1999). Social learning theory also suggests that a perpetrator uses justification in narratives of their transgressive conducts to exonerate themselves from self-criticism as well as avoid refutation and retaliation from others (Bandura, 1973). For example, the perpetrator utilizes neutralization techniques (e.g., denial of the victim, denial of responsibility) to legitimize their cyberbullying acts (Zhang & Leidner, 2018) or attempts to glorify bullying as a collective action of a self-serving and

socially inclusive ritual to confirm their “normality” (Thornberg, 2015). These justification processes may also operate among bystanders standing up for the perpetrator. Evidence shows that reinforcer/assistant bystanders of bullying judged it to be right or at least permitted, by employing various strategies (e.g., blaming the target, negating or alleviating harm caused by the perpetrator; Hara, 2002). This type of bystanders may also resort to justification to avoid negative moral feelings (e.g., guilt, regret, shame) resulting from their involvement in incivility, as outsider bystanders of bullying do (de Oliveira et al., 2019).

All the evidence cited above reflects widespread public indifference or laissez-faire attitudes regarding online incivility, including: seeing comments of personal attacks or public shaming as jokes, linking hateful speech to freedom of expression, and underestimating the negative impacts of incivility (Jubany, 2015). Given the fact that there has been a decrease in the public’s belief that online harassment is a serious societal issue (Pew Research Center, 2021), the assumption that desensitization and justification of incivility may provoke BCI behavior deserves special attention. Thus, the first set of research questions explores:

- RQ1a-b.** Do counter-intervening bystanders show (a) emotional desensitization and (b) justification of online incivility?

Senses of Responsibility and Efficacy for Counter-intervention

Much of the research in the tradition of Latané and Darley’s (1970) BI model has focused on personal senses of responsibility (Obermaier et al., 2016; Wong et al., 2021) and efficacy (e.g., Clark & Bussey, 2020; Hayashi & Tahmasbi, 2020) that drive prosocial bystander behavior. ‘To take responsibility’ in the bystander context has meant that bystanders feel a need to intervene due to their personal moral values as well as a social

obligation (Latané & Darley, 1968). This has been applied to online contexts, such that accepting personal responsibility for confronting cyberbullying (e.g., DiFranzo et al., 2018) and uncivil user comments (e.g., Naab et al., 2018) has a positive relationship with a likelihood of intervention. Prosocial bystander behavior is less likely when bystanders ascribe responsibility to other bystanders (Van Cleemput et al., 2014), platform providers or law enforcement (Riedl et al., 2021), rather than themselves. The so-called *diffusion of responsibility* effect (Latané & Darley, 1968) has predicted passive bystander behavior, for example, from adolescent bystanders of cyberbullying who believe it should be resolved by adults (Ferreira et al., 2020) and from bystanders of workplace mobbing who attribute responsibility to the target (Mulder et al., 2016).

However, the focus of the current research is not on the reduced or diffused responsibility for intervention. A novel notion is introduced in this research to explore the opposite of the conventionally understood notion of bystander responsibility, namely *counter-intervention responsibility*. It refers to the degree to which bystanders feel a need to reinforce and assist the perpetrator. This new notion of bystander responsibility may align with motives of perpetrators who say “because I need to” as the reason why they commit bullying (Konnikova et al., 2015). For example, some common motives for bullying include the need to gain or maintain popularity, dominance, or social status (Compton et al., 2014; Romera et al., 2021; Steer et al., 2020). Or, some perpetrators mention the need to get revenge and reestablish justice by eradicating different views (König et al., 2010). Perpetrators of hate speech comments on news websites are motivated by the belief that their behavior is a sort of “mission” that should be carried out to defend their political and ideological values through hatred against the outgroup (Erjavec & Kovačič, 2012).

Perpetrators also behave often due to relational concerns of seeking a sense of belonging and acceptance even if their behavior is antisocial (Pronk & Zimmer-Gembeck, 2010; Solomontos-Kountouri & Strohmeier, 2021). Bystanders might be driven by a similar set of motivations if they believe that support for the perpetrator is a means for fitting into the given situation or if they consider it the shared responsibility to enhance the power and status of their own social group by reinforcing and assisting the perpetrator. Such relational concerns or group-based needs for affiliation and dominance might be a motivator for BCI behavior, given the fact that some bystanders feel personally responsible to join in bullying if they are friends with the bully (Burns et al., 2008).

Along with the responsibility factor, the bystander literature has often focused on self-efficacy that represents an individual's belief in their own capabilities in performing actions for desirable outcomes (Bandura, 1997). This is an important concept in social cognitive theory (Bandura, 1986)—a framework that explains the influence of personal and environmental factors on human behavior—as a predictor of an individual's action especially in overcoming obstacles or problems encountered. This factor increases determination to act in accordance with perceived easiness of the intended action to be completed (i.e., perceived behavioral control; Ajzen, 1991). Self-efficacy has been stressed as a crucial factor in predicting bystander behavior in an ego-threatening situation, also as the next-to-last step in the original BI model (Latané & Darley, 1970). It indicates a bystander's belief in their abilities to understand the given incident and know how to act, as well as feel able to put their thoughts into action based on effective strategies (e.g., Gini et al., 2008; Pöyhönen et al., 2012; Song & Oh, 2017). Studies have revealed that bystanders' self-perceived abilities to sense the target's emotion (i.e., empathic self-efficacy) and

successfully defend the target (i.e., defending self-efficacy) are significantly related to high rates of intervention in antisocial online situations (Chen et al., 2022; Clark & Bussey, 2020). Some studies have examined self-efficacy that reflects task-specific competencies or skills closely related to the given situation, such as political self-efficacy and writing self-efficacy that bystanders might need to confront uncivil comments on news websites (Ziegele et al., 2020). These existing notions of efficacy predict other types of bystander behavior, such that low defending self-efficacy is more likely to elicit the choice of remaining passive (Thornberg & Jungert, 2013) or defending aggressively (Bussey et al., 2020).

Despite the importance of self-efficacy in bystander scholarship, it may lack utility in explaining BCI behavior because bystanders' decisions to support the perpetrator are not likely related to judging themselves as unable to defend the target. Indeed, evidence shows no relationship between defending self-efficacy and reinforcing behavior (Pöyhönen et al., 2012). In this sense, a more precise formulation of the efficacy notion is needed to capture the opposite belief held by counter-interveners regarding their capabilities to act in support of the perpetrator. One possible related factor is aggression self-efficacy that enables an individual to enact relational (e.g., isolating another, spreading a rumor) or verbal aggression (e.g., name calling, teasing; Barchia & Bussey, 2011) and increases perceived behavioral control over cyberbullying perpetration (Festl, 2016). Aligned with this notion, this research explores a self-efficacious belief related to antisocial bystander behavior. Namely *counter-intervention efficacy* refers to the degree to which bystanders feel confident and effective in reinforcing or assisting the perpetrator.

In attempting to devise a novel conceptual framework of BCI behavior, this research envisages potentially important yet understudied notions of counter-intervention

responsibility and efficacy that more accurately reflect antisocial bystanders. Direct evidence that could help establish these two new concepts, albeit plausible, have not been documented in the bystander literature yet. Those factors need attention for addressing online incivility as it is not only the product of an individual perpetrator's act but also the "problem of many hands (i.e., bystanders)" (Thompson, 2014). Thus, the second set of research questions explores these new concepts.

RQ2a-b. Do counter-intervening bystanders feel (a) the need to support online incivility perpetrators and (b) confidence in their ability to demonstrate that support?

Perceived Benefits and Costs of Counter-intervention

Bystanders are known to weigh expected CO consequences for intervention (Fischer et al., 2011; Latané & Nida, 1981). Depending on the nature of the incident in question, the risks of intervention can range from mere inconvenience or evaluative apprehension to possible psychological or physical harm, which constrain the behavior (Clark & Word, 1974). Thus, bystanders analyze the potential benefits and costs associated with possible actions and then make a decision that can incur as few net costs (i.e., costs minus benefits) as possible (Wang, 2021). For example, the costs of intervention for bystanders involve the fear of misinterpreting the incident and thus being embarrassed of making an inappropriate response, risk of getting into trouble or being another target themselves, or concern of being negatively evaluated by others, while the benefits involve such things as self-pride for helping the target, reputation or praise gained from others, and expectation for bullying to decrease (Burn, 2009; Pöyhönen et al., 2012; Wang, 2021; Ziegele et al., 2020). Studies have shown that behavioral intention and enactment of intervention is more likely to occur

when bystanders perceive greater benefits than costs of defending the target (Chen et al., 2022; DeSmet et al., 2016; Spadafora et al., 2020).

The current research applies this cost-reward component (Dovidio et al., 1991) to the investigation of BCI behavior. In this research, perceptions of the benefits and costs of BCI behavior are operationally defined as cognitive appraisals of the values and barriers of support for the perpetrator. Bullying studies suggest that bystanders who have negative or conflicting expectations about the outcome of intervention—for example, a minimal likelihood of the target feeling better or a greater likelihood of losing their own status—are likely to choose other behavioral options (Pöyhönen et al., 2012). This might be more likely to occur when bystanders expect positive consequences rather than undesirable outcomes when taking the side of the perpetrator. Considering that bystanders can experience positive and negative responses to their behavior from different parties involved in the incident (i.e., target, perpetrator, and other bystanders; Moschella et al., 2016), and thus the outcomes of their behavior are related to the social dynamics of those parties (Spadafora et al., 2020), they may take into account the social rewards of BCI behavior that are possibly derived from those on the same side. Also, as the desire for social status often motivates a perpetrator to engage in bullying (Salmivalli, 2014), this desire may also affect bystanders in support of the perpetrator. Unless such expected benefits outweigh the costs, bystanders might not take actions at all because they do not want to get themselves in trouble (Hassan et al., 2018).

Existing evidence corroborates the view that bystanders act in a self-interested way (Wang, 2021) by estimating the comparative merits and demerits of alternative behavioral options. From this perspective, the third set of research questions examines the potential

costs and benefits that bystanders may consider when deciding to counter-intervene in online incivility situations.

RQ3a-b. What (a) benefits and (b) costs do counter-intervening bystanders perceive from their behavior in support of online incivility perpetrators?

Pro-perpetrator Attitudes and Pro-incivility Social Norms

Again, the decision-making behind bystander behavior starts with an evaluation of the incident (Latané & Darley, 1970). This decision is often made under substantial ambiguity (Latané & Darley, 1968; Latané & Rodin, 1969), however, and in efforts to reduce such uncertainty, bystanders exploit any available social contextual cues that can help disambiguate the situation and inform what should be done (Hudson & Bruckman, 2004). When bystanders identify with either the perpetrator or the target based on their shared social identity or perceived psychological distance, this may provide a potent social context for bystander behavior. To explain this, Levine and his colleagues (2002, 2013) locate bystander research within self-categorization theory (Turner et al., 1987), which holds that the way in which bystanders categorize themselves in a particular group context determines their behavior. This may occur under specific conditions where one defines oneself in terms of stereotypical group characteristics (rather than in terms of unique personal characteristics) and thereby recognizes the self-other relationship in terms of a shared group membership (i.e., social identification), then senses of belonging (i.e., we-ness; Gaertner & Dovidio, 2000) and merged identities (i.e., one-ness; Cialdini et al., 1997) emerge. These entail in-group biased attitudes and constitute the psychological basis for human behavior.

Bystander research has yielded evidence supporting this approach, such that bystanders show prosocial behavior for their own social group (Panumaporn et al., 2020) specifically when they feel a greater sense of community and share similar attitudes with the target (Banyard, 2008; Bennett et al., 2016). Intervention is more likely to occur when the target is socially connected to bystanders (Palasinski, 2012) and considered worthy of support (Burn, 2009), or when the perpetrator has a negative or distant relationship with bystanders (Song & Oh, 2017, 2018). Given the findings that online community members who engage in trolling hold a strong sense of being like-minded (Gaudette et al., 2021), and that uncivil online comments receive more audience recommendations when they include partisan language that reflects commitments to a particular political group (Muddiman & Stroud, 2017), BCI behavior may be more likely among bystanders who feel strong empathy or positive attitudes toward the perpetrator based on a sense of group belonging (Pepler et al., 2021). Such attitudinal bias may encourage bystanders to act to support the perpetrator's behavior that aligns with their own belief or the collective interest of their group (Leidner et al., 2010). This factor may also reduce the psychological burden that bystanders might feel, such as regret, guilt, and stress, resulting from any adverse outcomes of participating in uncivil behavior (El Zein et al., 2019). The current research examines whether BCI behavior is driven by bystanders' favorable thoughts and feelings about the perpetrator and what factors underlie those positive attitudes.

Bystanders are also likely to rely on the reactions of other bystanders when defining an ambiguous situation (Latané & Darley, 1970). These social cues serve as quick decision rules for intervention (Schwartz, 1977) showing the high or low agreement in the bystander group that the incident is appropriate to intervene (Shotland & Huston, 1979). That being so,

bystander behavior is largely subject to social influences (Latané & Darley, 1968; Latané & Nida, 1981). This echoes the assumptions of theories on social norms. For example, according to social cognitive theory (Bandura, 1986), human behavior is influenced not only by personal-psychological factors (e.g., moral disengagement, self-efficacy) but also by observed others' behaviors and the outcomes of those behaviors. Relatedly, social norm theory (Cialdini & Goldstein, 2004) posits that observing others' prosocial behavior enhances an individual's likelihood of exhibiting the same behavior (Nook et al., 2016). Bystander research, which positions itself within these theoretical perspectives, indicates that social influences deriving from perceived descriptive and injunctive norms regarding intervention serve as the prevailing determinants of prosocial bystander behavior (Allison & Bussey, 2016; Hayashi & Tahmasbi, 2021; Nickerson et al., 2022). Those who believe intervention or non-intervention is the salient norm tend to reflect these behaviors to avoid the social risks associated with defying the norm (Panumaporn et al., 2020).

Research shows that injunctive norms in favor of bullying impose social pressure on bystanders to join in it (Bastiaensens et al., 2016), and that predictors of pro-bullying bystander behavior become more effective when bystanders perceive this behavior as normative for their peer group (Troop-Gordon et al., 2019). Similarly, bystanders with a higher level of personal acceptance of cyberbullying report greater intentions to reinforce the perpetrator when pro-cyberbullying social norms are also at a high level (Dang & Liu, 2020). Consistent with these findings, the current dissertation considers normative influences on BCI behavior that may be facilitated when perceived social climate is biased in favor of incivility perpetrators. In this research context, a greater number of likes (vs.

dislikes) on uncivil user comments can decrease perceived deviance of incivility and increase perceived social support for the comments (Kim & Gonzales, 2022).

Individual attitudes and social norms are key social psychological concepts that have often been considered as independent determinants of human behavior. The final set of research questions is thus to examine whether bystanders' own attitudes and perceived social norms toward online incivility perpetrators render them to engage in BCI behavior.

RQ4a-b. Do counter-intervening bystanders (a) have positive attitudes toward online incivility perpetrators and (b) perceive social norms that align with their attitudes?

Chapter III: Interview Method

A qualitative interview study was conducted to address the exploratory research questions stated above, which were aimed at generating preliminary insights into potential facilitators of BCI behavior. This study consisted of a series of one-on-one interviews with a small sample of people who have lived experiences of reacting favorably to uncivil online discussions. Findings were grouped around the variables that were expected to constitute the new BCI model.

Participants

To conduct interviews with individuals who have experiences counter-intervening in uncivil online discussions, the interviewee recruitment was initiated on *Reddit*—an anonymous online platform dedicated to user-generated content and discourse. This platform enables researchers to reach large samples of study participants who are relatively representative of the general adult U.S. population (Shatz, 2017), or target interest groups centered on subReddits that cover an expansive array of topics. For example, political

subReddits are found to attract more trolling comments than subReddits in other categories (e.g., entertainment, religion; Fichman & Sharp, 2020) and thus have been used to investigate uncivil online discussions (e.g., Gruzd et al., 2020; Marchal et al., 2020; Nithyanand et al., 2017). For this interview study, an interviewee recruitment announcement was posted on popular political subReddits (e.g., r/Politics, r/PoliticalDiscussion, r/Liberal, r/Democrats, r/Republican, r/Conservative), news-related subReddits (e.g., r/News, r/Inthenews, r/Discussion), and other subReddits available for study participant recruitment (e.g., r/Takemysurvey, r/Paidstudy, r/Paidresearchstudies). However, the recruitment posts on the political/news subReddits were mostly deleted or censored by the moderators due to rules banning the advertising of surveys and paywalls. As an alternative option, interview invitations were individually sent to 114 users of political/news subReddits who were found to engage in BCI behavior of interest to this study (e.g., posting positive replies to uncivil comments). Eight users responded to the invitation message, but eventually none of them decided to join the interview. Although this was surprising, the Reddit users might have felt uncomfortable to join the interviews due to the concern that their pseudonymity and online persona on Reddit would not be protected (Gagnon, 2013; Shelton et al., 2015).

For this reason, interviewees were recruited via two crowdsourcing websites widely used for research subject recruitment: Amazon Mechanical Turk (MTurk) and Prolific. Subject pools of these websites, which consist of individuals looking to participate in online studies, are appropriate for the present research because they can represent the Internet-using population (Ross et al., 2010) who spend more time online than the general population (Redmiles et al., 2019) and thus may more likely encounter online incivility (Harriman et al., 2020). A screening survey was conducted to ensure in advance that interviewees were fit

for this study based on the criteria set by the researcher. Participants were screened if they matched the inclusion criteria of having ever observed uncivil user comments during their online discussion activities, having ever reacted favorably to any of those comments, and having rarely been a perpetrator of online incivility (see Measures: *Online Incivility Exposure Experience*, *Bystander Counter-intervention Experience*, and *Online Incivility Perpetration Experience*). For the first and second criteria, participants were also asked to provide an open-ended response to describe one specific counter-intervening experience focusing on what kind of uncivil comment(s) they read and how they reacted. This procedure was necessary to confirm that interviewees understood the notion of counter-intervention and had the relevant experience. The reason for using the third criterion was to recruit interviewees who could provide a valid reflection of their pro-incivility behavior from the perspective of “bystanders” who assist it, not that of perpetrators who initiate it.

This purposive sampling resulted in a total of 26 interviewees. 13 interviewees were males (50.0%), 12 were females (46.2%), and 1 identified as non-binary/third gender (3.8%). The average age of interviewees was 34.46 years ($SD = 11.37$, $range = 20-57$). 19 interviewees indicated their ethnicity, education, and employment status (7 interviewees declined to give this information). 13 interviewees were White/Caucasian (50%), 4 were Asian/Pacific Islander (15.4%), and 2 were African-American (7.7%). As for the highest level of education completed, 11 interviewees reported a bachelor’s degree (42.3%), 4 reported a graduate or professional degree (15.4%), 3 reported a high school diploma or GED (11.5%), and 1 reported some college education (3.8%). Finally, as for political orientation, 9 interviewees indicated themselves as conservative/republican (34.6%), 8

indicated liberal/democrat (30.8%), and 7 indicated independent (26.9%). 2 interviewees indicated no particular political tendency (7.7%).

Procedure

The interviewee recruitment post on MTurk and Prolific introduced this study as aiming to understand online discussions based on users' experiences and perspectives in relation to online incivility. This study narrowed the scope of online incivility to focus on user-generated comments, on social media platforms and news websites that enable public discussions (e.g., Twitter, Facebook, Reddit, YouTube), that contain impolite, aggressive, or hateful expressions, especially those expressing disagreement, criticism, refusal, and negative emotions against a person or group holding a particular political tendency or social identity. In the case that "uncivil comments" might not be regarded as a familiar term, the alternative term "toxic comments" was used when communicating with participants throughout this research. The recruitment post informed the definition of "toxic comments" and provided some examples representing different types of such comments (e.g., name-calling, profanity, pejorative speech), based on Coe et al.'s (2014) and Papacharissi's (2004) understanding of online incivility. The post also explained the purpose of the screening survey and the eligibility requirements of participants. Participants were expected to be U.S. adults over 18 years of age and view and/or join online discussions on public affairs (see the Appendix A for the interview participant recruitment post).

Upon reading the recruitment post on MTurk and Prolific, participants were invited to a 3-minute screening survey questionnaire on Qualtrics. Participants were asked to indicate whether they have had experience responding to uncivil user comments in favorable ways (e.g., supporting, reinforcing or assisting the commenters) during their online

discussion activities. Those reporting such experience were asked to briefly describe one specific episode representing the typical situation in which they had engaged in such behavior. They were then given a few more questions about their online incivility-related experience, sociodemographic information, and an email address for researcher contact. Those who completed the screening survey received \$.50 as participation incentive (see the Appendix A for the pre-interview screening survey questionnaire).

For those who were deemed to have counter-intervening experience relevant to the current research, an interview invitation was sent to their email including a consent form. This study employed a written interview method with the use of an open-ended questionnaire to minimize the constraints typical of face-to-face interviews. That is, there was no need for interviewees to schedule a real-time meeting with the researcher. The written interview process allowed interviewees to have sufficient time to recall and reflect on their experiences with uncivil online discussions. Furthermore, the asynchronous setting was intended to encourage a relaxed and non-judgmental atmosphere to induce candid responses from interviewees. The researcher also carefully worded the questionnaire to help interviewees feel free and supported to share their own experiences and beliefs related to counter-intervention.

The interview questionnaire started by asking interviewees to recall and describe their counter-intervening experience that they briefly reported in the screening survey. A series of open-ended questions then followed to explore interviewees' thoughts and feelings about online incivility and their behaviors in response to it (see Measures below for the interview questions).¹ Interviewees were asked to send their answers back in a week. Some

¹ Before beginning the interview, interviewees were informed of the key terms needed to respond to the interview questions. The person who initiated the uncivil comment (in which interviewees

of them received follow-up questions to clarify or elaborate their initial answers. Once the questions were sufficiently covered, interviewees were asked to join a post-interview survey to report their online commenting behaviors and sociodemographic information that were not collected in the screening survey (see the Appendix A for the post-interview survey questionnaire). Those who completed all the interview and survey processes received \$15 as participation incentive.

Measures

Pre-Interview Screening Survey

Three variables were measured to screen participants before the interview, if they reported no experience of observing incivility in online discussions, no experience of counter-intervening, or frequent experiences of committing incivility as initial perpetrators.

Online Incivility Exposure Experience. Screening survey participants were asked how often in a typical week they have observed uncivil user comments made by others that targeted a person or group while viewing and/or joining online discussions about political or social issues. They were asked to indicate the exposure frequency, on a 6-point Likert scale (0 = *never* to 5 = *very often*). Those who answered “*never*” were filtered out during the screening phase.

counter-intervened) was called “the commenter” instead of “the perpetrator,” the person/group who was the focus of the uncivil comment was called “the target,” and other bystanders who reacted to the same uncivil comment were called “other comment-readers” or “others.” As mentioned in the Method section, the term “toxic comments” instead of “uncivil comments” was used in the interview questions. Finally, BCI behavior was instead referred to as “favorable reaction to uncivil comments” that includes any behavioral responses that support, defend, reinforce, or assist the comments. Note that the Method and Results sections use terms that are being consistently used throughout the dissertation: “uncivil comments,” “counter-intervention,” “perpetrator,” and “target.”

Bystander Counter-intervention Experience. As an indicator of BCI behavior, participants were asked how often they reacted favorably to observed uncivil comments. The question suggested some example actions (e.g., “clicking the like/upvote button on a toxic comment,” “posting a comment/reply to agree with, support, or defend a toxic comment,” “sharing a toxic comment with more people”) and also enabled participants to specify other types of BCI behavior they engaged in. They were asked to indicate the frequency of each behavior, on a 6-point scale (0 = *never* to 5 = *very often*). Those who answered “*never*” to all the items did not pass the screening.

An open-ended question followed then, as mentioned earlier, to determine if participants understood the notion of counter-intervention and had the experience relevant to this study. They were asked to recall a specific example of the typical online situation in which they reacted favorably to uncivil comments regarding political or social issues. They were then asked to briefly describe what the uncivil comment(s) said and how they reacted to the comment(s) in 3-5 sentences. These open-ended responses were manually reviewed to find suitable interviewees. Those who described irrelevant experiences were filtered out.

Online Incivility Perpetration/Victimization Experience. Two questions examined whether participants had experience perpetrating and being targeted by online incivility. One question asked how often they have personally made their own uncivil comments against someone else, while the other question asked how often they have personally received uncivil comments directed against themselves. They responded to each question on a 6-point scale (0 = *never* to 5 = *very often*). The former question on perpetration was used as the screening criterion to find interviewees who could better represent the perspective of counter-interveners (who “joined in” incivility) rather than perpetrators (who “initiated”

incivility). Accordingly, those who answered “often” or “very often” to this question were not deemed as suitable interviewees.

Sociodemographic Information. Participants were asked to indicate their age, gender (1 = *male*, 2 = *female*, 3 = *other*, 4 = *prefer not to answer*) and political orientation (1 = *liberal/democrat*, 2 = *conservative/republican*, 3 = *moderate*, 4 = *independent*, 5 = *other*, 6 = *no particular political orientation*, 7 = *prefer not to answer*). This information was used to obtain a balanced pool of interviewees in terms of gender and political orientation.

Interview

At the end of the screening survey, participants were asked to describe one specific episode representing their BCI behavior in uncivil online discussions. The interview started by asking them to recall this episode (e.g., *Describe the toxic comments to which you reacted favorably. What did the comments argue, and what kinds of toxic expressions were used in the comments? In what ways did you support, defend, reinforce, or assist in the comments?*). The questionnaire followed to explore their thoughts and feelings about the recalled experience. The questions were structured around the following topics in line with RQ1-4.

Perceptions about Incivility. Interviewees were asked to freely describe their emotional states before and during their BCI behavior (e.g., *What were the emotions that you felt at the moment you read and reacted to the toxic comments?*). They then described their thoughts and feelings about the use of incivility in online discussions by indicating their agreement with two competing positions: one that suggests justifiable aspects of incivility (e.g., *Some people might think it was okay for the commenter to make toxic language in that specific case. Did you think the same way?*) and the other that highlights

negative aspects of incivility (e.g., *In contrast, others might be worried that the toxic language could harm and hurt the target. Did you think the same way?*). They were also asked about their overall opinion on possible situations where incivility is acceptable or not (e.g., *Do you think there are certain circumstances in which toxic comments can be accepted or tolerated? In contrast, do you think there are certain circumstances in which toxic comments should be moderated or screened?*). These questions were intended to find evidence of desensitization and justification of incivility.

Perceptions about Counter-intervention. Some questions focused on what motivated interviewees to engage in their BCI behavior, especially whether they felt the senses of counter-intervention responsibility and efficacy (e.g., *Did you feel that you would need to react favorably to the toxic comments? How confident did you feel in doing what you did to show your favorable reaction to the toxic comments?*). Interviewees were also asked about their perception of the benefits and costs of BCI behavior (e.g., *What did you expect might be some possible positive and negative consequences of your favorable reaction to the toxic comments?*) and the actual consequences that resulted from the behavior.

Perceptions about Perpetrators. Further questions were asked to unveil interviewees' attitudes toward incivility perpetrators (e.g., *How did you feel about the commenters?*) and explore what factors contributed to those attitudes (e.g., *Did you agree or disagree with what the commenters argued? Did you know who the commenters were?*). Interviewees were also asked to recall the overall audience sentiment around incivility in the discussion, focusing on how others reacted to the focal uncivil comment(s) (e.g., *Did you see anyone else support or confront the toxic comments? Did others' reactions influence the*

way you responded to those comments?). This question was aimed to clarify the normative atmosphere that might affect BCI behavior.

Post-interview Survey

Additional survey questions were used to better contextualize online discussion experiences of interviewees, which were not asked in the pre-interview screening survey. These additional questions were provided once the interview was completed.

Online Discussion Platform. Interviewees were asked which digital platforms they usually use to view and/or join online discussions on public affairs. Several platforms were presented as options (e.g., Twitter, Facebook, Reddit, YouTube; 1 = *Yes, I use this platform*, 0 = *No, I do not use this platform*), and interviewees were also able to specify the platforms if not listed.

Online Comment Reading/Posting. Next, interviewees were asked, during their activities in online discussions, how often they read others' comments and posted their own comments in a typical week (0 = *never* to 5 = *very often*).

Sociodemographic Information. Two sociodemographic questions were added to identify interviewees' ethnic background (1 = *White/Caucasian*, 2 = *African-American*, 3 = *Native-American*, 4 = *Asian-Eastern*, 5 = *Asian-Indian*, 6 = *Hispanic*, 7 = *mixed race*, 8 = *other*, 9 = *prefer not to answer*) and the highest level of education they have completed (1 = *less than high school diploma or GED*, 2 = *high school diploma or GED*, 3 = *some college education*, 4 = *bachelor's degree*, 5 = *graduate or professional degree*, 6 = *other*, 7 = *prefer not to answer*). These questions were intended to be able to describe the baseline characteristics of interviewees.

Analytic Approach

All interview responses were archived in Microsoft Excel spreadsheets prior to analysis. Each interviewee response was given a unique identifier, with interviewees' names removed to protect their privacy. The data were analyzed using a *thematic analysis approach* (Braun & Clarke, 2006), "a method for systematically identifying, organizing, and offering insight into patterns of meaning (themes) across a data set" (Braun & Clarke, 2012, p. 57). This approach offers a highly accessible and flexible form of analysis that helps examine perspectives of different research subjects (King, 2004) and seek confirming or disconfirming evidence to support or negate theoretical ideas being tested in a study (Miles & Huberman, 1994; Strauss & Corbin, 1998). Previous bystander research using interviews has often employed a thematic analysis, for example, to investigate bystanders' perceptions about bullying (Macaulay et al., 2021; Patterson et al., 2016) and sexual harassment (Koelsch et al., 2012; Kruger et al., 2022).

Braun and Clarke (2006) and documented this method as an iterative process involving six phases: (1) familiarizing data; (2) generating initial codes; (3) searching for themes; (4) reviewing potential themes; (5) defining and naming themes; (6) producing the report. To minimize the potential disadvantages of this approach, such as lack of consistency and coherence in developing themes derived from data (Holloway & Todres, 2003), this study followed the step-by-step guidelines for establishing a rigorous thematic analysis.

First, the researcher read all interviewee responses in a repeated and deliberate fashion until becoming familiarized with all aspects of the interview data. In doing so, the researcher noted initial thoughts on latent patterns in the data for coding.

Next, the researcher quoted key portions of each interviewee response, which were relevant to this research, and explored the patterns underlying them, which provided evidence that captured the phenomena of interest. Through this process, initial codes were generated both inductively (driven by the raw interview data) and deductively (referring to the literature used to guide the interview). The codes had hierarchical levels of specificity, with broad higher-order codes providing an overview of possible core themes and detailed lower-order codes enabling fine distinctions to be made both within and between interviewee responses (King, 2004). When new codes were introduced, previously analyzed responses were revisited to identify if that code was overlooked during the initial coding. Using the resultant codes, the data organized into segments of related texts for corroborating the evidence.

To devise an initial set of themes, the coded data segments were examined in terms of whether they represented similar categories. By doing so, the codes were then condensed into several themes and subthemes as they mapped onto the literature concepts and research questions of this study. Some data segments that did not seem to fit into the researcher's initial analytic plan were organized under "miscellaneous" themes, as Braun and Clarke (2006) and King (2004) recommended, to produce a richer description of the overall data.

The validity and reliability of initial themes was then checked by scrutinizing whether each individual theme had enough and coherent data segments to support them. In doing so, some themes collapsed into each other or were broken down into separate subthemes. The refined themes were ensured to become clearly distinct between each other as well as broad enough to link substantial portions of the data together. This phase resulted

in a hierarchically structured set of significant themes that covered all major ideas from interviewees.

The researcher then assessed the scope and content of each individual theme to check whether they were sufficiently comprehensive and clear. The themes were re-organized to better reflect the overall story that unfolded in the entire dataset in relation to the research questions of the study. Once finalized, the themes were given labels using the literature concepts and interviewees' own words.

Once the final themes were established, the researcher articulated the meaning and implication of each theme and listed exemplar quotes. The quotes were selected based on their representativeness and relevance to the theme.

Chapter IV: Interview Results

Descriptive Information of Interview Participants

All 26 interviewees reported that they viewed and/or joined online discussions about political or social issues using the following digital platforms: Reddit ($n = 16$, 61.5%), Twitter ($n = 14$, 53.8%), Facebook ($n = 12$, 46.2%), YouTube ($n = 8$, 30.8%), TikTok ($n = 6$, 23.1%), Instagram ($n = 3$, 11.5%), and other platforms such as Discord, Fark, and Tagged ($n = 3$, 11.5%). On average, interviewees engaged in more frequent reading of others' comments ($M = 4.11$, $SD = .76$) than posting of their own comments ($M = 2.56$, $SD = 1.42$) during their online discussion activities. The raw comparison of self-report measures of incivility exposure, perpetration, and victimization indicated that interviewees observed uncivil user comments ($M = 4.31$, $SD = .62$; 0 = *never* to 5 = *very often*) as bystanders more frequently than they initiated their own uncivil comments ($M = 1.19$, $SD = 1.20$) or were the targets of uncivil comments ($M = 2.42$, $SD = 1.68$). This suggests that interviewees were

appropriate for the study to represent the perspective of bystanders rather than perpetrators and targets of online incivility.

BCI behavior took a variety of forms. All interviewees reported having clicked the like or upvote (e.g., thumbs-up, Smart/Funny) button on uncivil comments. 23 interviewees reported having responded with emojis (e.g., laughing; 88.5%); 19 posted their own comment or reply to show agreement (73.1%); 17 posted their own comment or reply to elaborate on uncivil comments (65.4%); 13 shared uncivil comments with more people (50%); and 14 engaged in other types of behavior such as posting memes, gifs, or pictures to express their positive emotion (53.8%). Liking/upvoting was the most frequently performed BCI behavior ($M = 2.69$, $SD = 1.12$; 0 = *never* to 5 = *very often*), followed by responding with emojis ($M = 2.46$, $SD = 1.42$), posting a comment or reply for elaboration ($M = 1.88$, $SD = 1.82$) and agreement ($M = 1.62$, $SD = 1.84$), sharing with more people ($M = 1.31$, $SD = 1.52$), and other types of BCI behavior ($M = 1.15$, $SD = 1.41$). The full descriptive statistics of interviewees are presented in Table 1 below.

Table 1. Descriptive Statistics of Interview Participants ($N = 26$)

	<i>n</i>	%
Gender		
Male	13	50
Female	12	46.2
Other (e.g., non-binary)	1	3.8
Age ($M_{age} = 34.46$ years, $SD = 11.37$; $Range_{age} = 20-57$ years)		
Ethnicity		
White/Caucasian	13	50
Asian/Pacific Islander	4	15.4
African-American	2	7.7
Prefer not to answer	7	26.9
Education		
High school diploma or GED	3	11.5
Some college education	1	3.8
Bachelor's degree	11	42.3
Graduate or professional degree	4	15.4
Prefer not to answer	7	26.9
Political affiliation		
Conservative/Republican	9	34.6
Liberal/Democrat	8	30.8

Independent	7	26.9
No political affiliation	2	7.7
Platform (used for online discussion activities)		
Reddit	16	61.5
Twitter	14	53.8
Facebook	12	46.2
YouTube	8	30.8
TikTok	6	23.1
Instagram	3	11.5
Other (e.g., Discord, Fark, Tagged)	3	11.5
Bystander Counter-intervention		
Click the like/upvote button	26	100
Respond with emojis (e.g., laughing)	23	88.5
Post a comment/reply for agreement	19	73.1
Post a comment/reply for elaboration	17	65.4
Share with more people	13	50
Other (e.g., posting memes, gifs or pictures)	14	53.8

Recall of Bystander Counter-intervention Experience

Interviewees were asked to recall one specific episode representing their typical BCI behavior as noted above. The most frequently reported BCI behavior, according to this specific recall, was liking/upvoting the uncivil comment ($n = 22$), followed by posting their own comment or reply ($n = 12$), responding to the uncivil comment with emojis and memes ($n = 6$), and sharing the uncivil comment with more people who might react to it in the same way ($n = 4$).

Interviewees reported the platform on which the recalled specific episode occurred. The most frequently reported platform was Facebook ($n = 7$), followed by Twitter ($n = 5$), Reddit ($n = 4$), TikTok ($n = 3$), YouTube ($n = 2$), and Fark.com ($n = 1$). 4 interviewees did not reveal the names of platforms. The root uncivil comments that prompted interviewees' BCI behavior were political or social in nature, which were about: left/right-wing ideology conflicts ($n = 9$), racial conflicts (e.g., white supremacy, black riots; $n = 5$), improper acts of public figures or influencers (e.g., celebrity scandals involving violence or alcohol abuse; $n = 4$), controversial social issues (e.g., abortion, same-sex marriage; $n = 3$), transgender

issues ($n = 2$), evaluations of particular social groups (e.g., boomers, Catholic people; $n = 2$), and social movements ($n = 1$). The recalled uncivil comments contained name-calling, profanity, insults, threats, or pejorative expressions, consistent with the types of incivility commonly identified from content analyses of uncivil comments on social media posts or online news articles (e.g., Coe et al., 2014; Oz et al., 2018).

Qualitative Analyses of Bystander Counter-intervention Facilitators

To sum up, the interview study was conducted with people who engaged in online discussions of political or social topics and had ample experience with uncivil user comments as bystanders rather than perpetrators and targets. All of them had experience showing favorable reactions to those comments by liking/upvoting, commenting/replying, sharing, and using emojis or memes. A major portion of the interview was aimed at identifying commonly experienced psychological factors that seemed to facilitate BCI behavior. The thematic analysis revealed several themes that were aligned with the exploratory research questions of this study, including: desensitization and justification of online incivility (RQ1a-b), responsibility and efficacy for counter-intervention (RQ2a-b), benefits and costs of counter-intervention (RQ3a-b), attitudes and social norms toward incivility (RQ4a-b). In reporting the results, the words and phrases in quotation marks were directly taken from the interviewee responses. Minimum descriptive words or phrases were inserted in brackets to complete interviewees' sentences for readability.

Desensitization to Online Incivility

Investigation of RQ1a ("Do counter-intervening bystanders show emotional desensitization to online incivility?") was done by exploring how interviewees felt about uncivil user comments. There was mixed evidence suggesting that interviewees experienced

emotional desensitization to incivility. On the one hand, interviewees recalled that the uncivil comment (they counter-intervened) contained moderate levels of toxicity (e.g., joke, mocking, sarcasm). They considered it “humorous,” “funny,” “entertaining,” and “comical.” This was the case when interviewees witnessed the comment in online communities where the normative tone was explicitly negative and mean and, thus, fostered an expectation that toxic discussions would occur. Some interviewees suggested that they deemed the comment as “somewhat of an inside joke or less to be taken seriously” in “a toxic place and a sort of free expression safe haven.” A 27-year-old man gave an example of the subReddit “r/wallstreet” as a community in which “everyone there is toxic” and “everyone understands what it is.” This interviewee elaborated that he was *not* always insensitive to online incivility, saying:

“I think that in situations where everyone isn’t on the same page toxic behavior is definitely not OK. For example, picking out certain people to target on open public forums such as Twitter or Facebook really seems negative to me. That’s just online bullying, compared to busting each other’s balls in a friendly sarcastic way.”

However, the majority of interviewees recognized the toxicity of the comment and its impact on the target. They described “the comment used harsher language,” “the [perpetrator] did not abbreviate the profane word,” and “the language may have been a little too intense.” They also acknowledged that the comment might have harmed or bothered the target, saying that “the [target] might feel terrible about themselves” and “the toxic comments further developed to physical threat to [the target]’s well-being.” Moreover, interviewees gave negative assessments to uncivil expressions in the comment separately from their agreement with it, according to a 23-year-old man who said: “When I reacted

favorably to the toxic comment, I did feel bad for a bit because it was rude of him to post that comment.” Thus, no firm evidence was found from the interviews to support the association between incivility desensitization and BCI behavior.

Notably, interviewees were unanimous in expressing their negative feelings about more severe forms of toxicity often found in online discussions, including “racist, sexist, misogynistic, and homophobic attacks,” “language that might trigger mental disorders, self-harm, or suicide,” and “discussion that encourages unlawful practices (e.g., child abuse, theft, murder).” They felt antipathy towards such types of incivility that “go against the standards of morality” and “drive off many of more reasonable members of the site.” A couple of interviewees were concerned about emotional desensitization among younger generations, as one 30-year-man noted:

“Our younger generations are the one seeing [toxic comments]. They see the ones they look up to doing it so it’s now deemed ok and a necessity to do. Is it circumstantial? Yes it is. There are some situations where civility and moral conduct have already been taken out of the equation leaving a hostile environment filled with hatred and words of perversion and vulgarities. If we are going to tolerate the allowance of such [toxic] comments, commentators need to explain as to why giving proper explanation so that younger eyes reading can understand why it was said.”

Justification of Online Incivility

Despite the weak evidence related to desensitization, investigation of RQ1b (“Do counter-intervening bystanders show justification of online incivility?”) revealed that the majority of interviewees tended to justify the perpetrator’s incivility with morally and

socially acceptable rationales. The commonly-cited rationales include the following five subthemes:

Targets Deserving Incivility. As the most often stated rationale, interviewees believed that incivility could be justified when the target’s “bad faith” and “unspeakable acts” was blameworthy and thus deserved to be punished. Interviewees applied this rationale mostly in two cases. First, they justified the uncivil comments that contained “a valid critique of someone who is acting in a problematic manner on a platform” or “posting contentious content.” This view was elaborated upon by a 22-year-old man who deemed the target “worse” than the perpetrator:

“[A toxic comment is] tolerable as long as the target is worse than the comment. The target is more dangerous and harmful to those around them [so] deserves the toxicity. He sought to do more harm to others than the commenter would do to those around them.”

As for another type of “deserving” target, many interviewees mentioned those who are subject to social monitoring and criticism, such as public figures, celebrities, and people “in positions of power or privilege.” Interviewees often applied this rationale to incivility targeting politicians showing “selfish and irresponsible actions” and “insensitivity to public opinion.”

Freedom of Expression. A second rationale used for justifying incivility was based on the free speech principle. Interviewees argued that the perpetrator had the “right to speak their mind” and even it contains toxicity. A 33-year-old man who stood in opposition to censorship practices elaborated on this rationale:

“Toxic comments can be accepted because I think censorship is worse... Even if comments are deemed toxic or problematic, I want individuals to express themselves freely on politics, sports, entertainment, religion or whatever.”

The overall responses concerning this rationale argued that a wide range of freedom of expression should be allowed even when it can come across in toxic ways. However, the findings of RQ1a (desensitization) suggest that interviewees might not extend the range of freedom of expression to extreme cases of incivility. No interviewee applied this rationale to the cases containing pure hate (e.g., racist, sexist, misogyny, homophobia), sexual harassment, or threats or unlawful acts (e.g., abuse, doxing, violence, murder). Only one 34-year-old man strongly claimed that “any comment at all should be allowed because free speech is the most important.”

Effective Discursive Strategy. Interviewees also suggested that incivility might play a useful role as a discursive strategy that could help people engage more in discussions. Consistent with previous research findings that show exposure to incivility often attracts more public attention (e.g., Kwon & Cho, 2016) and elicits better recall of the given discussion issue (Hwang et al., 2016), interviewees also considered incivility warranted if it served to “draw more attention” and “get a point across” among a certain type of people who might otherwise be less engaged in, or even ignore, the discussion. Some interviewees suggested that incivility could be an effective way to persuade others who stand against the perpetrator’s argument. They believed that a negatively intense tone of the comment could “potentially make people reconsider their views” and “open [the blind followers’] eyes.” One 40-year-old woman recalled a discussion in which comments speaking against anti-Asian hate were easily ignored. In such situations, the use of sarcastic and mocking

expressions was something necessary or even inevitable as “a quick way to elicit a response” from those who might be easily influenced by negativity:

“Sometimes presenting your opinion in a toxic manner is the only way to make your point... With reddit, sometimes you can’t state things nicely. You get ignored. Using harsh tones is sometimes the only way to get a point across... It might convince others to change their viewpoints.”

Uncivil comments were also considered an acceptable strategy when they were intended to correct those who were already being toxic or rude to others for no reason. Interviewees said the only way to influence such type of people is “to get down on their level.” Incivility might be the way “they should be approached” as “an eye for an eye and a tooth for a tooth” tactic.

Venting Negative Emotions. As another favorable function of incivility, interviewees suggested that incivility in online discussions could help people relieve negative emotions. The most often cited emotion was “frustration” felt from society, which is “hard to vent anywhere else.” The perpetrator’s uncivil tones and expressions were deemed appropriate in this sense at least as an outlet to allow them to “let off some steam” or “release a valve to deal with the annoyance” on behalf of the whole discussion group. Interviewees felt this venting effect from the comments that expressed anger, for example, against the president doing a poor job or political figures committing illegal acts. Interviewees might likely experience a cathartic emotional release by seeing and joining in the toxicity of the “like-minded” perpetrators.

Well-intentioned Incivility. Finally, interviewees agreed that incivility might be acceptable if it was perpetrated for the good of society. For example, uncivil comments were

considered worthwhile if they served to bring public awareness to important political or social issues. One 31-year-old man highlighted the “truthfulness” contained in toxicity, arguing that the recalled uncivil comment was warranted because it revealed truth to society. Some interviewees found incivility acceptable if it was used to “mobilize some kind of movements to make things better” or “speak up to the voice that isn’t typically heard.” That is, incivility was justified with the intention to claim socio-political transformation or speak up for a vocal minority in society.

Responsibility for Counter-intervention

To examine RQ2a (“Do counter-intervening bystanders feel the need to support online incivility perpetrators?”), the analysis focused on identifying motivational feelings that might underlie BCI behavior. Interviewees revealed a sense of responsibility that led them to act for the perpetrator, as interveners felt personal responsibility to act for the target (Latané & Darley, 1970). First of all, interviewees said they felt a need to support the perpetrator who shared a similar belief or opinion. This feeling more strongly motivated interviewees when they thought such shared belief or opinion “should have more weight in the world” and “deserve as much support as they can get in any way possible.” BCI behavior was a way to meet this goal according to one 20-year-old woman saying, “I thought more people should know how many people were thinking the same way as the commenter and myself.” Other interviewees described this feeling as a sort of “duty” or “calling” to “help make a difference” in society. This stemmed from the belief that their BCI behavior was intended to “help spread awareness of truth and knowledge to others.” Such ideas suggest that interviewees considered themselves as holding the moral or intellectual high ground. Indeed, one 25-year-old woman said, “I felt a sort of social responsibility to react favorably,

as I felt morally superior than the subjects of the comment.” These kinds of responses were found from the often-recalled online debate situation between pro-Trump and anti-Trump commenters. One 35-year-old man illustrated the idea of counter-intervention responsibility that operated in political discussions:

“I felt like I needed to give my support for someone who shared similar beliefs as I did. I felt like if I gave support more people wouldn’t be a scared to stand up to these blind Trump supporters. I felt like they are so loud and it was my duty to help spread awareness for people to question President Trump.”

A further finding was that such counter-intervention responsibility could become more pronounced based on other commenters’ reactions. Some interviewees said they would have felt *greater* need to support the uncivil comment if they saw many people dislike or reply to it negatively. The observed presence of the opposite party made interviewees feel the need to “get rid of the dislikes” and increase likes to prove to others that “[the perpetrator is] not alone and their opinion isn’t that unpopular.” One 40-year-old women described this feeling as a desire to show “camaraderie” to the perpetrator. Its underlying motivation seemed to be group-serving, according to what another 25-year-old women said, “If I saw that there were people disagreeing with them, I would have felt an obligation towards ‘my people’ to defend and further attempt to justify our commonly held beliefs.” Given these responses, it was more likely that a sense of counter-intervention responsibility arose when the belief or opinion shared by the perpetrator was incongruent with the overall sentiment of the discussion. A 27-year-old man’s response showed this possibility, suggesting that BCI behavior could be intended as a counter-attack against those that appear to be winning or controlling the discussion.

“Reddit tends to encourage people to pile on to things, so once a comment reaches a certain amount of downvotes I feel like it is doomed. I feel like as a sharer of the opinion it is up to me to support the sentiment and change the direction of the forum’s opinions. I’m a fan of underdogs and I’m way more likely to choose the side of whoever is losing.”

To sum up, the interview data suggest that bystanders would feel greater need to support the perpetrator when the status or position shared between them is threatened by other commenters that have an opposing perspective.

Efficacy for Counter-intervention

In relation to RQ2b (“Do counter-intervening bystanders have confidence in their ability to demonstrate the support for perpetrators?”), interviewees reported they felt highly confident about their behavior. Holding the desire to “stand up for [their] belief” and “share [their] opinion” boosted their counter-intervention efficacy. A 40-year-old woman said, “I was fully confident. I had an opinion and wanted to share it. It’s as simple as that.” Another 53-year-old woman said, “I feel pretty confident in doing what I did to show positive reaction. It was just an upvote, nothing revolutionary.”

However, it appeared throughout the data that the efficacy factor exerted a profound influence on the decision for BCI behavior. A notable finding was that more direct forms of BCI behavior, such as leaving a comment or reply, needed great self-confidence to do so. This was evidenced by a 21-year-old man saying, “I would be more hesitant to do anything other than press ‘like’ on it, because I don’t want to get dragged into an argument by expressing my agreement via something like a reply or retweet.” Relatively, those who chose commenting or replying as a way of BCI behavior indicated a greater sense of self-

competence in their responses, such as “I reacted to the issue that I knew well and accurately” and “I truly believed I was correct about all of the things I was saying.” Some interviewees talked about the emotional and psychological work behind reports of having high efficacy, saying that “my reaction was a byproduct of a lot of time spent on contemplating the issue” or disclosing that they did “fact checking” of the related information they had not been sure of. A 39-year-old man stated that his BCI behavior was based on enough understanding of the situation and those involved in it:

“I feel confident in my behavior when I am very sure about the subject being discussed and I am reasonably familiar with the parties involved in the toxic comment (toxic poster and target). In that case, I can be pretty sure that I’m not inadvertently supporting something I may not actually be on the side of.”

Counter-intervention efficacy also partly resulted from the strong belief that BCI behavior was “something that had to be done” and “nothing to be ashamed of.” A 31-year-old woman said, “I felt quite confident. I didn’t feel ashamed by the endorsement [of the toxic comment]. I did the minimum that I could do given the gravity of the issue.” Some interviewees also hinted that they had increased confidence when they saw others “chime in with support [of the toxic comment] as well.” This allowed the feeling of being validated by those on the same side as well as relieved from the concern about “facing backlash” by those on the opposite side. A 50-year-old woman said, “it is human nature for most to gain confidence when they feel others are behind them or support them.” Such responses were also evoked in the following findings concerning the benefits and costs of counter-intervention and the pro-incivility social norms.

Perceived Benefits of Counter-intervention

Investigation of RQ3a (“What benefits do counter-intervening bystanders perceive from BCI behavior?”) revealed that there were some commonly perceived positive consequences as result of support for incivility. Interviewees mentioned the following benefits and advantages not only for themselves and the perpetrator.

Connectedness to Like-minded People. First, interviewees felt being connected to “like-minded people” who reacted the same way to the uncivil comments. Their BCI behavior enabled a sense of “belongingness” or “social acceptance (vs. alienation) by a group of specific individuals.” This is well illustrated in a 50-year-old woman’s response:

“Seeing others’ toxic comments lets others who agree know they’re not alone in their feelings. It can also give a sense of validation to others who are too afraid to speak up themselves. It may also help similar minded people band together to create change. I could say it shows solidarity with the group the commenter represents.”

In addition to such emotional gains, interviewees further sought to obtain affirmation and reputation from those like-minded people. Through BCI behavior, interviewees anticipated they could gain popularity as made measurable by metrics such as likes/upvotes and followers/friends online. By doing so they experienced “collective support” and “rapport with each other.”

Empowerment for Perpetrators. Interviewees highlighted that their BCI behavior could help the perpetrator to feel empowered in the discussion. This was enabled not only by letting the perpetrator know that “others believe in what he/she was trying to express,” but also by letting those in opposition know “how many people were thinking the same way as the [perpetrator].” Interviewees believed their support for the perpetrator could signal “the

inertness of perceived toxicity,” that is, “the toxic comment might not be that toxic.” One 25-year-old man said:

“Supporting the toxic comment, giving it fame. Giving Fame to the toxic comment basically means so other people can see it and I think it’s a positive consequence because it will expose this comment to other people. [...] I gained the knowledge that I helped a toxic comment grow and strengthen into a viral comment.”

To sum up, interviewees expected that BCI behavior could increase the popularity and visibility of the perpetrator’s shared belief. The resultant benefit they perceived was to spread that shared belief to more people and thereby enhance its influence on the whole discussion.

Correction of Opponents. Finally, interviewees suggested that BCI behavior could help reframe a given debate so that the perpetrator was vindicated. A 39-year-old man anticipated that his support for the perpetrator could create a “chance that the [target] would refrain from posting in bad faith, which would allow the rest of the discussion to proceed in a better manner.” A 56-year-old man said his BCI behavior was intended to correct those who stood up for the target:

“I thought that maybe the people reading the comments would stop and think about what had happened [and] what they were doing was wrong... I felt like the comments that I made [to support the perpetrator] were toward a group that was acting irrationally.”

As other positive consequences of BCI behavior, interviewees mentioned “making people reconsider their views” and “convincing others to change their viewpoints.” In other

words, they viewed BCI behavior as a corrective action that could persuade others not to defend the target.

Perceived Costs of Counter-intervention

Meanwhile, interviewees also raised concerns about the negative consequences of supporting uncivil comments. Investigation of RQ3b (“What costs do counter-intervening bystanders perceive from BCI behavior?”) identified three types of counter-intervention costs that were commonly echoed among interviewees.

Criticism from Opponents. The most often reported concern was receiving “backlash” from those who disagreed with the perpetrator and in turn becoming another target of uncivil opponents. Interviewees mentioned negative evaluations that might follow BCI behavior, such as “being downvoted heavily,” “receiving awful direct messages,” and “losing followers/friends.” A 57-year-old man suggested a more severe type of risk, such as “having my account attacked or doxed by true nefarious actors.” However, such potential costs of BCI behavior might not necessarily be barriers to it if there was a strong antipathy toward the opponents. For example, a 31-year-old man said that he was not influenced by the backlash:

“Sometimes I get downvoted, sometimes I get reported, sometimes people give silly comments as a reply. Often times it’s never enough for me to learn a lesson, because I think my opposition is vacuous.”

Some interviewees further said that the initially received downvotes were eventually evened out by the upvotes they received from “like-minded people.” This led to the conclusion that the results of BCI behavior were good. This was evidenced by a 33-year-old man’s response:

“Usually I’ll get support most of the time. If I get evaluated it’s never negative. It’s usually a laugh or a comment reassuring my position. I’ll get support because my followers nine times out ten agree with me.”

Constraints on Online Activities. As another possible cost of BCI behavior, interviewees raised the possibility that their online activities could be censored and restricted. This concern was cited by those who had experience commenting or replying in support of the perpetrator. They said, “facing direct moderation or suspension,” “getting reported,” and “being blocked or banned” by platform practitioners or users for violating community guidelines. However, most interviewees believed that they had a low chance of being punished in these ways.

Deterioration of Discussion Quality. Finally, some interviewees were concerned that BCI behavior could possibly contribute to amplifying the spread of online incivility. They knew that their favorable reaction to the perpetrator could generate a contagious effect of incivility by causing “a lot more toxic comments and fighting” and “increased harassment and degraded talk.” A 33-year-old man said, “It’s almost like giving everyone a pass to join in and just escalates the situation more and more.” This concern was further echoed by another 39-year-old man who was apprehensive about the likely deterioration of the overall discussion quality due to incivility:

“I sometimes worry that it may lead to the general decline in reasoned discussion on the site... Supporting toxic comments in general may make it impossible to have actual calm discussion going forward.”

Despite these possible risks, throughout the data it was evident that interviewees perceived greater benefits than costs as result of BCI behavior. A 35-year-old man said

“negative consequences didn’t really concern me anymore” because the emotional satisfaction he gained mostly outweighed the risks. This was more the case when interviewees chose liking/upvoting rather than commenting/replying as a “safer” option, as reported by another 39-year-old man:

“Generally, the positive consequences outweigh the negative. Upvoting a toxic post is anonymous. So there are no negative consequences to me personally for doing so. Replying to a toxic post can have the consequence of drawing me into the argument with the troll being targeted.”

Attitudes toward Perpetrators

Investigation of RQ4a (“Do counter-intervening bystanders have positive attitudes toward incivility perpetrators?”) revealed that interviewees gave favorable evaluations of the perpetrator despite incivility. This was largely due to the fact that the perpetrator confirmed the beliefs and values that interviewees held. The analysis yielded the following three subthemes that help explain the contributors to such positive attitudes.

Similarity and Identification. First, interviewees identified with the perpetrator that they agreed with, as one 31-year-old woman said:

“I would say my thoughts about the commenter was more positive because he or she agreed with me and I could resonate with their feelings/emotions... I think I also felt a sense of resonance that came from identification with that person or some similarity I thought I felt with that person in terms of thinking and opinion.”

Some interviewees described that their feeling toward the perpetrator was based on “a sense of comradery” and “psychological closeness.” One 25-year-old woman further said

the shared belief led her to feel like she was in a “quasi-parasocial relationship” with the perpetrator.

Respect and Gratitude. Given that uncivil commenting behavior is possibly confronted or moderated (e.g., reporting, deleting; Goodman, 2013; Naab et al., 2018), interviewees deemed the perpetrator as brave for speaking out and expressed “respect” for them having to withstand criticism. A 25-year-old man said, “I felt that the person who posted the toxic comment took a big risk, and it paid off.” A 31-year-old woman also expressed this appreciation, saying:

“...someone else said it for me, so I don’t have to put in more effort to express my thoughts... To some extent, I think I felt some gratitude. Even though they were using some strong languages, I could understand how upset they would be and even appreciated their bravery in expressing those opinions.”

These interviewees thought that the perpetrator raised a voice on behalf of like-minded people or the group they belonged to. This yielded a feeling of “gratitude” for the perpetrator.

Compliment and Praise. In addition to the perceived bravery, interviewees suggested that there was much to praise and admire in the perpetrator’s speech. They gave high evaluations of the perpetrator’s discursive strategies and perspectives conveyed in the uncivil comments. For example, interviewees used adjectives such as “bold,” “clever,” “wise,” “credible,” and “intellectual” to describe their evaluations of the perpetrator. A 40-year-old woman further expressed satisfaction with the competent and critical aspect of the perpetrator, saying “I thought they were smart to point out the hypocrisy. I was proud of

them for doing so.” These reactions made interviewees conclude that the comments were warranted despite being considered toxic.

Social Norms Regarding Incivility

Finally, the analysis focused on RQ4b (“Do counter-intervening bystanders perceive social norms that align with their positive attitudes toward incivility perpetrators?”). Most interviewees perceived pro-incivility norms on the platforms they used, stating that the overall stance in the discussion toward uncivil comments was consistent with their own reaction to greater or lesser degrees. Some interviewees recalled an unanimously favorable sentiment toward incivility, while others recalled mixed reactions but the dominance of supportive over oppositional voices. A 27-year-old man said, “Most people supported the toxic sentiment of the comments. There were only a few responses pushing back against the general sentiment. Generally, everyone was on the same, toxic, page.” Another 35-year-old man, who recalled a political debate, noted incivility from both supporters and opponents of the perpetrator:

“A lot of the support would just be people agreeing with the original toxic comment. They would add on to the insults and share their experiences with other blind republicans. The confrontations were just name calling. They would just call us stupid liberals.”

Indeed, other interviewees also pointed out that the toxic sentiment in the discussion was further fueled by both groups of reactions to the comment.

Through a follow-up question, interviewees were explicitly asked whether the perceived discussion atmosphere affected the way they reacted to uncivil comments. Some of them agreed with the possibility of such social influence, but others did not acknowledge

their susceptibility to it. The former type of interviewees suggested a greater likelihood of BCI behavior in a less risky and more supportive atmosphere toward incivility, particularly in the presence of others who showed favorable reactions to toxic comments. This allowed the feeling of a “comfortable” environment and a sense of being “in a good company.” A 22-year-old woman testified to this by stating, “Since I saw that the overwhelming majority of people were attacking/insulting the [target], I joined in. I thought that it was low risk for me to react to the toxic comment because most people felt the same way.” This social influence effect was especially pronounced when one felt ambiguous about how to react, as suggested in previous research (Latané & Darley, 1970; Latané & Nida, 1981). To this end, a 23-year-old man said:

“Others did influence me because I did not understand the comment at first but after glancing what they saw, I was able to laugh which made me support the toxic comment.”

Interviewees also noted that a generally unfavorable atmosphere against incivility would make them feel “hesitant” or “discouraged” to engage in BCI behavior. The reasons mostly mentioned for refraining to show their reaction were to avoid “not sticking out like a sore thumb,” “being the odd man out,” and “not cause any controversy.” This led people to “less openly support” the comment (e.g., liking/upvoting or sending a private message). Seeing many dislikes or negative replies to the uncivil comment prompted interviewees to have “second thoughts” about their initial reaction. These interviewees mentioned that in this case they “just scrolled past” or even “downvoted the comment” to “keep the peace” in the discussion.

It is worth noting, however, that some interviewees reported the opposite effect of social influence, such that they became less willing to support the uncivil comment when they saw others react the same way. A 25-year-old man said, “If the majority had liked the toxic comment already, the chances would decrease that I would like that comment,” suggesting a possible bystander effect on BCI behavior. A couple of interviewees further revealed that their behavior run counter to the salient norms in the discussion. A 40-year-old woman said she felt more strongly motivated to advocate the comment when the majority of others had confronted it:

“If the toxic comment already has a decent amount of support, then I don’t feel the need to add on to that. Negative responses though, encourage me to react positively. It’s just to show that the toxic comment has some support.”

Chapter V: Interview Study Discussion

Interpretations

Qualitative data from the current interview study provided preliminary insights into potential facilitators of BCI behavior in the context of online incivility. The data were examined focusing on four sets of exploratory research questions that were posed to identify the candidate variables for predicting BCI behavior. In the following, the interview findings are summarized and then interpreted in terms of how they might inform the development of the new BCI model.

RQ1a-b explored whether counter-interveners are emotionally desensitized to online incivility and cognitively justify it. The interview findings did not provide strong evidence supporting the view that desensitization could explain BCI behavior. Although some interviewees made light of uncivil comments that were meant to be jokes or sarcastic

remarks, most interviewees recognized the negative and possibly harmful aspects of those comments that contained toxicity (e.g., name-calling, insults, profanity). They felt uncomfortable about severe forms of toxic comments (e.g., pure hate speech) that had the intent to victimize or dehumanize targets. Considering that interviewees reported an overall frequent exposure to online incivility ($M = 4.31$, $SD = .62$; 0 = *never* to 5 = *very often*), the current interview finding is somewhat inconsistent with the general belief that people become less emotionally responsive to aggressive behavior as result of habitually observing it, which was found in the research of bullying (Pabian et al., 2016; Rudy & Linz, 2008) and hate speech (Bilewicz & Soral, 2020; SELMA, 2019).

Irrespective of such emotional responses, however, it appeared consistently throughout the data that interviewees tended to justify incivility under certain circumstances. In general, they considered it tolerable and even appropriate for the purpose of correcting “blamable” targets or “irrational” opponents. This is consistent with the previous finding that reinforcer/assistant bystanders use justification strategies for their response to bullying to mitigate negative reactions from others (Hara, 2002).² It is also in line with the evidence that bystanders justify online harassment when it is committed for retributive justice (Blackwell et al., 2018).

Interviewees further deemed uncivil speech as part of freedom of expression or necessary means for obtaining certain legitimate ends of, for example, helping relieve one’s negative emotions (e.g., “reasonable anger”) and contributing to the good of society. Taken together, interviewees were aware of the negative implications of incivility but nevertheless

² Hara’s (2002) study revealed several justification strategies used by bystanders of bullying, including: denial of victim (putting blame on the victim’s personality or behavior), denial of injury (negating harm caused by the bully’s behavior), and denial of responsibility (attributing responsibility to external factors).

acknowledged that it could be functional. They might thus resort to justification strategies to avoid moral guilt or regret due to the support for incivility, as outsiders do (de Oliveira, 2019). These findings led the main survey study to place greater attention on justification processes related to incivility that might significantly affect decision making of counter-interveners.

As having long been a major focus of the traditional bystander literature (e.g., Clark & Bussey, 2020; Hayashi & Tahmasbi, 2020; Obermaier et al., 2016; Wong et al., 2021), bystander responsibility and efficacy were the key components of RQ2a-b. The current research attempted to envisage the opposite of the conventionally understood notion of bystander responsibility and efficacy, which could drive supportive actions toward the perpetrator, not the target, of online incivility. These novel notions were empirically evidenced by the interview findings. The interviewees' statements regarding the motivations underlying their BCI behavior reflected a sense of social duty or obligation to stand up for the belief or opinion they shared with the perpetrator and spread the "truth" or knowledge that they believed was important to others in the discussion. This was the case especially when interviewees perceived their stance as under threat or to be a minority in the discussion. This finding shows that behaviors of not only perpetrators but also reinforcer/assistant bystanders of online incivility are driven by the feeling of a "need" to do something against the challenges to their feelings of superiority (e.g., power, popularity; Compton et al., 2014; Romera et al., 2021; Steer et al., 2020) or to their views of justice and fairness (Erjavec & Kovačič, 2012; König et al., 2010). It suggests that counter-interveners are motivated by relational-interactional concerns or group-based needs for affiliation and dominance as far as they feel socially and psychologically connected with the perpetrator.

Additionally, it was evident that interviewees were also highly confident about their ability to demonstrate their support for the perpetrator. Most interviewees engaged in BCI behavior with competence because they believed they were knowledgeable about the discussion topic and the people involved in the discussion and thus correct about their own response. This finding suggests self-efficacy could be an important determinant of BCI behavior because bystanders may need to overcome obstacles or problems encountered in a possibly ego-threatening situation of online incivility (Bandura, 1997). As interveners and non-interveners show the gap in their ability to empathize with and defend the target (e.g., Chen et al., 2022; Thornberg & Jungert, 2013), self-perceived confidence levels of counter-interveners might have a decisive effect on their behavioral decision. Taken together, this interview study provides substantial qualitative evidence on two concepts, counter-intervention responsibility and efficacy, which might contribute to theorizing BCI behavior in the main survey study.

RQ3a-b was concerned with the anticipated positive and negative consequences of BCI behavior. The interview findings suggest that bystanders tend to engage in BCI behavior when it is expected to result in some benefits and advantages for themselves or others, including: being connected to like-minded people, empowering the perpetrator, and correcting or persuading others who have an opposing perspective. Meanwhile, the potential costs and risks of BCI behavior include: getting into trouble by receiving backlash from opponents, losing friends and followers, being restricted in platform usage, and degrading discussion quality. All these perceived benefits and costs are related with the social interaction dynamics of the bystanders involved.

Although much of bystander research has explored the benefit-cost perceptions of defending the target, many themes found in previous research were consistent within the current study on those of defending the perpetrator. For example, it is known that interveners desire increased social status through their defending behavior and thus expect to gain popularity or power (Pozzoli & Gini, 2013; Spadafora et al., 2020; van der Ploeg et al., 2017). The theme of social status and dominance came up consistently throughout the interview data, as both a benefit and a cost. The desire for social status and dominance is often found among bullying perpetrators (Salmivalli, 2014), and the current study shows that this desire may also be playing a role when it comes to the decision-making process of counter-interveners. Furthermore, the study demonstrated that the expected outcomes of BCI behavior are not only personal but also social. Interviewees weighed how their behavior would influence the perpetrator, target, and other bystanders, or the whole discussion atmosphere, as interveners expect that they can help the target feel better and the incident get resolved (Pöyhönen et al., 2012). All these findings suggest that counter-interveners are cognizant of the possible consequences of their planned behavior and are less likely to act if they anticipate nothing good to follow from the behavior or have conflicting expectations. Thus, the main survey study paid attention to how counter-interveners weigh the benefits and costs of their behavior adaptively.

Limitations

Although the current interview study was performed as a preliminary step toward the main survey study, certain limitations must be noted. First, this is a qualitative study based on the experiences of 26 U.S. adults. Thus, its findings cannot be extrapolated to the general population. All participants were volunteers, so self-selection bias cannot be ruled out. It is

also probable that the themes identified in their qualitative accounts of BCI behavior do not cover all potential attitudes and behaviors of bystanders of online incivility. Nevertheless, the focused inclusion criteria were enough to ensure that interviewees would be qualified to provide valuable insights for understanding the drivers of BCI behavior. In line with the constructivist grounded theory approach (Charmaz, 2006), the interview study does not claim to offer an exact picture but rather an interpretative portrayal of the phenomenon studied. It was explicitly intended to give voice to counter-interveners' own perspectives on how they and their "like-minded" group think and feel about online incivility. Furthermore, the interviews were conducted with people who could effectively represent the perspective of "pure" bystanders by trying to exclude those who had personally perpetrated incivility often in the past. It overcomes the limitation of previous research that has observed bystanders with histories of perpetration (e.g., Williams et al., 2018).

In addition, there were no analyzable data on what actual behavior interviewees performed in online incivility situations. Thus, interviewees might have reported things in order to place themselves in a favorable light instead of giving honest accounts. However, attempts were made to reduce this bias by conducting email-based interviews and emphasizing an anonymous and non-judgmental study setting to interviewees. To better address such possible inaccuracies in self-reports, the main survey study measured the *social desirability bias*—i.e., human tendency to present oneself in correct or ideal ways (Fisher, 1993)—of participants. As some of the previous bystander research did (e.g., Bennett et al., 2016; Hong & Lee, 2022), this factor was used as the statistical control in the subsequent data analysis (see "Survey Method" below).

Despite these limitations, the insights gleaned from the interviews are consistent with some bullying studies that suggest reinforcer/assistant bystanders decide their behavior by going through justification processes (e.g., Hara, 2002), cognitively appraising expected outcomes of the behavior (e.g., Pöyhönen et al., 2012), considering relationships with the perpetrator (e.g., Pepler et al., 2021), and observing the normative response in their peer group (e.g., Troop-Gordon et al., 2019). These interview findings guided the second portion of the dissertation, by informing the design of the main survey study and the interpretation of collected data.

Chapter VI: Tests of the Bystander Behavior Models

Mechanisms of Bystander Counter-intervention

Guided by preliminary insights on BCI behavior from the interview study, I returned to the literature to elucidate the decision-making process behind this behavior and propose the new model. Again, the classic BI model offers a useful theoretical basis for understanding moral and psychological mechanisms that underlie prosocial bystander behavior in a critical situation. This model posits a five-step sequence that bystanders go through to be able to intervene for the target. First, noticing a given incident is a prerequisite to taking action of any kind. The incident must be interpreted as a severe situation requiring action to support the target. Once bystanders perceive this, the next steps are to feel personal responsibility to intervene and feel efficacy to do so. The model assumes that the likelihood of intervention increases if all these steps are completed successfully. This process can also take place in a mediated context, as evidenced by studies using this model for the prediction of bystander intervention in online incivility (e.g., Naab et al., 2018; Watson et al., 2019) and cyberbullying (e.g., Clark & Bussey, 2020; Dillon & Bushman, 2015) situations.

Applying this approach to the prediction of the opposite behavioral outcome—counter-intervention—is a timely and much needed addition to not only bystander scholarship but also online incivility scholarship. Previous research has reported the existence of reinforcer/assistant bystanders in cyberbullying (e.g., Barlińska et al., 2018; DeSmet et al., 2016; Panumaporn et al., 2020), online harassment (e.g., Marwick, 2021), and online celebrity bashing (e.g., Ouvrein et al., 2018) on social media and news media platforms. These studies have attempted to identify the “why” of this behavior, often focusing on bystanders’ evaluations of the incident (e.g., Brody, 2021; Orue et al., 2023), perceptions of social norms (e.g., Bastiaensens et al., 2016; Dang & Liu, 2020), and relationships with perpetrators, targets, and other bystanders (e.g., Bastiaensens et al., 2014; Song & Oh, 2018). Particular attention has also been devoted to bystanders’ individual characteristics, such as their moral disengagement or empathy (e.g., Barlińska et al., 2013; Shen et al., 2022), previous experiences of perpetration and victimization (e.g., Kozubal et al., 2019; Zhao et al., 2023), and demographics (e.g., gender, age; Cao & Lin, 2015; Van Cleemput et al., 2014). However, no research currently exists to integrate these factors and identify the specific process leading to such negative bystander behavior.

Hence, a major contribution of this dissertation is to construct a theoretical model that outlines the steps that bystanders go through to be able to counter-intervene in online incivility situations. This newly proposed BCI model is aimed to present a comprehensive picture of how multiple facilitators of BCI behavior—which are suggested from the interview findings and previous literature—would interplay and sequentially contribute to increased engagement in this behavioral outcome. The BCI model is designed based on the original BI model’s assumptions, as it is hypothesized to consist of the following adapted

steps: being exposed to online incivility; justifying online incivility; feeling responsibility for counter-intervention; and feeling efficacy for counter-intervention. The investigation is focused on determining how each step is associated with BCI behavior as well as incorporates some situational factors that may moderate or mediate those relationships, such as how bystanders assess incivility perpetrators and what consequences are expected to follow from the support for the perpetrators. All these inquiries are integrated to demonstrate a possible path model that predicts BCI behavior in the context of online incivility.

Online Incivility Exposure and Pro-perpetrator Attitudes

The original BI model assumes for its first step that bystanders cannot intervene unless they are aware of a potentially critical situation (Latané & Darley, 1970). For the current research, this means that bystanders first must encounter and recognize incivility in online discussions. Previous research has shown that those who often observe uncivil online comments (Ziegele et al., 2020) and cyberbullying (Dillon & Bushman, 2015) are more likely to confront perpetrators. However, this exposure step might not always predict prosocial bystander behavior, given that a frequency of witnessing online hate is also positively related to being a perpetrator of it (Wachs & Wright, 2018; Wachs et al., 2019). This suggests that a frequency of observing online incivility might affect bystanders' personal involvement in negative or antisocial behavior. Given that a higher frequency of being exposed to cyberbullying, hate speech, or negative emotional SNS posts as a bystander is followed by decreased empathy (Bilewicz & Soral, 2020; Konrath et al., 2011; Pabian et al., 2016) and increased use of aggressive language over time (Ferreira et al., 2021), bystanders who often observe incivility might model the perpetrator's behavior (O'Connell et al., 1999).

These observations prompt the question motivating the current research: When, if ever, does exposure to online incivility lead to counter-intervention rather than intervention? One possibility is that bystanders might hold positive attitudes towards perpetrators. Considering that prosocial bystander behavior is more often found when bystanders feel socially connected to the target (Burn, 2009; Song & Oh, 2017, 2018), the opposite antisocial bystander behavior might be likely when they identify with the perpetrator. This assumption is consistent with the earlier interview finding that counter-interveners see the perpetrator as deserving of respect and gratitude for speaking up on behalf of like-minded people. It is thus possible to speculate that online incivility exposure is positively associated with engagement in BCI behavior, especially when perpetrators are highly evaluated. On this ground, the first set of hypotheses are proposed:

H1a-b. Bystanders who are more exposed to online incivility will show (a) greater engagement in counter-intervention, and (b) this relationship will be stronger among those who have more positive attitudes toward perpetrators of the observed incivility.

Justification of Incivility and Benefit-Cost Perception of Counter-intervention

As the findings of the interview study showed, counter-interveners might also engage in cognitive efforts to justify the acceptability of uncivil online discussions. Interviewees revealed that they enacted various justification strategies regarding the observed incivility (e.g., “Targets deserve it,” “Freedom of speech should be protected”), which is consistent with perpetrators’ own moral disengagement mechanisms (Gini et al., 2015; Romera et al., 2021), neutralization techniques (Zhang & Leidner, 2018) or framing of their behavior as a collective “normal” action (Thornberg, 2015). Bystanders may evaluate

the observed incivility as righteous or upright because they consider it to derive from shared outrage that is morally justifiable in the specific context. Much of moral disengagement work notes that perpetrators build narratives of their transgressive conduct as a retaliatory action or a lesson for correcting targets who deserve it (e.g., Konnikova, 2015). For example, online bullying and harassment is often framed as a retributive measure to try to reestablish justice (Gollwitzer, 2009). Typically, harassment occurring in an online social network starts when a perpetrator triggers moral outrage against a target and creates justification for harassment. Then, members of the network, as bystanders, amplify the harassment by reinforcing the justification (Marwick, 2021). This is consistent with the finding that bystanders who assist the perpetrator rather than the target show higher cyberbullying justification (Orue et al., 2023).

However, justification of incivility might more likely motivate BCI behavior when the perceived benefits of this bystander behavior outweigh the perceived costs. Traditional bullying research has shown that perpetrators justify their transgressive conduct mainly from an egocentric point of view, by focusing on the positive consequences and personal advantages they might attain from the situation (Menesini et al., 2003). It is also known that interveners seek social status through their defending behavior and thus consider potential benefits (e.g., gaining popularity or power) and costs (e.g., becoming less liked, getting into trouble) that can be obtained from the situation (Pozzoli & Gini, 2013; Spadafora et al., 2020; van der Ploeg et al., 2017). Similar desires may also affect counter-interveners. For example, if a perpetrator is considered popular and powerful, bystanders expect their status to be raised by endorsing and joining the perpetrator, and this expected personal benefit may thus increase BCI behavior (Pöyhönen et al., 2012). This is reflected in the previous finding

that cyberbullying bystanders were prone to reinforcing the perpetrator if they held positive beliefs about their response and expected to be rewarded with an elevated social status and extended social network (DeSmet et al., 2016). Furthermore, in the earlier interview study, consistent evidence was found suggesting that BCI behavior might more likely occur when bystanders perceive the potential risks to be outweighed by the potential benefits—for their own (e.g., feeling connected with like-minded people) and others (e.g., empowering the perpetrator, correcting the target). For the majority interviewees, it was evident that their belief that the observed uncivil comments were “justifiable” contributed to the perception of greater benefits from the support for those comments. However, if bystanders have conflicting expectations on their planned behavior, they are less likely to act (Pöyhönen et al., 2012). Given these considerations, the second set of hypotheses posit that bystanders who justify incivility will decide to counter-intervene through benefit-cost calculations regarding their behavior.

H2a-b. Bystanders who more strongly justify online incivility will show (a) greater engagement in counter-intervention, and (b) this relationship is mediated by the perception of greater benefits than costs of counter-intervention.

Responsibility for Counter-intervention and Pro-incivility Social Norms

Although no direct evidence exists in the bystander literature to corroborate on a new notion of counter-intervention responsibility, the earlier interview findings suggest bystanders might be driven by a sort of moral imperative or social obligation that they should protect their belief or identity shared with the perpetrator. In some cases, individuals who experience isolation due to personal or societal deprivation experiences (e.g., abuse history, lawbreaking) are likely to embrace deviant groups in society with the hope of

fulfilling the needs for social group affiliation and membership (Curtis & Curtis, 1993).

These types of individuals are more likely to support hate groups and show a greater willingness to become involved with such groups (Turpin-Petrosino, 2002). These observations suggest that BCI behavior may more likely occur when bystanders have a sense of loyalty or obligation to the group represented by the perpetrator.

In addition to the sense of counter-intervention responsibility, related studies suggest that bystanders' responses are further guided by varying levels of perceived social norms, such that those who believe the prevailing norm is to reinforce or assist the perpetrator might be more likely to reflect the same behavior (e.g., Bastiaensens et al., 2016; Dang & Liu, 2020; Troop-Gordon et al., 2019). This is evidenced by the earlier interview findings that showed that counter-interveners are responsive to social cues that inform how other bystanders react to uncivil comments, which suggests normative social influence effects on negative bystander behavior (e.g., Nickerson et al., 2022; Panumaporn et al., 2020). It is evident from the interview responses that counter-intervention responsibility could become more pronounced when the majority disapproves of the perpetrator. This is possible because the presence of such opponents might more strongly evoke feelings of obligation to protect and reinforce the shared "we-ness" between bystanders and the perpetrator.

In summary, the interview findings support the idea that pro-incivility social norms are an important consideration for determining BCI behavior. However, it is presently untenable to specifically predict how those norms would affect bystanders who feel counter-intervention responsibility. Thus, the third set of hypotheses are posed without a direction for the relationship:

H3a-b. Bystanders who feel greater counter-intervention responsibility will show (a) greater engagement in counter-intervention, and (b) this relationship will vary with the perceived pro-incivility social norms.

Self-Efficacy for Counter-intervention

The final potential factor investigated in the current research is counter-intervention efficacy—a bystander’s belief in their ability to effectively engage in BCI behavior. Previous research has shown that a perpetrator with greater efficacy towards antisocial behavior may be more likely to engage in online and offline bullying (Barchia & Bussey, 2011; Bussey et al., 2015). The present dissertation extends this observation to bystanders who feel greater efficacy to reinforce and assist the perpetrator. Along with counter-intervention responsibility, the interview study supports this new notion of bystander efficacy. Its findings suggest that BCI behavior would more likely occur when bystanders judge themselves as able to defend the perpetrator. Although to my knowledge there is no evidence directly demonstrating this notion of efficacy, findings that previous experience of cyberbullying perpetration increases the chance of negative bystander behavior (Barlińska et al., 2013; Bastiaensens et al., 2016; Zhao et al., 2023) lend support for the idea that a belief in one’s capacity for counter-intervention may predict such behaviors. Thus, the final hypothesis is intended to find out how, if any, counter-intervention efficacy works on bystanders in online incivility situations.

H4. Bystanders who feel greater counter-intervention efficacy will show greater engagement in counter-intervention.

A Multi-Step Path to Bystander Counter-intervention

Latané and Darley (1970) described the original BI model as a sequence of decisions that are made after noticing a given incident. Applying this assumption, the current research tests whether the new BCI model also represents a multi-step path that is comprised of the potential counter-intervention facilitators, articulated in H1-4, as the steps that increase the chance to achieve BCI behavior. The sequence of these steps is theoretically hypothesized based on that of the original BI model. The central addition of the current research is to involve some moderators in the proposed path to uncover the specific conditions that must be met for BCI behavior to occur. This will identify how the mechanisms of BCI behavior are contingent upon situational cues, such as perpetrator attributes and social norms in a given incivility incident. Putting together all the relationships stated in H1-4, this research assumes that BCI behavior occurs when bystanders go through the following step-by-step process (see Figure 1 below).

The first step of the model is to observe an online incivility incident—an uncivil user comment in the current research context—that attracts their attention. Situational ambiguity can pose a challenge to this initial step (Latané & Darley, 1968; Latané & Rodin, 1969) especially in an anonymous online setting (Holfeld, 2014; Smith, 2012). To reduce ambiguity, bystanders may focus their attention on the incident’s characteristics (Dovidio et al., 2006; Fischer et al., 2006)—in the current context, any relevant message cues (e.g., what is said) and source cues (e.g., who says it) that can be a basis for judgment of the uncivil comment. In doing so, bystanders may selectively concentrate on whether they agree and identify with the commenter—i.e., perpetrator (Macklem & Macklem, 2003; Warren, 2011).

This will then shape their attitudes toward the perpetrator, which may affect the way this first step leads to the next step.

As a result of observing what is said in the comment and by whom, bystanders may then determine whether the incivility incident can be perceived as justifiable and worthy of support. The current research posits that the second step of justifying the observed incivility can be further facilitated by bystanders' positive attitudes toward the perpetrator. For example, if bystanders empathize with the perpetrator's feelings of moral outrage (Marwick, 2021) and desire for retributive justice (Blackwell et al., 2018), this justification step may more likely take place. Supporting this reasoning, research has found that a close relationship with the perpetrator (e.g., friendship) could motivate bystanders to justify bullying (Forsberg et al., 2014). If this is applied to the present context, bystanders who feel social similarity and psychological connection with the perpetrator may more likely rationalize the perpetrator's uncivil act in order to approach the situation in a manner that protects and maintains the in-group represented by the perpetrator. Based on this reasoning, switching from the exposure step to the justification step is more likely to occur especially when bystanders have more favorable attitudes toward the perpetrator. The inclusion of this moderating attitudinal factor is unique to the new BCI model, not to the original BI model.

Those who undergo the justification step may then interpret the observed incivility as one requiring support. As for the next step, this research assumes that bystanders may anticipate and weigh the potential consequences of that support. This is in line with H2b, which predicts that justification would be positively related to BCI behavior through the benefit–cost perception of that behavior. Even if the observed incivility is directed toward a 'deserving' target and thus viewed as justifiable, bystanders may decide to support it only

when they hold positive beliefs about their choice of response (Jeyagobi et al., 2022; Pöyhönen et al., 2012). As the interview study suggests, bystanders may feel less motivated to enact BCI behavior if they do not succeed in resolving conflicting outcome expectations (e.g., support from proponents vs. criticism from opponents). Thus, the current research assumes that the third step involves a cognitive appraisal of the benefit-cost ratio of BCI behavior. This step also highlights where the flow of the BCI model is conceived somewhat differently from that of the original BI model. It is intended to uncover how transition occurs from the previous step (interpreting incivility as worthy of support) to the next step (deciding whether a bystander needs to be personally involved).

After justifying incivility and identifying greater benefits than costs of supporting it, the next step is accepting personal responsibility for acting accordingly. Again, the bystander literature has highlighted this responsibility factor as the critical point that explains why bystanders do *not* act when they notice an incident that needs intervention (Schroeder et al., 1995). For successful counter-intervention to occur, bystanders must feel a strong need to support the perpetrator and put their thoughts into action. This factor becomes more critical when bystanders see the presence of others (Darley & Latané, 1968). The consensus is that bystander behavior is influenced by what the majority does and approves (Bastiaensens et al., 2016; Dang & Liu, 2020; Troop-Gordon et al., 2019). However, such observation may cause a flaw in reasoning when an individual bystander thinks someone else could potentially act (i.e., diffusion of responsibility). Thus, taking *personal* responsibility for counter-intervention will be a crucial step in the BCI model. The social norm factor is considered to moderate the likelihood of bystanders moving to the last step

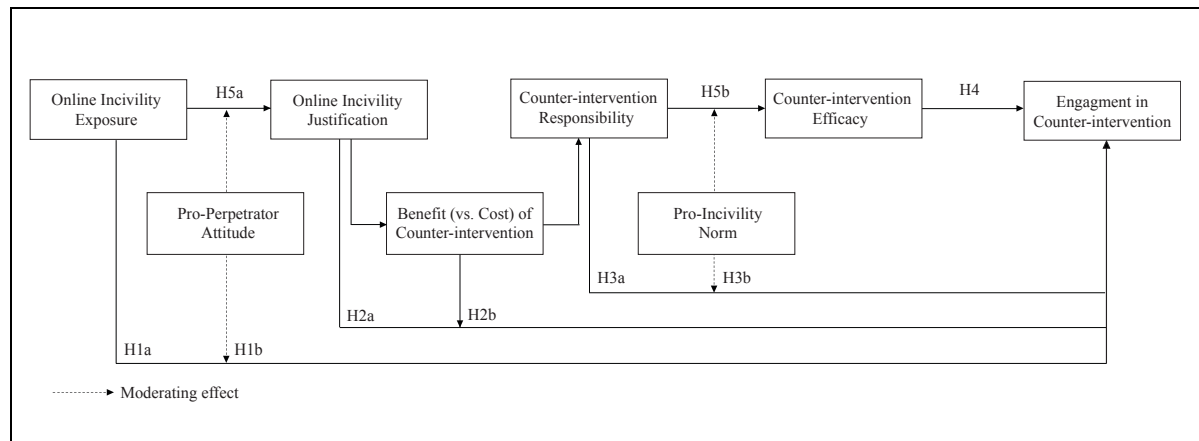
before enacting BCI behavior because it may affect their own perception of whether they need to be personally involved in the given situation and can act upon it with confidence.

Finally, bystanders must assess how much efficacy they have if they want to counter-intervene effectively. Those who go through the responsibility step can have their actions hampered if they lack knowledge and confidence in relation to their desired goal (Banyard, 2008; Burn, 2009). In this sense, social norms aligned with the attitudes of counter-interveners can remove the inhibitions that discourage this particular bystander behavior. Given this reasoning, it is predicted that those who feel that they “should” support the perpetrator are more likely to feel that they “can” do so, especially when they believe that the majority of others accept and approve of incivility. Such pro-incivility social norms may allow bystanders to feel comfortable to decide BCI behavior and thus enact it effectively.

Bringing all these steps together, the current research predicts that the BCI model involves an indirect path that starts from online incivility exposure and leads to BCI behavior. The model assumes six sequential decisions that should be made to incur BCI behavior: Counter-interveners (1) observe an online incivility incident, (2) justify the observed incivility, (3) perceive greater benefits than costs of supporting incivility, (4) accept personal responsibility for counter-intervention, (5) feel self-efficacy for counter-intervention, and (6) implement actions for counter-intervention. These sequential decisions are expected to be influenced by bystanders’ attitudes toward the perpetrator and perceived social norms regarding incivility. That is, this indirect path is hypothesized to occur in two conditions: when bystanders have positive attitudes toward the perpetrator and perceive the social norms align with their own attitudes.

H5a-b. There will be a positive indirect effect of online incivility exposure on counter-intervention via incivility justification, benefit-cost perception of counter-intervention, counter-intervention responsibility, and counter-intervention efficacy among bystanders who (a) hold pro-perpetrator attitudes and (b) perceive pro-incivility social norms.

Figure 1. A Hypothesized Model of Bystander Counter-intervention



Mechanisms of Bystander Intervention and Non-Intervention

Although the primary goal of the current research is to establish the new BCI model, another goal is to test the original BI model (Latané & Darley, 1970) in the context of online incivility. This classic model has been utilized to explain bystander behavior in multiple contexts, including traditional bullying (e.g., Fredrick et al., 2020; Jenkins & Nickerson, 2017; Nickerson et al., 2014), sexual and domestic violence (e.g., Burn, 2009; Ham et al., 2019; Hoefnagels & Zwikker, 2001), and many other situations that require helping behavior (e.g., refugee evacuation; Albayrak-Aydemir & Gleib, 2021). The step-by-step process of the BI model has also been used to explore the mechanisms of bystander behavior in cyberbullying (e.g., Ferreira et al., 2020; Jenkins et al., 2023; Obermaier et al., 2016;

Wang, 2021). These studies collectively suggest that the BI model is a useful theoretical framework for understanding prosocial bystanders both in offline and online settings.

To date, a relatively meager amount of research has adhered to this framework to study bystander behavior in relation to uncivil user comments and online hate speech (e.g., Leonhard et al., 2018; Obermaier et al., 2021, 2023; Ziegele et al., 2020). It is necessary to underscore that the types of social emergencies focused on by Latané and Darley (1970) have some relevant points in common with online incivility situations on public discussion platforms, which is of interest to the current research, including the occurrence of an unexpected uncivil incident perpetrated by a stranger in general, the presence of a person or group targeted by the perpetrator, the fact that the target needs intervention, and the potential risks for bystanders who intervene (Pozzoli & Gini, 2013). Yet, research has not matured to the point of having clear evidence supporting the full five-step sequence of the BI model as it is applied to the incivility context. Indeed, existing studies have applied this model while only focusing on a few particular steps (Leonhard et al., 2018; Obermaier et al., 2021, 2023) without examining the inter-relationships between the steps (Ziegele et al., 2020). Some studies have tested the direct and indirect paths of the BI model to explain bystander behavior in bullying (Ferreira et al., 2020; Jenkins et al., 2023; Nickerson et al., 2014). These studies regarded a low likelihood of intervention as equal to a high likelihood of non-intervention. However, the present dissertation suggests this might not always be the case because bystanders can choose the other behavioral option—counter-intervention.

Filling these gaps, this research applies the sequence of the BI model to the currently investigated online incivility context, in order to more fully validate this framework. In doing so, analyses are separately conducted on the direct and indirect effects of the steps on

bystander intervention behavior (hereafter “BI behavior”) and bystander non-intervention behavior (hereafter “BNI behavior”), instead of using a single behavioral outcome, as indicated in the two diagrams in Figure 2 below. In accordance with Latané and Darley’s (1970) assumptions, the following hypotheses examine whether each step is positively related to BI behavior (H6a-d), and, in another analysis of separate data, is negatively related to BNI behavior (H7a-d). Additional hypotheses examine whether all these variables constitute a positive indirect path leading to BI behavior (H6e) but a negative indirect path leading to BNI behavior (H7e). Taking all these model tests together, the current research aims to elucidate different decision-making paths leading to either prosocial, passive, or antisocial bystander behavior in the yet relatively less understood context of online incivility.

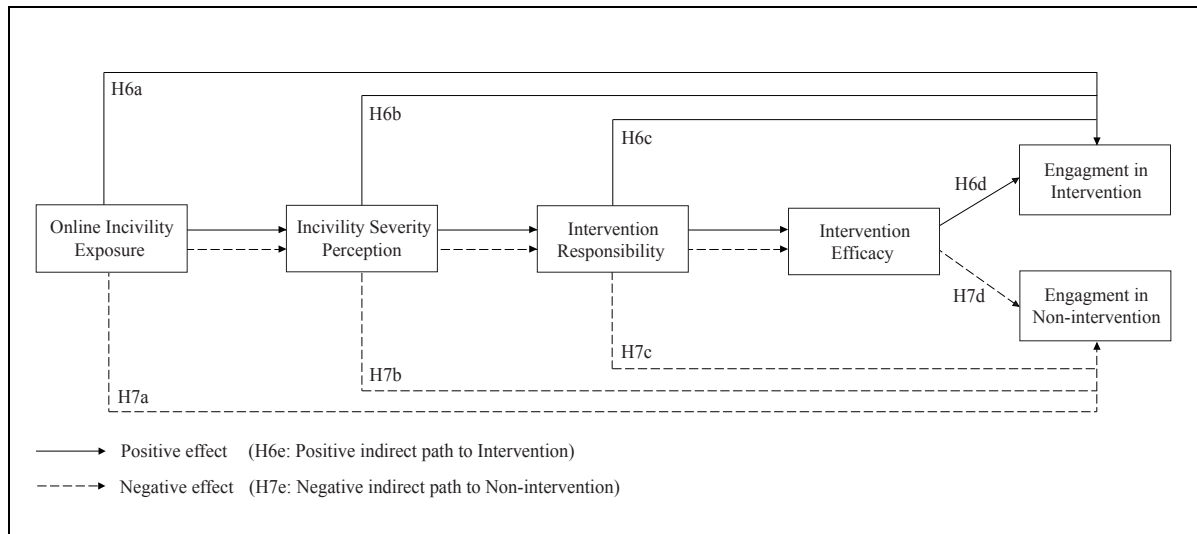
H6a-d. Bystanders will show greater engagement in intervention when they (a) are *more* exposed to online incivility, (b) perceive the observed incivility as *more* severe, (c) feel *greater* intervention responsibility, and (d) feel *greater* intervention efficacy.

H6e. There will be a *positive* indirect effect of online incivility exposure on bystander intervention via higher levels of perceived incivility severity, intervention responsibility, and intervention efficacy.

H7a-d. Bystanders will show greater engagement in non-intervention when they (a) are *less* exposed to online incivility, (b) perceive the observed incivility as *less* severe, (c) feel *less* intervention responsibility, and (d) feel *less* intervention efficacy.

H7e. There will be a negative indirect effect of online incivility exposure on bystander non-intervention via lower levels of perceived incivility severity, intervention responsibility, and intervention efficacy.

Figure 2. Hypothesized Models of Bystander Intervention and Non-Intervention



Summary of Hypotheses

To summarize the goals of the current research, five sets of hypotheses are posited to guide the development of the BCI model as indicated in Figure 1 above. Adapted from Latané and Darley's (1970) BI model that explains a sequential process leading to intervention, the new BCI model aims to illustrate an alternative sequential process that predicts the opposite behavioral outcome. The hypothesized process consists of the following variables that are assumed to have direct associations with BCI behavior: exposure to online incivility (H1a), justification of online incivility (H2a), responsibility for counter-intervention (H3a), and efficacy for counter-intervention (H4). The research also considers additional variables that might be involved in this process. It is further predicted that the exposure step would more likely lead to BCI behavior when bystanders have more positive attitudes toward perpetrators (H1b), and that the responsibility step would more

likely lead to BCI behavior when bystanders perceive more positive social norms toward incivility (H3b). Also, the justification step is assumed to indirectly lead to BCI behavior through a mediating step of perceiving greater benefits than costs of BCI behavior (H2b). Finally, the BCI model is assumed to involve an indirect path that is comprised of all these variables predicting BCI behavior, as the original BI model posits the sequential nature of its steps. This research proposes a path that would start from the exposure step and lead to BCI behavior through intermediate steps of incivility justification, benefit–cost perception, counter-intervention responsibility, and counter-intervention efficacy. Predictions are made that this indirect path would be observed especially among bystanders who have pro-perpetrator attitudes (H5a) and perceive pro-incivility social norms (H5b).

Another goal of the current research is to validate Latané and Darley’s (1970) original BI model in the online incivility context as indicated in Figure 2 above. The related hypotheses are posed to test the BI model’s assumptions by examining whether each step is positively related to intervention (H6a-d), and, in an analysis of separate data, is negatively related to non-intervention (H7a-d), and whether all these variables constitute a positive indirect path leading to intervention (H6e) and a negative indirect path leading to non-intervention (H7e).

Chapter VII: Survey Method

To collect data that could test the proposed BCI model, a series of online surveys were administered via Prolific with a sample of U.S. adults who have experiences of observing uncivil user comments in online discussions of sociopolitical issues. A screening survey was undertaken to recruit participants who would meet the prerequisites for three separate

surveys, each of which was aimed to examine their (a) BCI, (b) BI, and (c) BNI behaviors.³ A total of 460 participants joined the screening survey and then one to three main surveys (BCI survey: $N = 216$; BI survey: $N = 322$; BNI survey: $N = 323$) depending on their self-reported behaviors in online incivility situations. Given that bystanders are likely to play multiple roles (i.e., counter-intervener, intervener, and non-intervener) depending on different situational demands (DeSmet et al., 2014; Sun & Fichman, 2020), the current research conducted the three surveys separately so that participants could report their experiences pertaining to all types of bystander behavior.

Participants and Procedure

Screening Survey

Prolific allows filtering participants by their sociodemographic information collected at sign-up. These prescreening features are useful for the current research to ensure that the surveys could be administered with individuals who are over 18 years of age, located in the U.S., able to answer English-written surveys, and use online media platforms (e.g., Twitter, Facebook, Reddit). To collect quality participants, Prolific users who have an approval rate of at least 90% (determined by their response rate to other previous studies) were allowed to access the current screening survey. Finally, to avoid oversampling or undersampling people from a particular gender group and/or political affiliation, the screening survey sought to balance a sample across these two sociodemographic factors. For this, the screening survey invitation was sent to a total of 460 Prolific users (230 male and 230 female users; 120

³ Again, in the current research context, BCI behavior (bystander counter-intervention behavior) refers to potential behavioral strategies that can provide support for online incivility perpetrators, whereas BI behavior (bystander intervention behavior) refers to the opposite behavioral responses that can be used to defend online incivility targets. As the other behavioral option, BNI behavior (bystander non-intervention behavior) refers to inactions or alternative actions that are biased to neither targets nor perpetrators.

conservative, 120 liberal, 120 moderate, 50 other, and 50 N/A users). Considering the possibility that participants have made changes to their gender and political identity since sign-up, the screening survey asked them to provide their sociodemographic information again.

Potential survey participants were invited to a 5-min survey titled “Online Comment Research.” The recruitment announcement included a paragraph that described the purpose of this screening survey and the notion of online incivility in the same way that interview participants were informed (see the Appendix B for the survey participant recruitment post). Those who accessed a Qualtrics-based survey were asked to sign an online informed consent and then answer a series of questions designed to examine participants’ eligibility for the main surveys as well as collect their sociodemographic information (see the Appendix B for the screening survey questionnaire). Two attention check questions were also embedded within the survey to assess whether or not participants were attentive. At the end of the survey, participants were informed that they would receive an invitation to one to three surveys over several weeks once they were deemed eligible for those follow-up surveys. For those who completed the screening survey, regardless of eligibility, \$.7 was offered according to Prolific’s policy. To assess whether participants were eligible for the main surveys, the screening survey asked about their online discussion activity, online incivility exposure, and online bystander behavior (see Measures below). Four participants who reported having *never* used online discussion platforms, *never* observed uncivil user comments and/or having *never* engaged in any type of bystander behavior were filtered out. No participants failed the attention check questions.

Main Surveys

As a result of the screening survey, a total of 456 participants were deemed eligible for the main surveys, including 269 participants who reported having at least once engaged in BCI behavior; 424 participants who reported having at least once engaged in BI behavior; and 449 participants who reported having at least once engaged in BNI behavior. Among them, 261 participants reported engagement in all three types of bystander behavior; 2 participants reported both BCI and BI; 5 participants reported both BCI and BNI; 158 participants reported both BI and BNI; 1 participant reported BCI only; 4 participants reported BI only; 26 participants reported BNI only. The screening survey generated a list of Prolific IDs of these participants to invite them to the main surveys. For those who were deemed eligible for more than one main survey, each survey invitation was sent sequentially at a one-week interval to help prevent participants from being influenced by their previous responses.

Participants who received the invitation were directed to a 15-min survey titled “Follow-up Online Comment Research.” Upon accessing a Qualtrics-based survey, an online informed consent was taken from participants, and a questionnaire followed to ask their BCI, BI, or BNI behaviors depending on which survey they were directed to (see the Appendix B for each survey questionnaire). At the beginning of each survey, participants were asked to indicate what kind of behavior(s) they have ever engaged in when observing uncivil user comments. They were then asked to recall one typical situation where they engaged in the indicated behavior(s) and describe the situation in four open-ended questions asking: what the uncivil comment(s) said, to whom or on which topic, on which online platform, and what participants did in response to the comment(s). These questions were

also intended to check if participants understood BCI, BI, and/or BNI behaviors correctly. After the recall portion, participants were asked to answer a series of multiple-choice questions that further asked about the situation they described, focusing on the variables of interest in the current research. Two different attention check questions were also embedded within each survey, to ensure that participants were paying attention. After responding to the questionnaire, those who were eligible to participate in more than two main surveys were informed that they would also be invited to another survey in a week. For those who completed the assigned survey(s) and passed the attention check questions, \$2-6 (depending on the number of completed surveys) was offered according to Prolific’s policy.

Among the invited 456 participants, 378 participants responded to the main surveys, including: 209 to the BCI survey; 311 to the BI survey; and 300 to the BNI survey. Specifically, 174 participants responded to all three main surveys; 9 to the BCI and BI surveys; 3 to the BCI and BNI surveys; 122 to the BI and BNI surveys; 30 to the BCI survey only; 17 to the BI survey only; and 23 to the BNI survey only. Six participants who failed the attention check questions or who showed an incorrect understanding of BCI behavior were excluded from the analysis. Thus, 372 participants were included in the data analysis (206 participants in the BCI survey; 311 in the BI survey; 300 in the BNI survey). Table 2 below lists numbers of participants in each survey.

Table 2. Number of Survey Participants

Survey	<i>n</i>
Screening Survey	460
Eligible for the BCI survey	269
Eligible for the BI survey	424
Eligible for the BNI survey	449
Main Survey	378 (BCI <i>N</i> = 209; BI <i>N</i> = 311; BNI <i>N</i> = 300)
BCI + BI + BNI	174
BCI + BI	9
BCI + BNI	3
BI + BNI	122

BCI only	30
BI only	17
BNI only	23

Note. A total of 372 participants were included in the data analysis: BCI survey ($N = 206$); BI survey ($N = 311$); BNI survey ($N = 300$)

Table 3 below shows participants' sociodemographic information. A total of 372 participants consisted of 179 *male* (48.1%), 185 *female* (49.7%), and 8 *other gender-identified* participants (2.2%). The age of participants ranged from 18 to 81 years ($M_{age} = 37.06$ years, $SD = 13.26$), including 9 participants in their *10s* (2.4%), 117 in their *20s* (31.5%), 110 in their *30s* (29.6%), 68 in their *40s* (18.3%), 38 in their *50s* (10.2%), 21 in their *60s* (5.6%), 8 in their *70s* (2.2%), and 1 in their *80s* (.3%). As for ethnicity, 266 participants were *white/Caucasian* (71.5%), 30 were *Asian-Eastern* (8.1%), 29 were *Hispanic* (7.8%), 26 were *African-American* (7.0%), 15 were *mixed race* (4.0%), 3 were *Asian-Indian* (.8%), 2 were *Native-American* (.5%), and 1 did not answer (.3%). As for the education level, 128 participants reported *bachelor's degree* (34.4%), 128 reported *some college education* (34.4%), 62 reported *high school diploma or GED* (16.7%), 47 reported *graduate or professional degree* (12.6%), 5 reported *less than high school diploma or GED* (1.3%), and 2 did not answer (.5%). As for political affiliation, 126 participants indicated themselves as *liberal/democrat* (33.9%), 92 as *conservative/republican* (24.7%), 95 as *moderate* (25.5%), 16 as *other* (e.g., libertarian, socialist, communist, anarchist; 4.3%), and 7 as *independent* (1.9%). 28 participants indicated *no particular political affiliation* (7.5%) while 8 did not answer at all (2.2%).

Table 3. Descriptive Statistics of Survey Participants ($N = 372$)

Gender	<i>n</i>	%
Male	179	48.1
Female	185	49.7
Other (e.g., non-binary)	8	2.2

Age ($M_{age} = 37.06$ years, $SD = 13.26$)		
10s	9	2.4
20s	117	31.5
30s	110	29.6
40s	68	18.3
50s	38	10.2
60s	21	5.6
70s	8	2.2
80s	1	0.3
Ethnicity		
White/Caucasian	266	71.5
African-American	26	7.0
Native-American	2	0.5
Asian-Eastern	30	8.1
Asian-Indian	3	0.8
Hispanic	29	7.8
Mixed race	15	4.0
Prefer not to answer	1	0.3
Education		
Less than high school diploma or GED	5	1.3
High school diploma or GED	62	16.7
Some college education	128	34.4
Bachelor's degree	128	34.4
Graduate or professional degree	47	12.6
Prefer not to answer	2	0.5
Political affiliation		
Liberal/Democrat	126	33.9
Conservative/Republican	92	24.7
Moderate	95	25.5
Independent	7	1.9
Other	16	4.3
No particular political affiliation	28	7.5
Prefer not to answer	8	2.2
Total	372	100
Online discussion activity (1 = <i>never</i> to 7 = <i>everyday</i>)	$M = 4.58$	$SD = 1.31$
Online discussion platform		
Twitter	$n = 198$	53.2%
Facebook	$n = 195$	52.4%
YouTube	$n = 146$	39.2%
Reddit	$n = 122$	32.8%
Instagram	$n = 101$	27.2%
TikTok	$n = 84$	22.6%
News website	$n = 29$	7.8%
Snapchat	$n = 23$	6.2%
Quora	$n = 23$	6.2%
Online community	$n = 16$	4.3%
Blog	$n = 8$	2.2%
Discord	$n = 7$	1.9%
Other	$n = 19$	5.1%
Online incivility experience (1 = <i>never</i> to 7 = <i>always</i>)		
Exposure	$M = 5.15$	$SD = 1.24$
Counter-intervention	$M = 2.20$	$SD = 1.33$
Intervention	$M = 4.08$	$SD = 1.46$
Non-intervention	$M = 4.83$	$SD = 1.35$

Measures

The screening survey measured participants' previous experiences related to online incivility and their background information. Participants were considered unqualified for the main surveys if they reported never observing uncivil online discussions ($n = 4$). In the main surveys, participants were asked to recall one typical situation that could represent their BCI, BI, or BNI behaviors, according to which survey they joined, and then reported what they had thought and felt in that situation. All the questionnaires are presented in the Appendix B.

Online Bystander Behavior

Three types of bystander behavior (BCI, BI, and BNI) were measured by adapting the Participant Role Scale (Salmivalli et al., 1998), an instrument widely used to categorize bystander roles in bullying: reinforcer/assistant, defender, and outsider (e.g., González-Cabrera et al., 2019; Orue et al., 2023; Williams et al., 2018). Items were modified based on interview excerpts to appropriately reflect the online context of interest for this research. Using these measures, participants were asked to report how often they engaged in each type of behavior in uncivil online discussions (1 = *never* to 7 = *always*). Their responses were collected at the stage of screening survey to filter out those who reported “*never*” engaging in any type of behavior. Then, these measures were incorporated into the analyses of main survey data to indicate the outcome variable of each model. They were analyzed as continuous variables so that a higher score of each variable indicated greater engagement in the corresponding type of behavior.

Bystander Counter-intervention. Participants were asked to indicate how frequently they have reacted *favorably* to uncivil comment(s) they observed, despite containing

discursive incivility, by supporting, reinforcing, assisting, or defending the comments (1 = *never* to 7 = *always*). To help participants understand the question, some descriptors of

BCI behavior were provided by adapting the “reinforcer” and “assistant” items of the Participant Role Scale (Salmivalli et al., 1998): “clicked a like or upvote button on toxic comments,” “posted my own comment or reply to support, reinforce, assist, or defend toxic comments,” “posted emojis, gifs, or pictures to support, reinforce, assist, or defend toxic comments,” and “shared toxic comments with others who might also react favorably to them.” 185 participants reported having at least once engaged in liking or upvoting (90%); 70 reported commenting or replying (40%); 39 reported posting positive emojis, gifs, or pictures (19%); and 42 reported sharing (20%).

A content analysis was conducted on the open-ended responses in which participants described a specific situation where they engaged in BCI behavior. Most participants recalled experiences that occurred on Facebook ($n = 54$), followed by Reddit ($n = 45$), Twitter ($n = 44$), YouTube ($n = 11$), TikTok ($n = 11$), Instagram ($n = 7$), news websites ($n = 3$), and other platforms (e.g., Discord, Tumblr, Snapchat; $n = 16$). The most recalled topic of uncivil comments that participants counter-intervened was politics ($n = 94$), followed by identities (e.g., race/ethnicity, gender, religion, immigrants; $n = 38$), controversial social issues (e.g., abortion, vaccination, shooting; $n = 17$), and celebrities/public figures ($n = 13$). The rest of the uncivil comments were the ones directed at a specific individual (e.g., other commenters; $n = 36$) or general people ($n = 6$). 2 participants did not describe the content of the recalled comments.

Bystander Intervention. Participants were also asked to indicate how frequently they have reacted *unfavorably* to uncivil language and tone used in the observed comments, by

opposing, confronting, correcting, or stopping the comments (1 = *never* to 7 = *always*). The “defender” items of the Participant Role Scale were adapted to provide examples of BI behavior, including: “clicked a dislike or downvote button on toxic comments,” “reported or flagged toxic comments to signal to platform moderators,” “posted my own comment/reply to oppose, confront, correct, or stop toxic comments,” and “posted emojis, gifs or pictures to oppose, confront, correct, or stop toxic comments.” 246 participants reported having at least once engaged in disliking or downvoting (79%); 187 reported commenting or replying (60%); 150 reported reporting or flagging (48%); and 51 reported posting negative emojis, gifs, or pictures (16%). A few participants reported other behavioral options such as sending private messages to the uncivil commenters ($n = 2$; 0.6%), blocking the commenters ($n = 2$; 0.6%), and liking or upvoting others’ comments or replies that were posted to confront the commenters ($n = 1$; 0.3%).

The open-ended responses indicated that Facebook was the platform where most participants recalled their BI behavior ($n = 110$), followed by Reddit ($n = 79$), Twitter ($n = 46$), YouTube ($n = 15$), TikTok ($n = 15$), Instagram ($n = 11$), news websites ($n = 5$), and other platforms (e.g., Discord, Twitch; $n = 18$). The most recalled topic of uncivil comments that participants intervened was identities ($n = 133$), followed by politics ($n = 94$), controversial social issues ($n = 29$), and celebrities/public figures ($n = 6$). The rest of the comments were ones directed at a specific individual ($n = 45$) or general people ($n = 3$).

Bystander Non-intervention. Finally, participants were asked to indicate how frequently they have *remained silent* and *taken no action* (i.e., neither supportive nor oppositional reaction) after observing uncivil comments (1 = *never* to 7 = *always*). This non-intervention approach was described using the adapted “outsider” items of the Participant

Role Scale, such as: “read toxic comments but ignored them,” “looked on others’ responses to toxic comments,” “kept scrolling past toxic comments,” and “left the discussion or platform where toxic comments existed.” 255 participants reported having at least once ignored the uncivil comments (85%), 228 reported looking on others’ responses (76%), 167 reported scrolling (56%), and 110 reported leaving the platform completely (37%).

The open-ended responses indicated that most participants recalled their BNI behavior occurring on Facebook ($n = 112$), Twitter ($n = 61$), Reddit ($n = 58$), TikTok ($n = 17$), YouTube ($n = 13$), Instagram ($n = 13$), news websites ($n = 5$), and other platforms (e.g., StockTwits; $n = 8$). The most recalled topic of uncivil comments that participants non-intervened was politics ($n = 104$), followed by identities ($n = 93$), controversial social issues ($n = 33$), and celebrities/public figures ($n = 15$). The rest were the comments directed at a specified individual ($n = 52$) or general people ($n = 2$). 2 participants did not describe what the recalled comments were about.

Predictors of Bystander Counter-intervention

The following variables were measured to test the BCI model. Each variable was predicted to affect BCI behavior directly or indirectly.

Online Incivility Exposure. Participants were introduced to the definition of online incivility and its examples as interviewees were instructed in the same way. Throughout the whole survey, the alternative term “toxic comments” was used instead of “uncivil comments” with the assumption that this phrase would be more familiar to participants. Participants were then asked how often during their online discussion activities they have observed uncivil user comments made by others that targeted a person or group. They indicated the frequency on a 7-point Likert scale (1 = *never* to 7 = *always*; $M = 5.15$, $SD =$

1.24). One participant who answered “*never*” was filtered out after the screening survey. This variable was then used as the first-step factor of the BCI, BI, and BNI models, to test H1a-b, H5a-b, H6a, H6e, H7a, and H7e.

Online Incivility Justification. Participants were asked to indicate the extent to which they thought the recalled uncivil comments were justifiable. To help participants understand the notion of “justification,” some descriptions were provided by adapting Thornberg and Jungert’s (2013) items about bystanders’ justification of a bullying perpetrator’s behavior, and using common expressions found within interviewees’ lexicon regarding how they justify incivility (e.g., “The target deserved the comment,” “The comment was well-intended to correct or advise the target,” “The comment was well within the bounds of freedom of expression”). Participants responded to the question on a 7-point scale (1 = *not at all justifiable* to 7 = *extremely justifiable*; $M = 5.52$, $SD = 1.35$). This variable was used as the second-step factor of the BCI model to test H2a-b and H5a-b.

Benefit–Cost Perception of Counter-intervention. Participants were asked whether and how they took benefit and cost considerations into account when deciding to enact BCI behavior. Six items were constructed using the interviewees’ perceptions of the benefits and costs of BCI behavior. The benefit score was calculated by averaging responses on three items that measure the perceived positive consequences of BCI behavior (e.g., “I expected to feel connected and supported from like-minded people,” “I expected my reaction to positively contribute to the discussion”). The cost score was likewise obtained by averaging responses on three items that measure the perceived negative consequences of BCI behavior (e.g., “I expected to receive negative evaluations from other comment-readers [e.g., dislikes, harsh replies] and have conflict with them,” “I expected my reaction could be

moderated or my account could be suspended”). Participants responded to the items on a 7-point scale (1 = *not at all applicable* to 7 = *fully applicable*). They tended to perceive greater benefits ($M = 4.56$, $SD = 1.21$) than costs from BCI behavior ($M = 3.16$, $SD = 1.25$). The analysis used the benefit–cost ratio score, which was calculated by subtracting the averaged cost score from the averaged benefit score (i.e., the higher score means the greater benefits compared to costs; $M = 1.39$, $SD = 1.72$). This variable was used as the potential mediating factor linking the justification step to the outcome variable (H2b). It was also analyzed as the third-step factor of the BCI model when the indirect path was tested (H5a-b).

Counter-intervention Responsibility. This sense of bystander responsibility was measured as the extent to which participants felt they needed to support, reinforce, assist, or defend the recalled uncivil comments. To help participants understand this question, some descriptions were provided about this feeling, which were incorporated from the interview data (e.g., “I felt it was up to me to defend or assist the perpetrator against criticism from others,” “I felt I should show the perpetrator was not alone in their opinion”). The response was recorded on a 7-point scale (1 = *not at all needed* to 7 = *extremely needed*; $M = 3.53$, $SD = 1.73$). This variable was used as the fourth-step factor of the BCI model to test H3a-b and H5a-b.

Counter-intervention Efficacy. This sense of bystander efficacy was measured by asking participants to indicate the extent to which they felt they were confidently able to support, reinforce, assist, or defend the recalled uncivil comments. This question also incorporated the phrases that interviewees used to describe this feeling (e.g., “I was confident I knew what to say or do to reinforce and assist the perpetrator,” “I did not feel

reluctant to support the perpetrator despite the potential of backlash”). Participants responded on a 7-point scale (1 = *not at all confident* to 7 = *extremely confident*; $M = 4.80$, $SD = 1.47$). This variable was used as the fifth-step factor of the BCI model to test H4 and H5a-b.

Pro-perpetrator Attitude. This study assessed participants’ attitudes toward perpetrators of the observed incivility by focusing on three parameters: (a) identification, (b) agreement, and (c) sympathy with the perpetrators. Five items were constructed using interviewees’ words (e.g., “I felt good seeing the commenter express the same belief about the target that I share,” “I felt grateful that the commenter said what I wanted to say”). All the items were recorded on a 7-point scale (1 = *not at all applicable* to 7 = *fully applicable*), and the scores were averaged ($M = 5.19$, $SD = 1.04$). This variable was used as the potential moderating factor on the relationship between the exposure step and the behavioral outcome (H1b and H5a).

Pro-incivility Norm. Participants were asked about the perceived social norms regarding incivility, in two dimensions: *descriptive norm* (i.e., what most people do) and *injunctive norm* (i.e., what most people think one ought to do). Three items indicated the norms favoring the perpetrator (i.e., BCI behavior is considered normative in the given situation): “I saw a lot of other comment-readers like or reply favorably to the comment(s),” “I thought other comment-readers would consider the comment(s) acceptable,” and “I thought other comment-readers would *not* approve of me defending the comment(s) [reverse-coded].” All the items were recorded on a 7-point scale (1 = *not at all applicable* to 7 = *fully applicable*), and the scores were averaged ($M = 5.33$, $SD = 1.16$). This variable was

used as the potential moderating factor on the relationship between the responsibility step and the behavioral outcome (H3b and H5b).

Predictors of Bystander Intervention and Non-Intervention

Again, the primary goal of the current research is to establish the new BCI model. However, another goal is to test the Latané and Darley's (1970) BI model in the context of online incivility (see Figure 2 above). To investigate the processes of BI behavior and BNI behavior separately, two surveys—BI survey and BNI survey—were conducted in addition to the BCI survey. Both surveys included the measures of traditionally assumed predictors of bystander behavior. Those measures were incorporated into the BI model and BNI model, respectively, each of which was designed to validate whether Latané and Darley's assumptions could be applied to the context of online incivility. The first-step factor in each model was measured by the variable "Online Incivility Exposure" (see above). This variable was used in the tests of H6a, H6e, H7a, and H7e. The other step factors were measured by the following three variables.

Incivility Severity Perception. Participants were asked to indicate the extent to which they thought the recalled uncivil comments were severe enough to be potentially harmful to the target. To help participants understand this question, some descriptions were given that were adapted from the existing items on bystanders' perceived bullying severity (Nickerson et al., 2014; Obermaier et al., 2016) and the earlier interview excerpts (e.g., "The comment would harm the target," "The comment went too far in blaming the target"). Participants answered the question on a 7-point scale (1 = *not at all severe* to 7 = *extremely severe*; BI survey: $M = 3.95$, $SD = 1.73$; BNI survey: $M = 3.38$, $SD = 1.71$). This variable

was analyzed as the second-step factor of the BI model and BNI model. It was used in the tests of H6b, H6e, H7b, and H7e.

Intervention Responsibility. The sense of intervention responsibility was measured by the extent to which participants felt they needed to oppose, confront, correct, or stop uncivil comments, on a 7-point scale (1 = *not at all needed* to 7 = *extremely needed*). The descriptions of this feeling were provided by adapting the items used in previous bystander research (Greitemeyer et al., 2006; Liu et al., 2021) and the interview excerpts (e.g., “I felt it was my duty to do something against the perpetrator, such as disliking, reporting, and flagging,” “I felt I should help the target feel he/she/they were not alone”; $M = 5.26$, $SD = 1.51$). The question and descriptions were reversed in the BNI survey (e.g., “I felt I did *not* need to defend or assist the target against criticism from others”; $M = 2.94$, $SD = 1.72$). This variable was analyzed as the third-step factor of the BI model and BNI model. It was used to test H6c, H6e, H7c, and H7e.

Intervention Efficacy. This variable was measured by the extent to which participants felt they were confident to oppose, confront, correct, or stop uncivil comments (1 = *not at all confident* to 7 = *extremely confident*). This feeling was also described based on the items from previous bystander research (e.g., DeSmet et al., 2016) and the interview excerpts (e.g., “I was sure I had the skills and knowledge necessary to defend the target,” “I was confident I knew what to say or do to support and assist the target”; $M = 5.21$, $SD = 1.57$). Again, this question, including the descriptions, were tweaked in the BNI survey (e.g., “I felt reluctant to confront the perpetrator because of the potential of backlash”; $M = 3.73$, $SD = 1.85$). This variable was analyzed as the fourth-step factor of the BI model and BNI model, and used to test H6d, H6e, H7d, and H7e.

Screening and Control Measures.

Online Discussion Activity. To screen out participants with no experience with online discussions, participants were asked how often in a typical week they engaged in online discussions on sociopolitical issues by reading others' comments and/or posting their own comments. They responded on a 7-point scale (1 = *never* to 7 = *everyday*). Six participants who answered "*never*" were not invited to the main surveys. To collect descriptive background information about participants, another question was asked for them to indicate the platform(s) they use for those discussion activities. They were provided with some options (e.g., Twitter, Facebook, Reddit) and responded on a dichotomous scale (1 = *Yes, I use this platform*, 0 = *No, I do not use this platform*). They typed the platform they use if the given options did not include it.

Online Incivility Victimization/Perpetration. Previous research suggests that personal experience of bullying perpetration could increase the probability of negative bystander behavior (e.g., Barlińska et al., 2013; Bastiaenssens et al., 2016; Williams et al., 2018). Experience of bullying victimization has shown mixed effects, such that it could increase the probability of positive bystander behavior (Panumaporn et al., 2020; Song & Oh, 2017) as well as negative bystander behavior (Shen et al., 2022; Zhao et al., 2023). Thus, these two variables were separately incorporated as control variables in the analysis. Participants were asked about their personal experiences of (a) receiving uncivil comments from someone (victimization) and (b) initiating uncivil commenting toward someone (perpetration) during their online discussion activities. They responded, on a 4-point scale (1 = *never* to 4 = *several times*; perpetration: $M = 1.65$, $SD = .79$; victimization: $M = 2.35$, SD

= .91), to the questions that were adapted from the Youth Internet Safety Survey (Jones et al., 2013).

Social Desirability Bias. The human tendency to present oneself in correct or ideal ways has been found to distort self-report measures in social science research (Paulhus & Reid, 1991). This *social desirability bias* (Fisher, 1993) has been found to lead participants to underestimate their actual bystander behavior because they might become ego-defensive and feel unwilling to report accurately on sensitive topics like bullying or harassment (Banyard et al., 2019). For this reason, the current research used participants' socially desirable response patterns as a statistical control to avoid possible biases associated with self-reports, as previous studies did (e.g., Bennett et al., 2016; Hong & Lee, 2022). This variable was measured using the Balanced Inventory of Desirable Responding (Hart et al., 2015)—an instrument that measures the self-deception facet of social desirability bias (e.g., “I am a completely rational person,” “I have not always been honest with myself [reverse-coded],” “It’s hard for me to shut off a disturbing thought [reverse-coded]”). Responses to eight items were measured on a 7-point scale (1 = *strongly disagree* to 7 = *strongly agree*; $M = 4.06$, $SD = 1.08$, $\alpha = .80$).

Sociodemographic Information. The following five variables were measured to describe the baseline characteristics of participants: gender, age, ethnicity, education, and political orientation. Considering that gender and age have been found to affect the likelihood of negative or pro-aggressive bystander behavior (e.g., Evans & Smokowski, 2017; Orue et al., 2023) and thus have been often controlled in previous research (e.g., Barlińska et al., 2018; Hayashi & Tahmasbi, 2020), these two variables were also controlled in this study. Gender was dummy coded by using the male group as the referential category.

Ethnicity, education, and political orientation were measured for the purpose of obtaining descriptive background information of participants. Each variable was measured on the same questions used in the interview study.

Analytic Approach

The nature of the hypotheses in the current research is to explain *how* and *when* given effects occur. Specifically, this involves tests of (a) processes by which the effects operate (i.e., a pathway from predictor to mediator to outcome; *mediation*) and (b) boundary conditions of the process (i.e., for whom and in what circumstances the effect appears to be significant or insignificant, strong or weak, or positive or negative; *moderation*). As an analytical approach that can test these types of hypotheses, *conditional process modeling* (Hayes & Preacher, 2013) was adopted for the current research. This method integrates properties of *process modeling* (known as *mediation analysis* or *path analysis*) and *moderation analysis* in a *structural equation modeling* framework. This approach estimates (a) whether there are direct and indirect paths through which a predictor variable affects an outcome variable through one or more mediating variables as well as (b) how those effects depend on (or are *conditional* on) the values(s) of one or more moderating variables. Preacher and colleagues (2007) incorporated the term *conditional* from the principle of *interaction* in multiple regression to make inferences for conditional direct and indirect effects in a manner analogous to inferences about “simple slopes” in moderated regression analysis (Aiken & West, 1991). In other words, *conditional process modeling* is used to identify how paths that constitute direct and indirect effects vary across levels of more than one moderating variable. In a model with two moderators, the linear relationship between a direct/indirect effect and the first moderator is quantified at a given value of the second

moderator (i.e., *conditional moderated mediation*; Hayes, 2018). This approach has been commonly used in communication research (e.g., Antheunis et al., 2010; Clementson, 2019).

The current study employed this method to examine the hypothesized direct and indirect association of constructs in the proposed BCI model. All the analyses were conducted with Mplus structural equation modeling software using maximum likelihood estimation, which simultaneously calculates the coefficients of all the hypothesized paths that constitute the model. For example, an indirect path of “ $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ ” was calculated by multiplying four unstandardized coefficients indicating each segment of the path ($A \rightarrow B$; $B \rightarrow C$; $C \rightarrow D$; $D \rightarrow E$). This procedure was applied to the current study by examining the indirect effect of each path in the tested model (that includes the mediator variables) as the product of the unstandardized coefficient for the corresponding path across levels of the moderator variables. As a well-known strategy for inferences about indirect effects in mediation analyses (Hayes, 2009; MacKinnon et al., 2007; Preacher & Hayes, 2004), a bootstrapping technique was used in this study to generate 95% of bias-corrected confidence intervals (CI) on the basis of 1,000 bootstrap samples. The indirect effect was considered significant if the confidence interval did not contain zero, which means that the estimated effect of the respective path was significantly different from zero (Preacher & Hayes, 2004). While the 95% CI approach is traditionally widespread, some research in social sciences has also used the 90% CI approach when it is an exploratory study design with small sample sizes (e.g., Barbeau et al., 2019; Ferreira et al., 2020). Given that the current study aims for a test of a newly proposed BCI model ($n = 206$) and is thus of a very exploratory nature, the statistical significance of each path in the model was interpreted

based on the standard error (SE) estimates obtained from not only 95% CI but also those from 90% CI.

To test for moderation, the predictor and moderator variables were mean-centered before computing interaction terms. The nature of the moderated relationship was examined using the conventional ‘pick-a-point’ approach (Aiken & West, 1991; Rogosa, 1980) that defines a range of values of a moderator variable as “low,” “moderate,” and “high.” Research using conditional process modeling (e.g., Antheunis et al., 2010) has investigated the point estimates of the indirect effects at three different levels of the moderator variable: 1 SD below the mean, mean, and 1 SD above the mean (Cohen et al., 2003). The same approach was used for the current study.

The model fit was assessed using the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean square residual (SRMR). The model was considered to have a satisfactory fit when the analysis indicated $RMSEA < .08$, $CFI \geq .90$, and $SRMR < .08$ (Holbert & Stephenson, 2007; Hu & Bentler, 1999; Kline, 2015).

Chapter VIII: Survey Results

Descriptive Information of Survey Participants

On average, participants reported a moderate frequency of viewing and/or joining online discussions in a typical week ($M = 4.61$, $SD = 1.25$; 1 = *never* to 7 = *everyday*). When asked to indicate the platform(s) they use for their online discussion activities, 198 participants (53.2% of the whole survey sample) reported their use of Twitter, followed by Facebook ($n = 195$, 52.4%), YouTube ($n = 146$, 39.2%), Reddit ($n = 122$, 32.8%), Instagram ($n = 101$, 27.2%), TikTok ($n = 84$, 22.6%), news websites ($n = 29$, 7.8%), Snapchat ($n = 23$,

6.2%), Quora ($n = 23$, 6.2%), online communities ($n = 16$, 4.3%), blogs ($n = 8$, 2.2%), Discord ($n = 7$, 1.9%), and *other* platforms (e.g., *Twitch*, *Tumblr*, *4chan*; $n = 19$, 5.1%). During these online activities, participants have observed uncivil user comments as bystanders at a frequent level ($M = 5.15$, $SD = 1.24$; 1 = *never* to 7 = *always*) and, as their behavioral response, have engaged in BCI ($M = 2.20$, $SD = 1.33$), BI ($M = 4.08$, $SD = 1.46$), and BNI behaviors ($M = 4.83$, $SD = 1.35$).

Data Analyses

The data analysis process consisted of three parts. First, the BCI survey data were analyzed to test the new BCI model (see Figure 1 above) based on the five sets of hypotheses (see H1-5 above). However, the hypothesized direct and indirect paths in this model did not sufficiently work to predict BCI behavior (see “Test of the Initial Bystander Counter-intervention Model” below). An exploratory post-hoc test was thus conducted on the same survey data to present a more simplified and improved model (see Figure 4 below). This alternative BCI model was guided by research questions (see RQ1-3 below) that were formulated based on the initial model’s results as well as the additional review of the interview data (see “Test of the Alternative Bystander Counter-intervention Model” below). Finally, the BI survey data and the BNI survey data were analyzed to test the applicability of Latané and Darley’s (1970) BI model into the context of online incivility (see Figure 2 above) based on the two sets of hypotheses (see H6-7 above). The analysis was conducted on each dataset to test the hypothesized BI model and BNI model, respectively (see “Test of the Bystander Intervention/Non-intervention Models” below).

In all the analyses conducted, the following five variables were controlled: gender, age, online incivility victimization/perpetration, and social desirability bias. The intercorrelations for the key variables in each survey are presented in Table 4 below.

Table 4. Intercorrelations among the Key Variables

Bystander Counter-intervention Survey

	EXP	JUS	BEN	RES	EFF	ATT	NOR	BCI	VIC	PER	SDB
EXP		-.03	-.04	.03	.09	.07	.07	.22**	.31**	.18**	-.10*
JUS			.23**	.24**	.40**	.59**	.26**	.03	.002	.03	.11
BEN				.17*	.26**	.42**	.45**	-.13	-.21**	-.21**	.11
RES					.26**	.47**	.07	.15*	.01	-.05	.11
EFF						.36**	.32**	.06	.07	.13	.30**
ATT							.43**	-.01	-.08	-.11	.14*
NOR								-.11	-.23**	-.12	.09
BCI									.26**	.50**	-.03
VIC										.45**	.02
PER											-.12*

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; EXP (Online Incivility Exposure); JUS (Justification of Online Incivility); BEN (Benefit–Cost Perception of Counter-intervention); RES (Counter-intervention Responsibility); EFF (Counter-intervention Efficacy); ATT (Pro-Perpetrator Attitude); NOR (Pro-Incivility Norm); BCI (Bystander Counter-intervention); VIC (Online Incivility Victimization); PER (Online Incivility Perpetration); SDB (Social Desirability Bias)

Bystander Intervention Survey

	EXP	SEV	RES	EFF	BI	VIC	PER	SDB
EXP		.11	.16**	.02	.34**	.30**	.17**	-.13*
SEV			.31**	.03	.05	.10	-.07	-.01
RES				.40**	.22**	.14*	-.11	-.002
EFF					.13*	.11	-.01	.20**
BI						.31**	.09	-.05
VIC							.41**	.02
PER								-.09

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; EXP (Online Incivility Exposure); SEV (Incivility Severity Perception); RES (Intervention Responsibility); EFF (Intervention Efficacy); BI (Bystander Intervention); VIC (Online Incivility Victimization); PER (Online Incivility Perpetration); SDB (Social Desirability Bias)

Bystander Non-intervention Survey

	EXP	SEV	RES	EFF	BNI	VIC	PER	SDB
EXP		.08	.10	.04	-.07	.32**	.20**	-.16**
SEV			.51**	.01	-.02	.09	-.11	-.18**
RES				.16**	-.17**	.08	-.05	-.08
EFF					-.18**	.11	.06	.24**
BNI						-.27**	-.16**	-.02
VIC							.46**	-.01
PER								-.09

Note. *** $p < .001$; ** $p < .01$; * $p < .05$; EXP (Online Incivility Exposure); SEV (Incivility Severity Perception); RES (Intervention Responsibility); EFF (Intervention Efficacy); BNI (Bystander Non-intervention); VIC (Online Incivility Victimization); PER (Online Incivility Perpetration); SDB (Social Desirability Bias)

Test of the Initial Bystander Counter-intervention Model

The initially proposed BCI model (see Figure 1 above) was examined using a conditional process modeling approach. First, the analysis estimated the separate direct paths of each step variable to the outcome variable. This included the assessments of whether online incivility exposure (“Exposure”), online incivility justification (“Justification”), counter-intervention responsibility (“BCI Responsibility”) and counter-intervention efficacy (“BCI Efficacy”) would be directly and positively related to the engagement in BCI Behavior. The analysis also tested the moderation effects of bystanders’ positive attitudes toward perpetrators (“Pro-perpetrator Attitude”) and perceived positive social norms toward incivility (“Pro-incivility Norm”) on the relationship of Exposure and BCI Behavior and the relationship of BCI Responsibility and BCI Behavior, respectively. It was also tested whether the relationship of Justification and BCI Behavior would be mediated by the perceived benefits (vs. costs) of counter-intervention (“BCI Benefit–Cost”). Finally, the analysis examined whether the hypothesized indirect path from Exposure to BCI Behavior, through Justification, BCI Benefit–Cost, BCI Responsibility, and BCI Efficacy, would be moderated by Pro-perpetrator Attitude and Pro-incivility Norm.

All these direct and indirect paths were examined in the conditional process modeling analysis conducted on the initially proposed BCI model. The results did not indicate a good model fit ($X^2 = 140.17$, $df = 25$, $p < .001$; CFI = .63; RMSEA = .15 [90% CI = .13 to .17]; SRMR = .10). In a statistical sense, it is not advisable to make meaningful inferences from the results of a poor-fitting model, even when significant paths appear in it (Browne & Cudeck, 1992; West et al., 2012). However, it is also true that a good-fitting model is not necessarily a valid model because all of its estimated parameters are not

significantly different from zero (Kenny, 2015). Given the truth that “no model can be literally true” (Preacher & Yaremych, 2022, p. 207), some researchers attempt to interpret a poor-fitting model with a cautious exploration of the observed relations between the variables that are consistent with previous research, especially when they conduct a purely exploratory search or wish to improve on the poor model fit (e.g., Bard, 2006). Thus, although the overall fit of the initially proposed BCI model was not within the conventional range of acceptance, its findings are still worth consideration given their theoretical origins. Given that some components of this model were associated with BCI behavior, some insights could be gleaned into how to improve the model fit. Thus, the findings regarding the initially proposed BCI model are reported below (see Figure 3 below), followed by a proposed alternative BCI model derived from these findings (see Figure 4 below).

Direct Paths to Counter-intervention

H1a predicted that participants who were more exposed to online incivility would show greater engagement in counter-intervention. H1b assumed this relationship to be more pronounced among those who held positive attitudes toward perpetrators. The results revealed a significant direct relationship between Exposure and BCI Behavior ($b = .15$, $SE = .06$; $\beta = .15$, $p = .01$), supporting H1a. This relationship was not however moderated by Perpetrator Attitude ($b = .06$, $SE = .07$; $\beta = .07$, $p = .34$), rejecting H1b. These findings suggest that participants who experienced more exposure to online incivility engaged in counter-intervention more frequently regardless of their attitudes toward perpetrators.

H2a posited that participants who strongly justified online incivility would show greater engagement in counter-intervention. H2b tested whether this relationship would be mediated by perceived benefits (vs. costs) of counter-intervention. H2a was rejected due to

no relationship between Justification and BCI Behavior ($b = .02$, $SE = .07$; $\beta = .02$, $p = .79$). Justification was positively related to BCI Benefit–Cost ($b = .29$, $SE = .08$; $\beta = .23$, $p = .001$), but BCI Benefit–Cost was not significantly related to BCI Behavior ($b = -.04$, $SE = .05$; $\beta = -.07$, $p = .40$), rejecting H2b. That is, participants perceived greater benefits than costs of supporting incivility when they strongly justified it, but such positive perception did not increase counter-intervention.

H3a assumed that participants who felt greater counter-intervention responsibility would show greater engagement in BCI behavior. H3b posited that this relationship would be moderated by perceived social norms about incivility. BCI Responsibility was found to be positively related to BCI Behavior ($b = .12$, $SE = .05$; $\beta = .19$, $p = .01$), supporting H3a. This relationship was not moderated by Pro-incivility Norm ($b = -.05$, $SE = .04$; $\beta = -.09$, $p = .19$) though, rejecting H3b. These findings indicate that participants who felt greater need to support the perpetrator engaged in that behavior more frequently, but this was not contingent upon the social norms.

Finally, H4 predicted that participants who felt greater counter-intervention efficacy would show greater engagement in that behavior. It was rejected though due to no significant relationship between BCI Efficacy and BCI Behavior ($b = -.04$, $SE = .05$; $\beta = -.06$, $p = .43$).

Indirect Paths to Counter-intervention

H5a-b proposed that the path from Exposure to BCI Behavior would involve a sequential mediation via Justification, BCI Benefit–Cost, BCI Responsibility, and BCI Efficacy. This path was assumed to be contingent upon Pro-perpetrator Attitude and Pro-incivility Norm as moderator variables. The bootstrapping analysis indicated that the

indirect path of Exposure to BCI Behavior was not significant at any of the three levels (*mean-1SD*, *mean*, and *mean+1SD*) of each moderator variable, rejecting H5a-b.

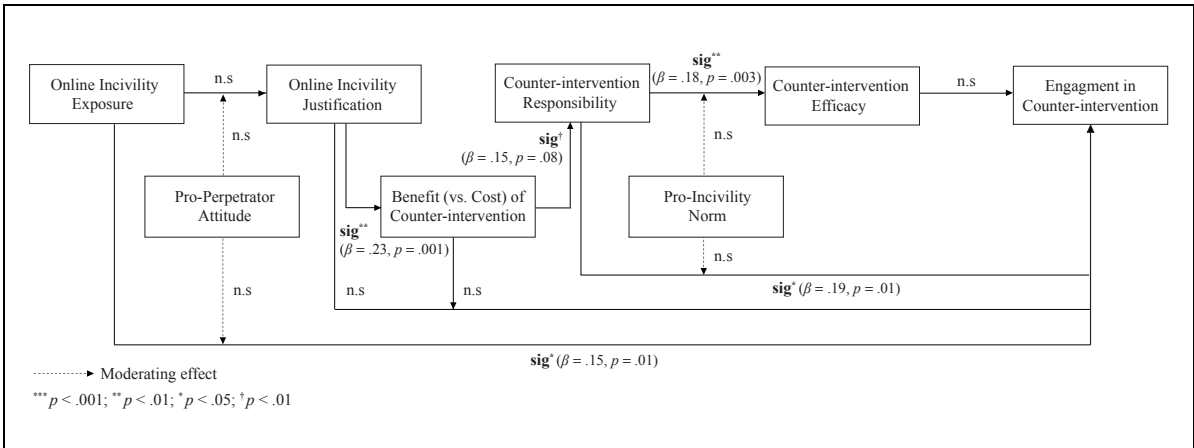
Specifically, the relationship between Exposure and Justification was not significant when it was assessed with Pro-perpetrator Attitude ($b = .08$, $SE = .08$; $\beta = .07$, $p = .32$) as well as without it ($b = -.06$, $SE = .07$; $\beta = -.04$, $p = .45$). Justification was positively related to BCI Benefit–Cost, as reported above, and in turn BCI Benefit–Cost was positively related to BCI Responsibility at a marginally significant level ($b = .15$, $SE = .09$; $\beta = .15$, $p = .08$). BCI Responsibility and BCI Efficacy were positively related to each other ($b = .15$, $SE = .05$; $\beta = .18$, $p = .003$), but this relationship disappeared when it was assessed with Pro-incivility Norm ($b = .03$, $SE = .05$; $\beta = .05$, $p = .52$). As reported above, BCI Efficacy was not significantly related to BCI Behavior.

Summary

The analysis of the initially proposed BCI model evidenced that its model fit was not acceptable although some of the hypothesized paths were significant (see Figure 3 below). The results showed positive relationships of Exposure and BCI Responsibility to BCI Behavior, respectively. These findings suggest that bystanders would more frequently engage in counter-intervention when they experience more exposure to online incivility and when they feel greater need to support the perpetrator personally. There were also positive associations among the intermediate step variables—Justification, BCI Benefit–Cost, BCI Responsibility, and BCI Efficacy—although not directly leading to BCI Behavior. The related findings suggest that bystanders who strongly justify the observed incivility would tend to perceive greater benefits than costs regarding the supportive response to it, and such expected positive consequences of counter-intervention would likely lead bystanders to feel

greater need and confidence to engage in that behavior. No further evidence was found to suggest that Pro-perpetrator Attitude and Pro-incivility-Norm moderate any of the direct and indirect paths to counter-intervention.

Figure 3. Test Results of the Initial Bystander Counter-intervention Model



Note. Gender, age, incivility victimization/perpetration, and social desirability bias were controlled.

Test of the Alternative Bystander Counter-intervention Model

Given that the initially proposed BCI model was not well supported by the survey data, a further effort was made to improve it through an exploratory post-hoc test. Previous studies using Latané and Darley's (1970) BI model have used simpler or alternative models that can better explain and pinpoint factors contributing to a given bystander situation (e.g., Ferreira et al., 2020; Ham et al., 2019; Obermaier et al., 2023). These studies have attempted to identify other possible relationships or paths between the variables of the original BI model or selectively emphasize some variables over others. A similar attempt was made in the current research based on the results of the initial model test and the additional review of qualitative data from the earlier interview study. Resultantly, an alternative BCI model was developed focusing on five variables that provided meaningful information in the initial BCI model. Exposure and BCI Responsibility were incorporated in the alternative model in the same fashion as had previously been used, as they showed positive relationships to BCI

Behavior in the initial model. Justification, BCI Benefit–Cost, and BCI Efficacy were also incorporated but reconsidered in terms of their roles in predicting BCI Behavior. Given that these variables were directly and indirectly related to BCI Responsibility that served as the major predictor of BCI Behavior, according to the initial model test, the alternative model leaves open the possibilities that they might be involved in the partial path that links Exposure to BCI Responsibility. These ideas formulated the following post-hoc research questions with the aim of identifying an alternative but equally plausible pattern of relationships between the variables analyzed.

Joint Effects of Exposure and Justification

The initial model test revealed a positive direct relationship between Exposure and BCI Behavior but no significant relationship between Justification and BCI Behavior. Yet Justification was the much-talked-about factor among interviewees as the dominant mental process behind their reaction to incivility. Justification stimulates moral emotions and cognitive dissonance reactions in relation to the observed incivility by, for example, recognizing it as legitimate or feeling justified outrage (Blackwell et al., 2018; Carlsmith & Darley, 2008). In doing so, bystanders may make a narrative conclusion biased toward BCI behavior. From these conjectures, a speculative but logical possibility is raised that Justification might be involved simultaneously in the Exposure step, as the factor determining bystander' immediate and convincing impression of the given online incivility situation. To test this, the first post-hoc research question was intended to examine whether the relationship between Exposure and BCI Behavior would vary with the levels of Justification:

RQ1. Do bystanders more often exposed to online incivility show increased engagement in counter-intervention when they more strongly justify the observed incivility?

Joint Effects of BCI Benefit–Cost and BCI Efficacy

Also, according to the initial model test on the remaining steps, neither BCI Benefit–Cost nor BCI Efficacy directly led to BCI Behavior but both showed positive relationships with BCI Responsibility. Specifically, BCI Benefit–Cost was found to increase BCI Responsibility at a marginally significant level. This finding tentatively supports the possibility that bystanders feel greater need for counter-intervention when they expect the net positive consequences of that behavior. Through the post-hoc analysis, a further assessment was made to see whether this effect would be enhanced by BCI Efficacy that was found to contribute to BCI Responsibility as well. Again, given that self-efficacy is known as a crucial factor in deciding behavior, especially in an ego-threatening situation that involves obstacles or problems to be overcome (Bandura, 1997), its role still needs to be verified in the current research context. Although the initial model test did not demonstrate the direct impact of self-efficacy, it might nevertheless be the factor that explains the engagement in BCI Behavior, as previous research has consistently highlighted a bystander’s belief in their abilities of putting the intended strategy into action (Gini et al., 2008; Song & Oh, 2017; Van Cleemput et al., 2014). The earlier interview study also suggests that, even though bystanders expect more benefits than costs from their BCI behavior, such outcomes could be achieved only when they feel confident that they could successfully carry out the intended behavior. From synthesizing the qualitative responses about BCI Benefit–Cost and BCI Efficacy, it is reasonable to speculate that the former factor

might largely depend on the latter factor in provoking BCI behavior. In this sense, the second post-hoc research question examined whether BCI Efficacy moderates the relationship between BCI Benefit–Cost and BCI Behavior.

RQ2. Do bystanders who perceive greater benefits (vs. costs) of counter-intervention show increased engagement in counter-intervention when they have greater counter-intervention efficacy?

A Multi-Step Path to Bystander Counter-intervention

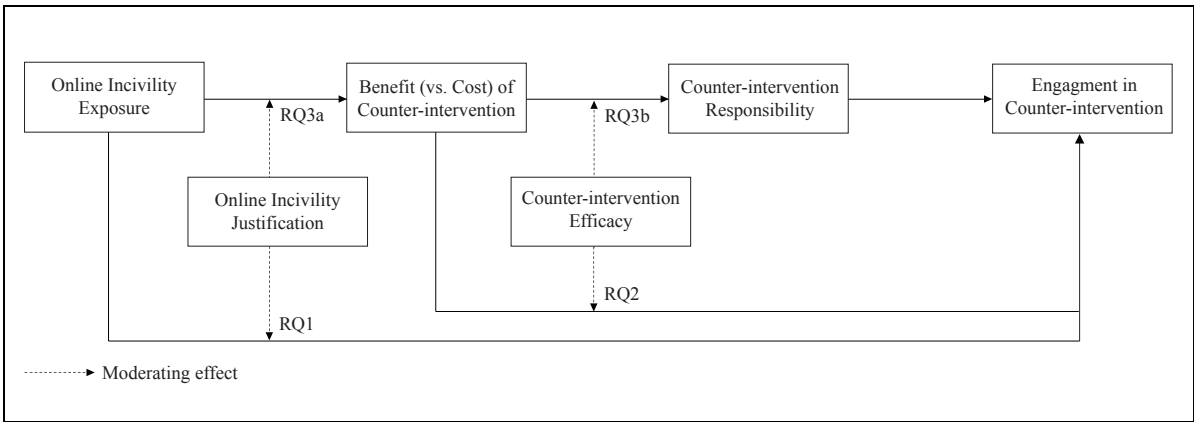
The amended final inquiry builds on the proposed research questions to explore a possible alternative sequence of steps leading to BCI Behavior. Consistent with the initial model, the first step is that bystanders are often exposed to online incivility incidents. Instead of assuming that bystanders try to assess the incident based on the impression formed of the perpetrator, the alternative model pays attention to how bystanders interpret the incivility itself. As far as bystanders consider the observed incivility acceptable or even appropriate in the given situation, they may determine that benefits outweigh costs when supporting it. Then, they may realize a need to act in support of the perpetrator and feel personally motivated to engage in BCI behavior on their own. Considering that diffusion of responsibility occurs as a means of reducing the psychological costs associated with non-intervention (Fischer et al., 2006; Markey, 2000), bystanders may more likely take counter-intervention responsibility as a means of reaping the positive benefits for both the perpetrator and themselves. This enhanced sense of responsibility is central to implementing BCI Behavior, but bystanders may demonstrate such motivation when they feel self-confident and self-effective in their capabilities to reinforce or assist in the perpetrator successfully. Taking all these conjectures together, the final post-hoc research question

explored the path including BCI Benefit–Cost and BCI Responsibility as two mediating steps in the relationship between Exposure and BCI Behavior. This question focused on whether this indirect path would be contingent upon the extent to which bystanders justify the observed incivility and feel efficacious for enacting BCI Behavior. The purpose of this inquiry was to thus identify whether Justification and BCI Efficacy would moderate the decision-making process to implement BCI Behavior.

RQ3a-b. Does a positive indirect effect of online incivility exposure on counter-intervention via benefit–cost perception of counter-intervention and counter-intervention responsibility appear among bystanders who (a) more strongly justify the observed incivility and (b) have greater counter-intervention efficacy?

Based on these three sets of post-hoc research questions, the alternative BCI model is proposed as shown in Figure 4 below.

Figure 4. A Proposed Alternative Bystander Counter-intervention Model



To test these research questions, a conditional process modeling analysis was conducted again on the same BCI survey data. Based on the alternative model, the analysis re-estimated the direct paths of Exposure, BCI Benefit–Cost, and BCI Responsibility to BCI

Behavior. Then, the indirect path of Exposure to BCI Behavior was assessed with a sequential mediation by BCI Benefit–Cost and BCI Responsibility. Within this path, Justification was incorporated as the first moderator variable into the relationship between Exposure and BCI Benefit–Cost, and BCI Efficacy was incorporated as the second moderator variable into the relationship between BCI Benefit–Cost and BCI Responsibility. As a reminder, the five variables (gender, age, online incivility victimization/perpetration, and social desirability bias) were controlled in the analysis.

The results from this post-hoc test indicated that the alternative BCI model achieved a satisfactory model fit ($X^2 = 11.13$, $df = 5$, $p = .05$; CFI = .93; RMSEA = .07 [90% CI = .01 to .14]; SRMR = .03). The detailed results are provided below and in Figure 6.

Direct Paths to Counter-intervention

The results of the direct paths showed that BCI Behavior had positive associations with Exposure ($b = .13$, $SE = .06$; $\beta = .13$, $p = .04$) and BCI Responsibility ($b = .10$, $SE = .05$; $\beta = .16$, $p = .03$) but not with BCI Benefit–Cost ($b = -.06$, $SE = .05$; $\beta = -.09$, $p = .27$). No significant interaction effects on BCI Behavior were found between Exposure and Justification ($b = -.01$, $SE = .05$; $\beta = -.01$, $p = .92$; RQ1) and between BCI Benefit–Cost and BCI Efficacy ($b = .01$, $SE = .03$; $\beta = .02$, $p = .80$; RQ2). That is, the two moderator variables—Justification and BCI Efficacy—did not work jointly with Exposure and BCI Benefit–Cost, respectively, to directly predict BCI Behavior.

Indirect Paths to Counter-intervention

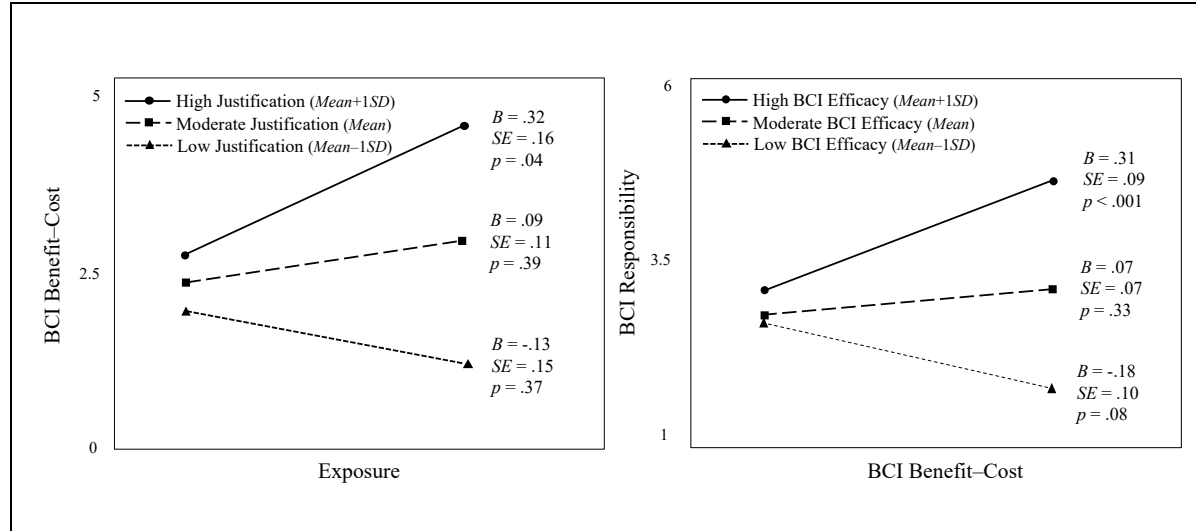
To assess the indirect paths, the analysis examined how each step led to its subsequent step. The results revealed that the interaction term of Exposure x Justification was significantly related to BCI Benefit–Cost ($b = .18$, $SE = .09$; $\beta = .15$, $p = .04$; RQ3a); the

interaction term of BCI Benefit–Cost x BCI Efficacy was significantly related to BCI Responsibility ($b = .16$, $SE = .04$; $\beta = .25$, $p < .001$; RQ3b); and BCI Responsibility was significantly related to BCI Behavior ($b = .10$, $SE = .05$; $\beta = .16$, $p = .03$).

Simple effect analyses were conducted to clarify the patterns of (a) how Exposure and Justification jointly affect BCI Benefit–Cost (see Figure 5 left) and (b) how BCI Benefit–Cost and BCI Efficacy jointly affect BCI Responsibility (see Figure 5 right). The first interaction (Exposure x Justification on BCI Benefit–Cost) was assessed by computing simple slopes at three levels ($M-1SD$, *mean*, and $M+1SD$) of Justification. This revealed a positive association between Exposure and BCI Benefit–Cost at the high level ($M+1SD$) of Justification ($b = .32$, $SE = .16$, $p = .04$), showing that participants who more often exposed to incivility perceive greater benefits (vs. cost) of supporting it especially when they more strongly justified it.

The second interaction (BCI Benefit–Cost x BCI Efficacy on BCI Responsibility) was also decomposed into three simple slopes at three levels ($M-1SD$, *mean*, and $M+1SD$) of BCI Efficacy. A positive association between BCI Benefit–Cost and BCI Responsibility was found at the high level ($M+1SD$) of BCI Efficacy ($b = .31$, $SE = .09$, $p < .001$). But this relationship became opposite at the low level ($M-1SD$) of BCI Efficacy, although marginally significant ($p < .10$), such that BCI Benefit–Cost was negatively related to BCI Responsibility ($b = -.18$, $SE = .10$, $p = .08$). These findings suggest that participants who perceived greater benefits (vs. costs) of supporting incivility felt greater need to act personally, but this effect only applied to those with greater counter-intervention efficacy. Those who lacked such efficacy, however, felt less need for BCI Behavior even when they expected positive consequences from that behavior.

Figure 5. Exposure x Justification and Benefit–Cost x Efficacy Interactions



Note. Interaction between Exposure and Justification on BCI Benefit–Cost (left) and Interaction between BCI Benefit–Cost and BCI Efficacy on BCI Responsibility (right)

These two interaction effects were involved in the moderated mediation effects on the indirect path of Exposure to BCI Behavior via Benefit–Cost and BCI Responsibility. The bootstrapping results showed that this path was significant in the following two conditions that involved different levels of Justification and BCI Efficacy.

First, Exposure positively led to BCI Behavior via BCI Benefit–Cost and BCI Responsibility at the high levels ($M+1SD$) of Justification and BCI Efficacy ($b = .01$, $SE = .01$, 95% CI = .001 to .04, 90% CI = .002 to .04). This result revealed a condition that could increase BCI Behavior among participants who more strongly justified online incivility. When this type of bystander experienced more exposure to incivility, they perceived greater benefits than costs of supporting it. Such perception of greater benefits (vs. costs) led bystanders to feel greater need to show that support personally, but this effect was found only among those who felt greater efficacy to take such actions. Finally, the resultant increase in BCI Responsibility led to greater engagement in BCI behavior.

The other condition for the significance of the indirect path was found at the high level ($M+1SD$) of Justification and the low level ($M-1SD$) of BCI Efficacy ($b = -.01$, $SE = .01$, 95% CI = $-.03$ to 0 , 90% CI = $-.03$ to $-.001$). This condition produced the opposite consequence, showing when BCI Behavior would less likely occur even though participants justified online incivility strongly. As shown in the former condition, this type of participant perceived greater benefits (vs. costs) of supporting incivility when they were more exposed to it. However, despite such increased perceived benefits (vs. costs), participants felt less need to support incivility personally as far as they felt less efficacy for taking such actions. The resultant decrease in BCI Responsibility led to less engagement in BCI behavior consequently.

Summary

The alternatively proposed BCI model obtained a good model fit by demonstrating significant direct and indirect paths to counter-intervention (see Figure 6 below). Consistent with the initial model, Exposure and BCI Responsibility showed significant positive relationships to BCI Behavior, respectively. An additional and more important finding from the alternative model was that online incivility exposure could affect counter-intervention through the following process: Bystanders who are more often exposed to online incivility would tend to perceive greater benefits than costs of counter-intervention as far as the observed incivility was considered warranted or well-grounded. This could lead bystanders to feel greater need for counter-intervention and in turn engage more frequently in that behavior as far as they feel greater efficacy to do so. However, the lack of counter-intervention efficacy could cause the opposite consequence (i.e., less counter-intervention),

by making bystanders feel it necessary to personally act to support the observed incivility even though they justify it.

Note. Gender, age, incivility victimization/perpetration, and social desirability bias were controlled.

intervention behavior (“BNI Behavior”). It was assumed that Exposure, Severity, BI Responsibility, and BI Efficacy would have negative relationships to BNI Behavior (H7a-d), and that the indirect path from Exposure to BNI Behavior would be mediated by lower levels of Severity, BI Responsibility, and BI Efficacy (H7e). The same five variables (gender, age, online incivility victimization/perpetration, and social desirability bias) were controlled in these path analyses.

The analysis results from both the BI and BNI survey datasets suggested an acceptable model fit for each model. The BI model, which was tested using the BI survey, showed a good model fit ($CFI = .97$; $RMSEA = .08$ [90% CI = .03 to .15]; $SRMR = .02$) although the chi-square result was slightly less than the desirable value ($X^2 = 9.42$, $df = 3$, $p = .02$). An even better model fit was found for the BNI model, which was tested using the BNI survey ($X^2 = 3.25$, $df = 3$, $p = .35$; $CFI = 1.0$; $RMSEA = .02$ [90% CI = 0 to .10]; $SRMR = .01$).

Direct Paths to Intervention and Non-intervention (H6a-d and H7a-d)

Examination of the direct paths to BI Behavior indicated that this behavioral outcome was significantly related with Exposure ($b = .24$, $SE = .06$; $\beta = .25$, $p < .001$) and BI Responsibility ($b = .12$, $SE = .06$; $\beta = .15$, $p = .03$), supporting H6a and H6c. No significant associations were found between Severity and BI Behavior ($b = -.05$, $SE = .04$; $\beta = -.07$, $p = .25$) and between BI Efficacy and BI Behavior ($b = .06$, $SE = .05$; $\beta = .07$, $p = .25$), rejecting H6b and H6d. These findings showed that participants intervened more often in online incivility when they experienced more exposure to it and when they felt greater need to confront it.

Meanwhile, the significant direct path to BNI Behavior was found only with BI Responsibility ($b = -.14$, $SE = .05$; $\beta = -.18$, $p = .01$), supporting H7c. No significant associations were found between Exposure and BNI Behavior ($b = .03$, $SE = .06$; $\beta = .02$, $p = .69$), between Severity and BNI Behavior ($b = .06$, $SE = .05$; $\beta = .08$, $p = .19$), and between BI Efficacy and BI Behavior ($b = -.06$, $SE = .05$; $\beta = -.09$, $p = .17$), rejecting H7a, H7b, and H7d. These findings suggest that participants chose non-intervention more often when they felt less need to confront the observed incivility.

Indirect Paths to Intervention and Non-intervention (H6e and H7e)

H6e predicted that the relationship between Exposure and BI Behavior would be mediated by perceiving greater severity of incivility and feeling greater responsibility and efficacy for intervention. However, the bootstrapping results showed that the indirect path from Exposure to BI Behavior via Severity, BI Responsibility, and BI Efficacy was not significant at both 95% and 90% CIs ($b = .001$, $SE = .001$; 95% CI = 0 to .004; 90% CI = 0 to .004), rejecting H6e. Instead, a partial indirect path among the intermediate steps—Severity, BI Responsibility, and BI Efficacy—was found to be significant at both 95% and 90% CIs ($b = .10$, $SE = .02$; 95% CI = .07 to .15; 90% CI = .07 to .16). Each step in this partial path predicted its subsequent step significantly, such that Severity positively predicted BI Responsibility ($b = .26$, $SE = .05$; $\beta = .30$, $p < .001$), and in turn BI Responsibility positively predicted BI Efficacy ($b = .40$, $SE = .07$; $\beta = .39$, $p < .001$).

In contrast, H7e predicted that the relationship between Exposure and BNI Behavior would be mediated by perceiving less severity of incivility and feeling less responsibility and efficacy for intervention. The bootstrapping procedure revealed that the indirect path from Exposure to BNI Behavior via Severity, BI Responsibility, and BI Efficacy was not

significant either at both 95% and 90% CIs ($b = 0$, $SE = .001$; 95% CI = $-.003$ to 0 ; 90% CI = $-.003$ to $.001$), rejecting H7e. However, the same partial indirect path from Severity to BI Responsibility to BI Efficacy was also found to be significant at both 95% and 90% CIs in this BNI model ($b = .10$, $SE = .02$; 95% CI = $.05$ to $.17$; 90% CI = $.04$ to $.18$). That is, for non-intervening participants, Severity positively predicted BI Responsibility ($b = .50$, $SE = .06$; $\beta = .50$, $p < .001$), and in turn BI Responsibility positively predicted BI Efficacy ($b = .20$, $SE = .07$; $\beta = .18$, $p = .004$) as well.

Reduced Model Tests

The findings above were obtained from the analysis that considered all the possible direct and indirect associations among the variables simultaneously. The current research attempted a further examination to obtain a reduced model (i.e., the model with a smaller number of parameters; Preacher & Yaremych, 2022) to find the most parsimonious model that has the best explanatory power for the same data (Kline, 2005). With the sole focus on the direct and indirect effects of Exposure on each type of behavior, a reduced model was constructed without consideration of the direct associations between the mediator variables and the outcome variable. That is, the reduced model on BI behavior was designed to involve (a) the direct path from Exposure to BI behavior and (b) the indirect path from Exposure to BI behavior via Severity, BI Responsibility, and BI Efficacy. The same procedure was applied to testing the reduced model on BNI behavior.

The analysis of the BI survey data still indicated an acceptable model fit for the reduced BI model ($CFI = .95$; $RMSEA = .08$ [90% CI = $.04$ to $.13$]; $SRMR = .03$) although the chi-square result was slightly less than the desirable value ($X^2 = 15.56$, $df = 5$, $p = .01$). The reduced BI model revealed a significant direct path from Exposure to BI behavior ($b =$

.26, $SE = .06$; $\beta = .26$, $p < .001$); however, the five-step indirect path from Exposure to BI behavior was not significant though ($b = .001$, $SE = .001$; 95% CI = 0 to .006; 90% CI = 0 to .004). Instead, the bootstrapping results of this reduced model showed the significance for the *partial* (four-step) indirect path from Severity to BI behavior at both 95% and 90% CIs ($b = .01$, $SE = .01$; 95% CI = .002 to .022; 90% CI = .003 to .02). Each step significantly predicted its subsequent step in this partial path, such that Severity positively predicted BI responsibility ($b = .26$, $SE = .05$; $\beta = .30$, $p < .001$), BI responsibility positively predicted BI efficacy ($b = .40$, $SE = .07$; $\beta = .39$, $p < .001$), and then BI efficacy positively predicted BI behavior ($b = .10$, $SE = .04$; $\beta = .13$, $p = .02$).

Likewise, the analysis of the BNI survey data indicated an acceptable model fit for the reduced BNI model ($CFI = .97$; $RMSEA = .07$ [90% CI = .01 to .12]; $SRMR = .03$) although the chi-square result was slightly less than the desirable value ($X^2 = 11.66$, $df = 5$, $p = .04$). The reduced BNI model revealed that neither the direct path ($b = .01$, $SE = .07$; $\beta = .01$, $p = .83$) nor the indirect path ($b = 0$, $SE = .001$; 95% CI = -.004 to .001; 90% CI = -.003 to 0) was significant between Exposure and BNI behavior. However, the *partial* (four-step) indirect path from Severity to BNI behavior was significant in this reduced model at 90% CI ($b = -.008$, $SE = .005$; 95% CI = -.022 to 0; 90% CI = -.02 to -.001). Each step predicted its subsequent step significantly in this partial path, such that Severity positively predicted BI responsibility ($b = .50$, $SE = .06$; $\beta = .50$, $p < .001$), BI responsibility positively predicted BI efficacy ($b = .18$, $SE = .06$; $\beta = .18$, $p = .004$), and BI efficacy negatively predicted BNI behavior at a marginally significant level ($b = -.12$, $SE = .06$; $\beta = -.12$, $p = .054$).

Summary

The analysis results of the current BI and BNI survey datasets partially validated Latané and Darley's (1970) assumptions (see Figure 7). Significant direct paths to BI and BNI behaviors were found with two factors: Exposure and BI responsibility. Participants who experienced more exposure to online incivility showed greater engagement in BI behavior. Also, those who felt greater responsibility for intervention showed increased BI behavior but decreased BNI behavior. Meanwhile, the reduced model tests confirmed that the three steps—Severity, BI responsibility, and BI efficacy—shaped a significant indirect path leading to both BI behavior and BNI behavior (see Figure 8). The results showed that participants who perceived greater severity from the observed incivility felt greater responsibility to confront it, which was also predictive of the feeling of greater efficacy to do so. The resultant increased intervention efficacy finally led to increased BI behavior but decreased BNI behavior.

Figure 7. Test Results of the Full Bystander Intervention/Non-Intervention Models

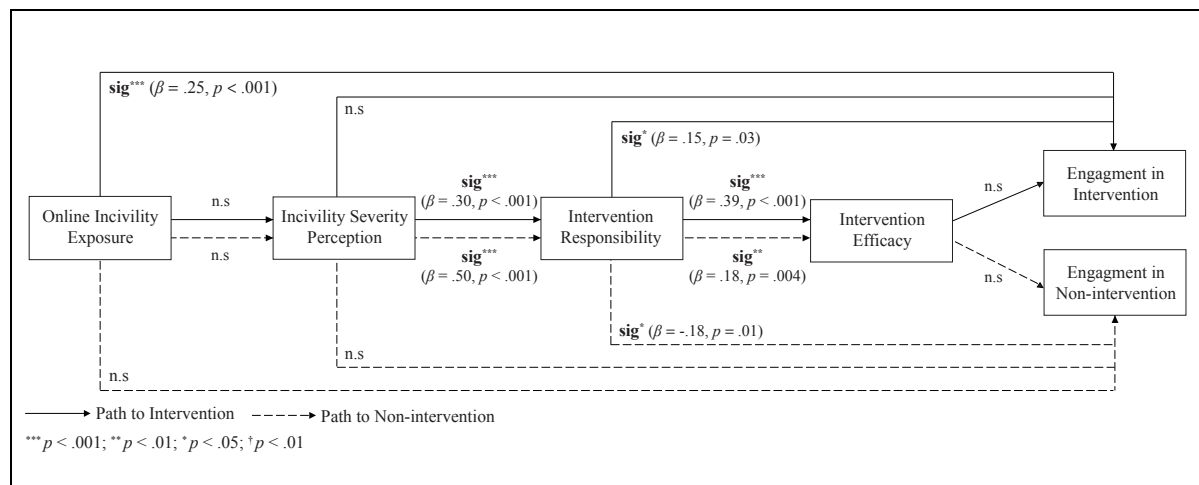
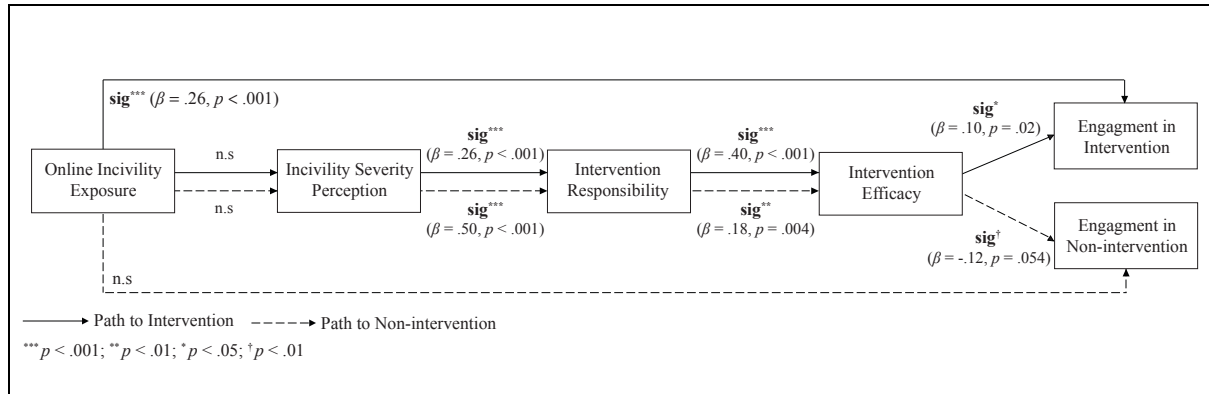


Figure 8. Test Results of the Reduced Bystander Intervention/Non-Intervention Models



Note. Gender, age, incivility victimization/perpetration, and social desirability bias were controlled.

Chapter IX: Survey Study Discussion

Summaries and Interpretations

The Initial Bystander Counter-intervention Model

The current research aimed to establish a new theoretical model of bystander counter-intervention that illustrates a decision-making process of bystanders reinforcing or assisting incivility perpetrators. The hypothetical process consisted of a series of multiple steps that were inspired by Latané and Darley's (1970) BI model. The initially designed model represented a sequential path starting from exposure to online incivility and leading to BCI behavior through justification and the benefit-cost perception of the observed incivility and then feelings of responsibility and efficacy for the intended behavior. Also, two factors were assumed to moderate this path: bystanders' attitudes toward perpetrators and the perceived social norms regarding incivility. However, the survey data did not yield a good model fit. The rejected hypotheses indicated no direct relationships from the justification and efficacy steps to BCI behavior. These two factors did not individually predict the behavioral outcome but instead were significantly related to the responsibility step. Given the qualitative findings from the earlier interview study regarding the important roles of

these factors, the current research did not rule out the alternative possibility that the justification and efficacy factors might jointly contribute to impacts on the behavioral outcome. Some exploratory research questions were thus posed to see whether these two factors interact with other factors on the path to further increase engagement in BCI behavior. These questions were answered by the alternative model, which will be further discussed below. However, I will first discuss possible explanations for why pro-perpetrator attitudes and pro-incivility norms did not have an effect on bystander behavior in either the initial or alternative models.

Non-significant Effect of Pro-perpetrator Attitude. Previous research suggests that the amount of social or psychological closeness between a bystander and a target can encourage appropriate situational interpretation (Dovidio et al., 2006) and thus increase the likelihood of BI behavior (Levine et al., 2002). The current research applied this same approach to predict BCI behavior by focusing on the bystander-perpetrator relation. Although the interview findings showed that counter-interveners tend to have more positive attitudes of and perceive high social norms toward incivility perpetrators, neither the main nor joint effects were found from these two factors to influence BCI behavior, according to survey data analyses. A possible explanation can be inferred from conflicting evidence present in the interview responses. The positive attitudes held by interviewees toward incivility perpetrators were largely due to what the perpetrators said and how they said it, which demonstrates their beliefs and viewpoints, or discursive strategies and ways of talking. This could also help generate favorable impressions toward the perpetrators, but it may not necessarily in itself completely explain the decision to counter-intervene. An illustration of this complexity is reflected in the responses by some interviewees saying, “I

thought that [the perpetrators] were more credible due to their beliefs. However, I did not have any specific positive or negative thoughts about them” and “I didn’t feel anything towards [the perpetrator] other than it was a clever response. I liked the response and not him.” As such, an additional review of the interview data suggests that attitudes toward the observed incivility, not toward perpetrators, may be a more likely basis for BCI behavior.

Non-significant Effect of Pro-incivility Norm. Given the interview findings that counter-interveners tend to perceive that social norms endorse uncivil discussion, the initial model deemed compliance with such pro-incivility norms to be one of the important drivers for BCI behavior. This prediction was in line with previous research indicating that bystanders’ responses are influenced by perceived social norms, such that those who believe that most others are more tolerant to cyberbullying (i.e., pro-cyberbullying) tend to reflect this behavior and show a higher intention to reinforce the bully (Dang & Liu, 2020; Panumaporn et al., 2020). However, the survey data did not reveal any evidence of this social norm factor influencing BCI behavior. The hypothesis that the association between counter-intervention responsibility and BCI behavior might vary with perceived pro-incivility social norms was not supported. The social norm factor itself did not show a statistically significant relationship with BCI behavior in this survey study. A possible explanation for this result is that the perception that social norms exist and one’s sense of obligation to those norms may need to be addressed separately. It is notable that not all interviewees agreed on the possibility that their behavior was influenced by norms, regardless of whether they knew the existence of others and how they responded. Although not the majority, there were several interviewees who denied their susceptibility to normative social influence. They said, for example, “I don’t believe that other responses

encourage or discourage me from reacting with my opinion, especially favorably” and “The way that others responded to this toxic comment did not influence how I reacted at all. That was my personal choice as to how to react to it.” These interviewees believed that their support for the perpetrator was solely their own decision. Thus, the effect of the pro-incivility norm on BCI behavior would need to be addressed jointly with bystanders’ responsiveness or their sensitivity to social influence.

The Alternative Bystander Counter-intervention Model

Despite an unsatisfactory model fit, the initial analysis revealed some notable patterns that emerged from the survey data as follows: (1) bystanders with more frequent exposure to online incivility showed greater engagement in BCI behavior, (2) bystanders with a strong justification for observed incivility perceived greater benefits of BCI behavior and in turn felt greater BCI responsibility, and (3) bystanders with greater BCI responsibility engaged in more BCI behavior. These inconclusive but promising findings from the initial model test, along with an additional review of qualitative data from the earlier interview study, guided the design of the alternative BCI model as a reduced path model with non-significant variables eliminated from the model. The alternative model demonstrated a superior fit to the survey data, meaning that it depicted patterns or dynamics of bystander responses present in the data relatively well.

Consistent with the initial model test, the alternative BCI model showed that bystanders who experienced more frequent exposure to online incivility engaged more in BCI behavior. At the same time, the test of the original BI model also found that a higher level of incivility exposure led to greater engagement in BI behavior, supporting previous evidence (e.g., Dillon & Bushman, 2015; Ziegele et al., 2020). Such findings from the two

models highlight that there are certain conditions that accompany witnessing online incivility incidents, which determine whether bystanders show prosocial or antisocial behavioral responses. The alternative BCI model revealed both direct and indirect paths that connect exposure to BCI behavior in positive and statistically significant ways. According to the indirect path shown in the model, more frequent exposure to uncivil online discussions could evoke the perception that there are greater benefits than costs involved in a supportive reaction to the observed incivility, especially when it is considered justifiable in the given discussion. Given this finding, the importance of the justification process cannot be emphasized enough when taking into consideration what may lay the groundwork for BCI behavior.

The current research uncovers that counter-intervening bystanders tend to frame incivility as “morally justifiable” or a “normal” collective action, as its perpetrators do, and unite in virtual outrage on online platforms. This finding suggests that a crucial part of the justification process involves assessment of the benefits (vs. costs) caused by support for incivility. If the intent or impact of observed incivility is sufficiently justified, in the name of moral correctness, free speech, or open debate, it can nurture expectations regarding the merit of BCI behavior—for example, social support and acceptance from like-minded people or self-pride about contributing to the discussion. These beliefs outweigh concerns over the potential negative evaluations or platform moderations. This can be further enhanced as bystanders are heavily exposed to incivility, suggesting that the consequential effect of exposure frequency is to move beyond simply being tolerant of incivility toward viewing it as less costly and even a pleasurable and potentially rewarding act.

The current research newly conceives a notion of counter-intervention responsibility as a fundamental part of this motivational process by which bystanders decide to personally engage in BCI behavior. The more that bystanders deem it advantageous to support the observed “justifiable” uncivil act, the more they feel a need to act in support of perpetrators who share a similar opinion or orientation about how the target should be addressed. Counter-intervention responsibility reflects that BCI behavior is driven by a homophily or in-group bias toward perpetrators. The finding that bystanders are motivated to support perpetrators whose incivility they believe is justifiable and beneficial suggests that they deem BCI behavior as a means of fulfilling their need to protect and extend their shared interests and goals. This motive is largely in line with that of bullying perpetrators who use aggressive means to reach a particular goal (i.e., proactive or instrumental aggressiveness; Roland & Idsøe, 2001), such as achieving power or enhancing social affiliation (Solomontos-Kountouri & Strohmeier, 2021), and seek social or self-rewards. For example, some types of perpetrators of hate speech comments are organized and have a “mission” to defend the interests of their group and attack enemies (see the “soldier” and “believer” types of perpetrators; Erjavec & Kovačič, 2012). This motive can be extended to counter-interveners who are dedicated to reinforcing the belief and group identity shared with the perpetrators. In this way, online incivility can be amplified through many helping hands.

A notable finding of the present research is that the function of counter-intervention responsibility is subject to whether bystanders have sufficient confidence in demonstrating their support for the perpetrator. This sense of counter-intervention efficacy does not directly influence BCI behavior, rejecting the initial model’s hypothesis. Instead, the alternative model reveals that counter-intervention efficacy is crucial for evoking an adequate level of

counter-intervention responsibility that can determine BCI behavior. Previous research has consistently reported that low self-efficacy in relation to the intended goal, as well as the perceived costs to individual bystanders, influence whether they decide to intervene (e.g., Banyard, 2008). With regard to bullying, bystanders often believe that their possible actions will be ineffective or that they will suffer retaliation from the perpetrator or other bystanders (Lodge & Frydenberg, 2005). This efficacy factor operates similarly but in the opposite way in the process of BCI behavior, along with the benefit-cost perception. The current research reveals that, even when bystanders anticipate the intended behavior would yield greater benefits than costs, it is important for them to ascertain whether they are truly able and knowledgeable enough to obtain those benefits. For example, not knowing what actions to take in order to support the perpetrator or lacking necessary counter-intervention “skills” may inhibit bystanders from becoming involved personally, even when they consider the perpetrator’s act acceptable and socially beneficial. This critical role of efficacy for antisocial bystander behavior has not been documented in past work and must be studied further.

The Bystander Intervention and Non-intervention Models

Two additional surveys were conducted in the current research to further test the validity of Latané and Darley’s (1970) BI model in the context of online incivility. Although the survey findings did not sufficiently support the original assumptions, some substantive findings are noted. Two factors were found to positively impact the chance of BI behavior. First, frequency of incivility exposure was found to increase BI behavior. Given that there was also a positive association between exposure and BCI behavior, but no significant association between exposure and BNI behavior, it is possible to conclude that those who

habitually observe online incivility become a kind of “responsive” bystander, either the prosocial or antisocial type. Second, another important factor determining BI and BNI behaviors was a sense of intervention responsibility. Assuming this responsibility personally to defend the target and confront the perpetrator increases the chance of BI behavior but decreases that of BNI behavior. Again, this finding—along with the finding related to counter-intervention responsibility—highlights the responsibility factor as a crucial boundary condition for bystander behavior.

The originally assumed path (Latané & Darley, 1970) was largely confirmed when reduced models were tested with a sole focus on the direct and indirect effects of exposure on each behavioral outcome. Although the five-step sequence was not fully supported by the survey data, a significant partial indirect path was found that starts with perception of incivility severity (i.e., Step 2) and ends with each behavioral outcome. This suggests that bystanders who perceive greater severity from the observed incivility feel greater senses of intervention responsibility and in turn intervention efficacy. This process increases BI behavior but decreases BNI behavior.

Given all these findings, the original BI model (Latané & Darley, 1970) was partially validated in the context of online incivility. Except for exposure as the supposed first step of the model, the remaining steps occur sequentially to yield either BI behavior or BNI behavior. According to the current BI model applied in an online incivility situation, bystanders must, primarily, (1) assess the observed incivility as severe; (2) perceive that they are personally responsible for intervening; and finally (3) feel that they have sufficient self-efficacy for intervening. Those who undergo these three steps will make the decision to intervene and act accordingly. That is, passive responses can be inhibited in this process,

such that bystanders are less inclined to non-intervene as far as all these three steps are sufficiently taken.

Again, the exposure step was found to predict BI behavior directly but not indirectly. Some clarifications may be needed as to why the exposure step does not lead to its hypothesized second step, i.e., perception of incivility severity. The three surveys on BCI, BI, and BNI behaviors measured the extent to which their corresponding participants were desensitized to online incivility. This factor was measured to illustrate the different aspects between counter-interveners, interveners, and non-interveners; however, it was not used to predict bystander behavior due to a lack of support in the qualitative data.⁴ The common view is that extensive and cumulated exposure to aggressive behavior causes a reduction in cognitive, physiological, emotional, or behavioral responsiveness (Funk et al., 2004). For example, people become less sensitive to bullying (e.g., Gini et al., 2008; Pabian & Vandebosch, 2016; Rudy & Linz, 2008) or hate speech (Bilewicz & Soral, 2020; SELMA, 2019) as they are habitually exposed to it. In the current research, this desensitizing effect to cumulative incivility exposure was not found among any of the bystander types. Rather, non-interveners were found to be *less* desensitized to incivility as they experienced more frequent exposure to it ($b = -.17$, $SE = .06$; $\beta = -.17$, $p = .003$). For interveners ($b = -.03$, $SE = .05$; $\beta = -.03$, $p = .62$) and counter-interveners ($b = .02$, $SE = .06$; $\beta = .02$, $p = .76$), no significant relationship between exposure and desensitization was detected.⁵ This suggests

⁴ As a reminder, the earlier interview study addressed the desensitization factor in its investigation but did not provide convincing evidence to support its potential effect on BCI behavior. Thus, this factor was not included as the main focus of the survey study. However, it was measured to identify whether it might indicate notable differences across bystander groups.

⁵ Participants' desensitization to incivility was measured in two ways. First, three items were extracted from the interview responses to assess the degree of participants' emotional desensitization to incivility in the recalled uncivil comment(s): "I felt the toxic comment(s) were humorous and funny," "I felt the toxic comment(s) were offensive and harsh (reverse-coded)," and "I did not feel

that greater exposure frequency leads bystanders to become more “responsive” to uncivil online discussions, either in a prosocial or antisocial way, but it serves to neither increase nor decrease their sensitivity to incivility. This can help partially explain why the exposure step does not lead to the severity step in the tested models.

Limitations and Future Research Directions

Although the current research adds to research demonstrating the novel decision-making process of BCI behavior, its results and contributions must be understood in light of certain limitations. First, the survey data were gathered from using a non-representative sample of U.S. adults who use online platforms (especially comment sections of social media and news media) as a recourse of political or social discussion and observe a certain amount of incivility in their everyday online activities. However, the sample is not generalizable. Also, given that the current survey sample was predominately *white/Caucasian* (71.5%), future studies should aim to recruit more ethnically diverse samples, especially increasing *Hispanic*, *African-American*, and *Asian* participants who may have experience with online incivility from a “target” perspective.

Second, this study used cross-sectional data to validate the proposed models to predict bystander behavior. While a similar approach was often taken in previous studies that administered surveys to test the BI model (e.g., Ferreira et al., 2020; Nickerson et al., 2014), the direction of the causal path between different steps could not be proved conclusively with cross-sectional data only. It would have been ideal to have more waves of

any particular way about the toxic comment(s) (reverse-coded).” Participants responded to each item on a 7-point scale (1 = *not at all applicable* to 7 = *extremely applicable*). An additional question asked the extent to which participants thought they were desensitized to how uncivil the comment(s) were (1 = *not at all desensitized* to 7 = *extremely desensitized*). The scores were averaged to give an index of desensitization.

bystander data regarding all included variables, in order to more fully examine how engagement in BCI, BI, and BNI behaviors over time would be associated with step variables from the corresponding model. However, this was not feasible in the current surveys, thus future studies would benefit from using longitudinal methods to better understand the causal dynamics of these processes. Among additional limitations regarding the survey data is the study's use of self-reported rather than direct behavioral measures. Self-reporting participants may not understand the given questionnaire or give socially desirable, rather than behaviorally accurate, responses. These limitations might not be totally excluded, although the current research tried to address these issues by providing participants with definitions and descriptions of the studied variables and statistically controlling their social desirability biases. Researchers would benefit from incorporating quasi-experimental or other observational methods of bystander behavior into future studies on this topic.

Third, it is conceivable that bystander roles could change over time or across situations and that people could have multiple roles as counter-interveners, interveners, and non-interveners in online incivility situations. This was considered in the current research, so that those who reported experiences of playing more than one bystander role completed all the surveys that applied to them. Since the focus of the current study was to identify the underlying processes leading to involvement in each behavior, an examination of whether those with multiple roles undergo those processes differently from other types of bystanders was not explored. This was beyond the scope of the current study and is an area for future research.

Last but not least, the current survey questionnaire could be improved to increase measurement reliability and generalizability. Considering that a wide range of antisocial bystander behaviors are available in online incivility situations, further investigation should be made to identify the additional behavioral options adopted by counter-interveners. The current study only considered positive responses toward the perpetrator, but counter-interveners may also negatively respond to the target or the target's supporters (e.g., disliking/downvoting replies that confront the perpetrator or defend the target). Providing more specific behavioral descriptions can help participants not to underestimate their BCI behavior. Also, given that current survey items were mostly adapted from excerpts from the interview data, a more reliable scale should be designed to further validate the BCI model.

Chapter X: General Discussion

The current dissertation aimed to illustrate processes by which the bystander public reinforces and assists perpetrators of online incivility. I have done this by proposing a new notion of bystander counter-intervention. With this newly coined term, the dissertation used an alternative typology of three different types of bystander behavior—intervention, counter-intervention, and non-intervention—to provide a general overview of the roles of bystander public in online incivility phenomena. This typology is intended to explore different behavioral responses that appear within and among individual bystanders rather than assign fixed and defined roles to them. The dissertation sheds light on bystanders as dynamic actors who might adapt to situational demands and move across different points on a behavioral spectrum between interveners, counter-interveners, and non-interveners. In this mixed-methods sequential exploratory research, qualitative data were obtained first followed by further quantitative data collection to establish the prediction model for BCI behavior. An

interview study with a small but informative sample of counter-interveners was conducted to guide the main survey study. This preliminary investigation explored what motives and reasons counter-interveners might have for deciding their behavioral response and helped formulate the specific hypotheses and research questions about the decision-making process they might follow. The resultant findings were synthesized into an integrative theoretical model that specifies the relative influence and interplay of multiple drivers of bystander counter-intervention. The insights offered by this research have important implications from both theoretical and practical perspectives.

Theoretical Contributions

Much of the focus of the current dissertation was inspired by the tremendous progress that has been made over the past decades in understanding the psychological processes of bystander behavior. Special interest was paid to Latané and Darley's (1970) BI model, derived from social psychology, that emphasizes that bystanders are likely to decide their behavior along a causal chain whereby they first notice and interpret a given incident, then accept responsibility to be involved, and finally have efficacy in how to act. This model has been successfully applied (totally or partially) in different areas, but to date no studies have adapted this traditional framework to predict antisocial bystander behavior. The current dissertation filled this gap by developing and testing an alternative sequential path model detailing the steps one must take in order to counter-intervene. After multiple tests, the results obtained through this approach suggest that BCI behavior cannot be always reduced to an impulsive antisocial behavior but entails a series of considerations and decisions to achieve a desired goal. Findings from each component of the new BCI model can contribute to advancing the bystander research paradigm.

By adapting the traditional BI model, the present research was able to clarify the mixed evidence that frequency of online incivility exposure is not only positively related to showing prosocial bystander behavior (e.g., Dillon & Bushman, 2015; Ziegele et al., 2020) but also positively related to being a perpetrator (e.g., Wachs & Wright, 2018; Wachs et al., 2019). The current survey findings also revealed that those with a higher frequency of online incivility exposure engaged in both BI and BCI behaviors at a greater level compared those with less exposure. Higher exposure frequency becomes the driving factor for counter-interveners when they see incivility as justifiable. This is especially true when the observed incivility is deemed to have appropriate intentions (e.g., correction or retribution), even if it sometimes involves personal attacks or public shaming. This finding supports the argument that an individual's evaluation of a situation is a vital factor in determining whether that person decides to offer aid (Fazio, 1990). BCI behavior in the face of online incivility can be considered a function of the individual's interpretation or definition of incivility in their everyday life (Taylor et al., 2006). Those who perceive the observed incivility as normal or acceptable, and who consider targets as people who deserve to be punished, are more likely to engage in counter-intervention. Future research should consider different levels of online incivility in the investigation of counter-intervention. Although the current interview study suggests that counter-interveners do not necessarily experience desensitization to incivility, it is uncertain whether their behavioral intention decreases with a more severe incivility incident (Bastiaensens et al., 2014).

The next factor driving BCI behavior is the perception of the benefits (vs. costs) of this behavior, which reflects the desires of counter-interveners seeking to enhance social status through this behavior. This was evidenced by interviewees who expected that BCI

behavior would allow them to gain likes/upvotes and followers/friends or increase the popularity and visibility of their belief over that of their opponents. This finding is in line with previous studies that have demonstrated that the benefits of bystander behavior involve the gratification of their need for social approval (e.g., Pouwels et al., 2017; van der Ploeg et al., 2017). Engaging in BCI behavior can be a way for people to increase positive self-regard and avoid being a minority in the discussion by empowering or validating the perpetrator's opinion to exert more influence on the discussion.

The current research also highlights self-efficacy as a critical factor in determining counter-intervention. It impacts the degree of personal commitment and mental effort that people invest in supporting and defending a perpetrator, which then affects their likelihood of counter-intervening. The current research focused on self-efficacy as specific to BCI behavior, but other dimensions of self-efficacy could be further examined within this context. For example, political self-efficacy and writing self-efficacy, or the perceived group efficacy of solving collective problems through unified efforts, have been considered in research on uncivil political online discussions (e.g., Ziegele et al., 2020). Internet usage efficacy (Xiao & Wong, 2013) may also be necessary conditions to help bystanders cope with the potential costs of BCI behavior. Further research would be needed to identify the relevant antecedents to this sense of efficacy, which might help mitigate the subtle motivation for antisocial bystander behavior.

The other part of the current investigation was devoted to validating Latané and Darley's (1970) five-step BI model in the context of online incivility. The core assumption to be tested was that each step of this model predicted the next step in the model. The present research found partial support for this idea, such that the level of perceived severity

of observed incivility (Step 2) boosts bystanders' senses of responsibility (Step 3) and efficacy (Step 4) to confront it, and that this process leads to an increased likelihood of BI behavior and a decreased likelihood of BNI behavior (Step 5). These findings are in harmony with those from previous research that identified the indirect relationships of those psychological factors and the prosocial behavioral outcome from bystanders of cyberbullying or hate speech on social media (Beavon et al., 2022; Ferreira et al., 2019; Leonhard et al., 2018). The present research demonstrates that this 1970's bystander theory is still worth to be used for research of prosocial aspects of bystander behavior, given that it does not solely focus on the end behavior of intervening but identifies important steps leading to intervention. To better evaluate the boundary of applicability of the model, future research can be directed toward exploring the associations between individual characteristics of bystanders and each step of the BI model for online incivility contexts. Similar attempts have been made in some studies that suggest that the BI model's process might occur differently with individual bystanders' social identities (Jenkins et al., 2023), past bullying experiences (Beavon et al., 2022), and social-emotional wellness (Aliberti et al., 2022). Activating those particular characteristics of bystanders may prove useful for facilitating the process for BI behavior.

Meanwhile, the results of this study additionally informed that bystanders having ever experienced online incivility as victims showed an increase in BI behavior and a decrease in BNI behavior. This victimization factor, however, was not significantly associated with the steps of the BI model and BNI model, except for the first step. That is, bystanders with their own victimization experience showed higher exposure to incivility (directed at someone else) and greater intervention against it, but their behavior might occur

through other mechanisms than the path that is supposed by the model. Further research is needed to better understand why this type of bystanders engage in BI and BNI behaviors in different way, and in a practical sense, how we can incorporate that this experiential factor as a meaningful source for motivating them to take the lead in incivility intervention.

Practical Suggestions

Bystanders are ubiquitous in the context of online incivility. An understanding of bystander engagement is needed to establish how to mitigate this problem. A recent survey shows that the majority of the American public agrees that laws should be strengthened to hold perpetrators (82%) and online platforms (77%) accountable, policies should have more training and resources to help targets (79%), and online platforms should do more to counter online hate (78%; Anti-Defamation League, 2022). However, the public is also highly critical of the role of social media companies, including the ways that they are handling abuse and user harassment on their platforms (79%; Pew Research Center, 2021) and how the government is policing them (87.5%; Anti-Defamation League, 2020). Formal social control executed by professional moderators and authorities (e.g., content screening, user monitoring; Watson et al., 2019) involves labor-intensive resources and lacks effectiveness (Goodman, 2013); this has led many news media companies (e.g., CNN, NPR) to reduce or even disable the potential for user input on their platforms as a more cost-effective approach to managing online hate (Finley, 2015). Some have also turned their attention to user-driven efforts to correct and combat incivility perpetrators. Given that a small number of users confronting online hate posts can influence a larger audience (Schieb & Preuss, 2016), and that those users who have a moderator role can break the aggressive cycle (Ouvrein et al., 2018), there have appeared some practical training programs aimed to invite the bystander

public as potential allies in efforts to reduce uncivil online discourses and cyberbullying in digital environments (DiFranzo et al., 2019; NICD, n.d.).

However, the promise of such programs might depend on a more complete understanding of when potential bystanders fail to act in support of targets and instead choose to support perpetrators. Thus, the focus of this dissertation was on bystanders who play a critical role in a social reality that empowers online incivility. Achieving greater knowledge about what fosters BCI behavior will allow better development of strategies to cope with this problem. One possible way of preventing BCI behavior is to promote introspection of one's own internal biases (e.g., confirmation bias, homophily) that may lead to the justification of online incivility and the sense of counter-intervention responsibility. It may be difficult to instill this type of self-awareness among adult bystanders, though interventions may have more success by educating children. In sum, multilayered and multifaceted behavior-modification guidelines are needed to successfully address this complex problem (Bradshaw, 2015).

A step-by-step process leading to BCI behavior, informed by this research, can help practitioners to develop initiatives, such as bystander training programs or consent-based workshops that are tailored for potential counter-interveners. Different specific strategies can be used to design preventive and deterrent messages on BCI behavior by focusing on elements associated with each step. Given that more frequent exposure to incivility can elicit not only BCI behavior but also BI behavior, we would need to implement appropriate tactics to increase the chance of the latter behavior. For example, it might be possible for platform content moderators to highlight the negative implications of uncivil user postings so that bystanders can better recognize the perpetrator's intent to harm or the power imbalance that

may exist between the perpetrator and target. To undermine the justification of observed incivility and the perception benefit of BCI behavior, it is necessary to use message frames (e.g., in platform policies and community guidelines) that highlight the severity of incivility (e.g., its harmful consequences) or social disapproval and condemnation of incivility. Transporting bystanders into an engaging narrative to promote moral engagement (e.g., encouraging civil discourse and netiquette) and emphasizing the costs associated with BCI behavior (e.g., being reported or receiving harsh responses) may help lessen the sense of counter-intervention responsibility and efficacy expectations. Using media messages that negatively portray cases of fruitless BCI behavior and outcomes may also demoralize potential counter-interveners. Indeed, existing prevention or intervention programs have been designed to try to educate people about the risks of online bullying or harassment and how to help stop it (e.g., Chen et al., 2023; Choi & Park, 2021; Salem et al., 2023) through skill-building, psychoeducation, and multimedia materials (see Polanin et al., 2022). But findings from this dissertation may yield more nuanced and specific tactics to hinder the bystander public from making behavioral choices toward uncivil interactions online.

Along with education programs targeting individual bystanders, another actionable solution may come from technological responses. Particular attention should be given to the role of online media companies in structuring social environments that may encourage BCI or BI behavior. Over 4.62 billion people, 58% of the world's total population, were estimated to be active social media users as of January 2022 (Hootsuite & We Are Social, 2022). As part of these online communities, people share opinions and information at an unprecedented pace and volume (McCormick et al., 2017). Researchers, media practitioners, and policy makers should work together to identify and analyze the motivations, cultures, or

rules of online communities that regulate interactions among users. Designing social media applications to mitigate BCI behavior can also be another solution. An example is found from the project by Fan and colleagues (2016), which presented a software solution ('Feelbook') with a range of suggestions for the design of social media to counter cyberbullying (Bowler et al., 2015), including those that can help create greater equity amongst users in terms of controlling power (design for personal empowerment), expressing emotion and care for others (design for empathy), and provoking self-assessment about the user's own online behavior and motivation (design for reflection). The application feature also included a reward system for positive online behavior (e.g., having no offensive post). These principles could also be applied to the design of social media platforms in order to create a nurturing environment and cultivate digital citizenship among the bystander public.

Future Research

The dissertation concludes by suggesting two agendas to develop this line of research. The first provides contextual depth to the current set of questions about when and how bystanders choose to intervene in negative and positive ways. Given that bystander behavior is largely subject to social influence (Latané & Darley, 1968), future investigations can extend understanding of how the BCI model is conditioned by content-related and context-related aspects of online incivility. For example, message attributes (e.g., argument strength) and audience responses (e.g., like/dislike ratings) associated with uncivil speech acts can generate two forms of social influence cues that may affect bystander responses: informational and normative social influence (Deutsch & Gerard, 1955). Informational social influence is based on the desire to make the best decision possible in a given situation, while normative social influence is based on the desire to belong to a group (Kaplan, 1989).

Research has suggested that informational social influence emanates when a message provides compelling arguments and supportive evidence, and that message recipients who are prone to this type of influence are more likely to change their own opinions and accept others' opinions (Price et al., 2006). If this is the case in the current research context, the likelihood of counter-intervention may increase in the case where bystanders perceive greater informational social influence from uncivil speech acts with high argument strength. Meanwhile, research has also shown that bystanders are subject to normative tones of online communities (e.g., being sensitive or tolerant to negative comments; Allison et al., 2019). If this effect is translated into an uncivil comment, a greater number of likes on it can increase ambiguity about its deviance, while a greater number of dislikes can confirm its problematic nature (Naab et al., 2018). Of interest from this new research will be examining how these two influence factors moderate the BCI process.

Another direction for future work is to identify the product of BCI behavior for bystanders themselves. Understanding what happens afterward for bystanders may help us predict their future behavior. Evidence has shown that successful intervening experiences create a positive effect, whereas unsuccessful or failed attempts at intervening had an unpleasant effect (Rosen et al., 1987). This suggests that bystanders with prior negative experiences of BCI behavior are likely to be demotivated to act the same way in the future. Given the intense and possibly ego-threatening effects of being a bystander (Krauss et al., 2017), bystanders may experience self-effects (i.e., message effects on the cognitions, attitudes, and behaviors of the message senders themselves; Valkenburg, 2017) from their own behavior that may enhance counter-intervention responsibility and efficacy. Based on these ideas, the BCI model can be extended to include a further step after the behavioral

response. A test of this newly added step will demonstrate whether bystanders' self-perceptions are altered by their own behavior, and whether such changes then affect the likelihood of future counter-intervention. This new step could have important theoretical implications because uncivil online exchanges are incidents that happen repeatedly, providing constant opportunities for counter-intervention. To test this theoretical extension, future work can draw on the identity shift framework—the idea that online self-presentation results in self-transformation through self-perception of one's own behavior (Gonzales & Hancock, 2008). This work can also incorporate two important situational factors from the framework—publicness and feedback—as moderating variables to further explore: whether those who publicly perform BCI behavior, with supportive audience feedback, show greater transformative effects on their self-concepts that lead to becoming more determined counter-interveners. These areas of future research will provide crucial pieces in understanding BCI mechanisms and ultimately more effective BCI prevention.

Conclusion

The UN General Assembly proclaimed June 18 as the International Day for Countering Hate Speech, highlighting global concerns over the exponential rise of hate speech in whole societies around the world (United Nations, n.d.). It is not an overstatement to say that online incivility has become an increasingly prominent factor shaping today's sociopolitical climate, and that it has often constituted crisis situations in need of research and prevention efforts. Bystanders are ubiquitous in the online incivility dynamic and can have a profound effect on it, both negative and positive. According to the screening survey of the current research, 58% of respondents have ever acted to support online incivility, 98% have overlooked or tolerated it, while 92% have confronted it. The current dissertation

highlights the fluidity of online bystander roles and uncovers the decision-making processes leading to the different behavioral responses. To my knowledge, this research is among some of the first work that theorizes specific socio-psychological mechanisms that account for the negative implications of the bystander public and their role in fueling the spread of uncivil online discourse. This aspect of online incivility sensitizes researchers to the complexity of this problem, extending the focus of attention beyond perpetrators to include many invisible actors, which can be insightful for scholars studying the dark side of mediated public conversation.

References

- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. doi: 10.1016/0749-5978(91)90020-T
- Albayrak-Aydemir, N., & Gleibs, I. H. (2021). Measuring global bystander intervention and exploring its antecedents for helping refugees. *British Journal of Psychology*, 112(2), 519-548. doi: 10.1111/bjop.12474
- Aliberti, M., Jenkins, L., & Monk, M. (2022). Predictors of cyberbystander intervention among adolescents. *Journal of Interpersonal Violence*, 38(9-10), 6454-6479. doi: 10.1177/08862605221132784
- Allison, K. R., & Bussey, K. (2016). Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, 65, 183-194. doi: 10.1016/j.chilyouth.2016.03.026
- Allison, K. R., Bussey, K., & Sweller, N. (2019). 'I'm going to hell for laughing at this' norms, humour, and the neutralisation of aggression in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-25. doi: 10.1145/3359254
- Amadeu Antonio Foundation. (n.d.). *Civic.net - Aktiv gegen Hass im Netz*.
<https://www.amadeu-antonio-stiftung.de/projekte/civic-net-aktiv-gegen-hass-im-netz/>

- Antheunis, M. L., Valkenburg, P. M., & Peter, J. (2010). Getting acquainted through social network sites: Testing a model of online uncertainty reduction and social attraction. *Computers in Human Behavior*, 26(1), 100-109. doi: 10.1016/j.chb.2009.07.005
- Anti-Defamation League. (2020, June 9). *Online hate and harassment report: The American experience 2020*. ADL. <https://www.adl.org/online-hate-2020>
- Anti-Defamation League. (2021, March 22). *Online hate and harassment: The American experience 2021*. ADL. <https://www.adl.org/online-hate-2021>
- Anti-Defamation League. (2022, June 20). *Online hate and harassment: The American experience 2022*. <https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2022>
- Antoci, A., Bonelli, L., Paglieri, F., Reggiani, T., & Sabatini, F. (2018, February 6). *Civility and trust in social media*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3117274
- Badalge, K. N. (2017, May 25). *Our phones make us feel like activists, but they're actually turning us into bystanders*. Quartz. <https://qz.com/991167/our-phones-make-us-feel-like-social-media-activists-but-theyre-actually-turning-us-into-bystanders>
- Bandura, A. (1973). *Aggression: A social learning analysis*. Prentice-Hall. doi: 10.2307/1227918
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall. <https://psycnet.apa.org/record/1985-98423-000>
- Bandura, A. (1997). *Self-Efficacy: The exercise of control*. Worth Publishers. <https://psycnet.apa.org/record/1997-08589-000>

- Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209. doi: 10.1207/s15327957pspr0303_3
- Bandurski, D. (2011, January 4). *The surrounding gaze*. China Media Project.
<https://chinamediaproject.org/2011/01/04/the-surrounding-gaze-%E5%9B%B4%E8%A7%82/>
- Banyard, V. L. (2008). Measurement and correlates of prosocial bystander behavior: The case of interpersonal violence. *Violence and Victims*, 23(1), 83-97. doi: 10.1891/0886-6708.23.1.83
- Banyard, V. L., Moschella, E., Grych, J., & Jouriles, E. (2019). What happened next? New measures of consequences of bystander actions to prevent interpersonal violence. *Psychology of Violence*, 9(6), 664-674. doi:10.1037/vio0000229
- Barbeau, K., Boileau, K., Sarr, F., & Smith, K. (2019). Path analysis in Mplus: A tutorial using a conceptual model of psychological and behavioral antecedents of bulimic symptoms in young adults. *The Quantitative Methods for Psychology*, 15(1), 38-53. doi: 10.20982/tqmp.15.1.p038
- Barchia, K., & Bussey, K. (2011). Individual and collective social cognitive influences on peer aggression: Exploring the contribution of aggression efficacy, moral disengagement, and collective efficacy. *Aggressive Behavior*, 37(2), 107-120. doi: 10.1002/ab.20375
- Bard, M. K. T. (2006). *The relation between history of abuse, personal characteristics, and sexual offense characteristics and recidivism among a sample of juveniles who have sexually offended*. <https://www.proquest.com/dissertations-theses/relation-between-history-abuse-personal/docview/305260193/se-2>

- Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology*, 23(1), 37-51. doi: 10.1002/casp.2137
- 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. doi: 10.3389/fpsyg.2018.00799
- Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites: An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, 31, 259-271. doi: 10.1016/j.chb.2013.10.036
- Bastiaensens, S., Pabian, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2016). From normative influence to social pressure: How relevant others affect whether bystanders join in cyberbullying. *Social Development*, 25(1), 193-211. doi: 10.1111/sode.12134
- Bates, L. (2014, July 17). #JadaPose: The online ridiculing of a teen victim is part of a sickening trend. The Guardian. <https://www.theguardian.com/lifeandstyle/womens-blog/2014/jul/17/jadapose-online-ridiculing-rape-victims-sickening-trend>
- Beavon, E., Jenkins, L., Bradley, M., & Verma, K. (2022). Understanding adolescents' bystander intervention in cyberbullying. *Contemporary School Psychology*, 1-12. doi:10.1007/s40688-022-00437-1.
- Beckert, J., & Ziegele, M. (2020). The effects of personality traits and situational factors on the deliberativeness and civility of user comments on news websites. *International*

Journal of Communication, 14, 3924-3945.

<https://ijoc.org/index.php/ijoc/article/view/13422>

- Bennett, S., Banyard, V. L., & Garnhart, L. (2016). To act or not to act, that is the question? Barriers and facilitators of bystander intervention. *Journal of Interpersonal Violence*, 29(3), 476-496. doi: 10.1177/0886260513505210
- Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41(S1), 3-33. doi: 10.1111/pops.12670
- Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived as justified. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 22-31. doi: 10.1609/icwsm.v12i1.15036
- Borah, P. (2014). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research*, 41(6), 809-827. doi: 10.1177/0093650212449353
- Borah, P., Keib, K., Trude, B., Binford, M., Irom, B., & Himmelboim, I. (2022). “You are a disgrace and traitor to our country”: Incivility against “The Squad” on Twitter. *Internet Research*, 32(5), 1646-1661. doi: 10.1108/INTR-06-2021-0363
- Bowler, L., Knobel, C., & Mattern, E. (2015). From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media. *Journal of the Association for Information Science and Technology*, 55(5), 1274-1293. doi: 10.1002/asi.23270
- Bradshaw, C. P. (2015). Translating research to practice in bullying prevention. *American Psychologist*, 70(4), 322-332. doi: 10.1037/a0039114

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. doi: 10.1191/1478088706qp063oa
- Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57-71). American Psychological Association. doi: 10.1037/13620-004
- Brody, N. (2021). Bystander intervention in cyberbullying and online harassment: The role of expectancy violations. *International Journal of Communication*, 15, 647-667. <https://ijoc.org/index.php/ijoc/article/view/14169>
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230-258. doi: 10.1177/0049124192021002005
- Burn, S. M. (2009). A situational model of sexual assault prevention through bystander intervention. *Sex Roles*, 60(11-12), 779-792. doi: 10.1007/s11199-008-9581-5
- Burns, S., Maycock, B., Cross, D., & Brown, G. (2008). The power of peers: Why some students bully others to conform. *Qualitative Health Research*, 18(12), 1704-1716. doi: 10.1177/1049732308325865
- Bussey, K., Fitzpatrick, S., & Raman, A. (2015). The role of moral disengagement and self-efficacy in cyberbullying. *Journal of School Violence*, 14(1), 30-46. doi: 10.1080/15388220.2014.954045
- Bussey, K., Luo, A., Fitzpatrick, S., & Allison, K. (2020). Defending victims of cyberbullying: The role of self-efficacy and moral disengagement. *Journal of School Psychology*, 78, 1-12. doi: 10.1016/j.jsp.2019.11.006

- Cao, B., & Lin, W.-Y. (2015). How do victims react to cyberbullying on social networking sites? The influence of previous cyberbullying victimization experiences. *Computers in Human Behavior*, 52(4), 458-465. doi: 10.1016/j.chb.2015.06.009
- Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, 40, 193-236. doi: 10.1016/S0065-2601(07)00004-4
- Chan, T. K., Cheung, C. M., Benbasat, I., Xiao, B., & Lee, Z. W. Y. (2022). Bystanders join in cyberbullying on social networking sites: The deindividuation and moral disengagement perspectives. *Information Systems Research*, 34(3), 1-19. doi: 10.1287/isre.2022.1161
- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.
- Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Springer. doi: 10.1386/ajr.41.1.142_5
- Chen, G. M., & Lu, S. (2017). Online political discourse: Exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media*, 61(1), 108-125. doi: 10.1080/08838151.2016.1273922
- Chen, H., Huang, Q., & Jiang, M. (2022). Empowering Chinese college students to prevent sexual assault in post-MeToo Era: An empirical study of the bystander intervention approach. *Journal of Interpersonal Violence*, 37(1-2), NP449-NP472. doi: 10.1177/088626052091751

- Chen, L., Liu, X., & Tang, H. (2023). The interactive effects of parental mediation strategies in preventing cyberbullying on social media. *Psychology Research and Behavior Management, 16*, 1009-1022. doi: 10.2147/PRBM.S386968
- Choi, E., & Park, N. (2021). Can online education programs solve the cyberbullying problem? Educating South Korean elementary students in the COVID-19 ERA. *Sustainability, 13*(20), 11211. doi: 10.3390/su132011211
- Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy–altruism relationship: When one into one equals Oneness. *Journal of Personality and Social Psychology, 73*(3), 481-494. doi: 10.1037/0022-3514.73.3.481
- Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591-621. doi: 10.1146/annurev.psych.55.090902.142015
- Clark, M., & Bussey, K. (2020). The role of self-efficacy in defending cyberbullying victims. *Computers in Human Behavior, 109*. doi: 10.1016/j.chb.2020.106340
- Clark, R. D., & Word, L. E. (1974). Where is the apathetic bystander? Situational characteristics of the emergency. *Journal of Personality and Social Psychology, 29*(3), 279-287. doi: 10.1037/h0036000
- Clementson, D. E. (2019). Why won't you answer the question? Mass-mediated deception detection after journalists' accusations of politicians' evasion. *Journal of Communication, 69*(6), 674-695. doi: 10.1093/joc/jqz036

- Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4), 658-679. doi: 10.1111/jcom.12104
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Taylor & Francis Ltd. doi: 10.4324/9780203774441
- Compton, L., Campbell, M. A., & Mergler, A. (2014). Teacher, parent and student perceptions of the motives of cyberbullies. *Social Psychology of Education*, 17, 383-400. doi: 10.1007/s11218-014-9254-x
- Council of Europe. (n.d.). *No hate speech youth campaign website - no hate speech youth campaign*. No Hate Speech Youth Campaign. <https://www.coe.int/en/web/no-hate-campaign>
- Curtis, J. M., & Curtis, M. J. (1993). Factors related to susceptibility and recruitment by cults. *Psychological Reports*, 73, 451-460. doi: 10.2466/pr0.1993.73.2.451
- Dang, J., & Liu, L. (2020). Me and others around: The roles of personal and social norms in chinese adolescent bystanders' responses toward cyberbullying. *Journal of Interpersonal Violence*, 37(9-10). doi: 10.1177/0886260520967128
- Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4), 377-383. doi: 10.1037/h0025589
- De Oliveira, W. A., Caravita, S. C., Colombo, B., Donghi, E., Da Silva, J. L., & Silva, M. A. I. (2019). Self-justification processes related to bullying among Brazilian

- adolescents: A mixed methods study. *Frontiers in Psychology*, 10, 1086. doi: 10.3389/fpsyg.2019.01086
- DeSmet, A., Veldeman, C., Poels, K., Bastiaensens, S., Van f, K., Vandebosch, H., & De Bourdeaudhuij, I. (2014). Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents. *Cyberpsychology, Behavior, and Social Networking*, 17(4), 207-215. doi: 10.1089/cyber.2013.0027
- DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, 57, 398-415. doi: 10.1016/j.chb.2015.12.051
- Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3), 629-636. doi: 10.1037/h0046408
- Dickter, C. L., & Newton, V. A. (2013). To confront or not to confront: Non-targets' evaluations of and responses to racist comments. *Journal of Applied Social Psychology*, 43(S2), 262-275. doi: 10.1111/jasp.12022
- DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018). Upstanding by design: Bystander intervention in cyberbullying. *Proceedings of the 2018 CHI conference on Human Factors in Computing Systems*, (pp. 1-12). doi: 10.1145/3173574.3173785
- DiFranzo, D., Choi, Y. H., Purington, A., Taft, J. G., Whitlock, J., & Bazarova, N. N. (2019). Social Media TestDrive: Real-world social media education for the next

- generation. *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-11). doi: 10.1145/3290605.3300533
- Dillon, K. P., & Bushman, B. J. (2015). Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, 45, 144-150. doi: 10.1016/j.chb.2014.12.009
- Dovidio, J. F., Piliavin, J. A., Gaertner, S. L., Schroeder, D. A., & Clark, R. D. III. (1991). The arousal: Cost-reward model and the process of intervention: A review of the evidence. In M. S. Clark (Ed.), *Prosocial behavior* (pp. 86-118). Sage.
- Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., & Penner, L. A. (2006). *The social psychology of prosocial behavior*. Lawrence Erlbaum Associates.
- El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behavior*, 3(6), 554-559. doi: 10.1038/s41562-019-0596-4
- Erjavec, K., & Kovačič, M. P. (2012). “You don't understand, this is a new war!” Analysis of hate speech in news web sites’ comments. *Mass Communication and Society*, 15(6), 899-920. doi: 10.1080/15205436.2011.619679
- Erreygers, S., Pabian, S., Vandebosch, H., & Baillien, E. (2016). Helping behavior among adolescent bystanders of cyberbullying: The role of impulsivity. *Learning and Individual Differences*, 48, 61-67. doi: 10.1016/j.lindif.2016.03.003
- Evans, C. B. R., & Smokowski, P. R. (2017). Negative bystander behavior in bullying dynamics: Assessing the impact of social capital deprivation and anti-social capital. *Child Psychiatry & Human Development*, 48(1), 120-135. doi: 10.1007/s10578-016-0657-0

- Evans, C., Smokowski, P. R., Rose, R. A., Mercado, M. C., & Marshall, K. J. (2018). Cumulative bullying experiences, adolescent behavioral and mental health, and academic achievement: An integrative model of perpetration, victimization, and bystander behavior. *Journal of Child and Family Studies*, 27, 2415-2428. doi: 10.1007/s10826-018-1078-4.
- Fan, M., Yu, L., & Bowler, L. (2016, May). Feelbook: A social media app for teens designed to foster positive online behavior and prevent cyberbullying. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1187-1192. doi.org/10.1145/2851581.2892398
- Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology*, 23, 75-110. doi: 10.1016/S0065-2601(08)60318-4
- Ferrás, S. D., & Selman, R. (2014). How students' perceptions of the school climate influence their choice to upstand, bystand, or join perpetrators of bullying. *Harvard Educational Review*, 84(2), 162-187. doi: 10.17763/haer.84.2.h488313410l651mm
- Ferreira, P. D. C., Veiga Simão, A. M., Paiva, A., & Ferreira, A. (2020). Responsive bystander behaviour in cyberbullying: A path through self-efficacy. *Behaviour & Information Technology*, 39(5), 511-524. doi: 10.1080/0144929X.2019.1602671
- Ferreira, P. D. C., Veiga Simao, A. M., Pereira, N. S., Paulino, P., & Oliveira, S. (2021). Online verbal aggression, social relationships, and self-efficacy beliefs. *New Media & Society*, 23(5), 960-981. doi: 10.1177/1461444820905531

- Festl, R. (2016). Perpetrators on the Internet: Analyzing individual and structural explanation factors of cyberbullying in the school context. *Computers in Human Behaviour*, 59, 237-248. doi: 10.1016/j.chb.2016.02.017
- Fichman, P., & Sharp, S. (2020). Successful trolling on Reddit: A comparison across subreddits in entertainment, health, politics, and religion. *Proceedings of the Association for Information Science and Technology*, 57(1). doi: 10.1002/pra2.333
- Finley, K. (2015, October 8). *A brief history of the end of the comments*. Wired.
<https://www.wired.com/2015/10/brief-history-of-the-demise-of-the-comments-timeline>
- Fischer, P., Greitemeyer, T., Pollozek, F., & Frey, D. (2006). The unresponsive bystander: Are bystanders more responsive in dangerous emergencies? *European Journal of Social Psychology*, 36(2), 267-278. doi: 10.1002/ejsp.297
- Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., ... & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137(4), 517-537. doi: 10.1037/a0023304
- Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2), 303-315. doi: 10.1086/209351
- Forsberg, C., Thornberg, R., & Samuelsson, M. (2014). Bystanders to bullying: Fourth- to Seventh-grade students' perspectives on their reactions. *Research Papers in Education*, 29(5), 557-576. doi: 10.1080/02671522.2013.878375

- Fredrick, S., Jenkins, L., & Ray, K. (2020). Dimensions of empathy and bystander intervention in bullying in elementary school. *Journal of School Psychology, 79*, 31-42. doi: 10.1016/j.jsp.2020.03.001
- Fujiwara, T., Kohyama, T., Andreu, J. M., & Ramirez, J. M. (1999). Justification of interpersonal aggression in Japanese, American, and Spanish students. *Aggressive Behavior, 25*(3), 185-195. doi: 10.1002/(SICI)1098-2337(1999)25:3<185::AID-AB3>3.0.CO;2-K
- Funk, J. B., Baldacci, H. B., Pasold, T., & Baumgardner, J. (2004). Violence exposure in real-life, video games, television, movies, and the Internet: Is there desensitization?. *Journal of Adolescence, 27*(1), 23-39. doi: 10.1016/j.adolescence.2003.10.005
- Gaertner, S. L., & Dovidio, J. F. (2000). Reducing intergroup bias. The common ingroup identity model. *Psychology Press*.
- Gagnon, T. (2013). The disinhibition of Reddit users. *Adele Richardson's Spring*.
- Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society, 23*(12), 3491-3508. doi: 10.1177/1461444820958123
- Gini, G., Albiero, P., Benelli, B. & Altoe, G. (2008). Determinants of adolescents' active defending and passive bystanding behavior in bullying. *Journal of Adolescence, 31*(1), 93-105. doi: 10.1016/j.adolescence.2007.05.002
- Gini, G., Pozzoli, T., & Bussey, K. (2015). The role of individual and collective moral disengagement in peer aggression and bystanding: A multilevel analysis. *Journal of Abnormal Child Psychology, 43*, 441-452. doi: 10.1007/s10802-014-9920-7

- Gollwitzer, M. (2009). Justice and revenge. In M. E. Oswald, S. Bieneck, & J. Hupfeld-Heinemann (Eds.), *Social psychology of punishment of crime* (pp. 137-156). Wiley.
- Gonzales, A. L., & Hancock, J. T. (2008). Identity shift in computer-mediated environments. *Media Psychology, 11*(2), 167-185. doi: 10.1080/15213260802023433
- González-Cabrera, J. M., León-Mejía, A., Machimbarrena, J. M., Balea, A., & Calvete, E. (2019). Psychometric properties of the cyberbullying triangulation questionnaire: A prevalence analysis through seven roles. *Scandinavian Journal of Psychology, 60*(2), 160-168. doi: 10.1111/sjop.12518
- Goodman, E. (2013). *Online comment moderation: Emerging best practices*. World Editors Forum. http://netino.fr/wp-content/uploads/2013/10/WAN-IFRA_Online_Commenting.pdf
- Graf, J., Erba, J., & Harn, R.-W. (2017). The role of civility and anonymity on perceptions of online comments. *Mass Communication and Society, 20*(4), 526-549. doi: 10.1080/15205436.2016.1274763
- Gruzd, A., Mai, P., & Vahedi, Z. (2020). Studying anti-social behaviour on Reddit with Communalytic. *Advance Preprint*. doi: 10.31124/advance.12453749.v1
- Gutzwiller-Helfenfinger, E., & Perren, S. (2021). Tempted to join in or not? Moral temptation and self-reported behaviour in bullying situations. *British Journal of Developmental Psychology, 39*(1), 98-124. doi: 10.1111/bjdp.12348
- Ham, L. S., Wiersma-Mosley, J. D., Wolkowicz, N. R., Jozkowski, K. N., Bridges, A. J., & Melkonian, A. J. (2019). Effect of alcohol intoxication on bystander intervention in a vignette depiction of sexual assault. *Journal of Studies on Alcohol and Drugs, 80*(2), 252-260. doi:10.15288/jsad.2019.80.252

- Han, B., & Bowen, J. P. (2021). The weiguan culture phenomenon in Chinese online activism. *Proceedings of EVA London 2021*, 129-136.
doi:10.14236/ewic/EVA2021.22
- Hara, H. (2002). Justifications for bullying among Japanese schoolchildren. *Asian Journal of Social Psychology*, 5(3), 197-204. doi: 10.1111/1467-839X.00104
- Harriman, N., Shortland, N., Su, M., Cote, T., Testa, M. A., & Savoia, E. (2020). Youth exposure to hate in the online space: An exploratory analysis. *International Journal of Environmental Research and Public Health*, 17(22), 8531. doi: 10.3390/ijerph17228531
- Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The balanced inventory of desirable responding short form (BIDR-16). *Sage Open*, 5(4). doi: 10.1177/2158244015621113
- Hassan, S., Yacob, M. I., Nguyen, T., & Zambri, S. (2018). Should I intervene? The case of cyberbullying on celebrities from the perspective of the bystanders. *Proceedings of the 9th Knowledge Management International Conference (KMICe)*, 382-387.
- Hayashi, Y., & Tahmasbi, N. (2020). Decision-making process underlying bystanders' helping cyberbullying victims: A behavioral economic analysis of role of social discounting. *Computers in Human Behavior*, 104, 106-157. doi: 10.1016/j.chb.2019.106157
- Hayashi, Y., & Tahmasbi, N. (2021). Psychological predictors of bystanders' intention to help cyberbullying victims among college students: An application of theory of planned behavior. *Journal of Interpersonal Violence*, 37(13-14), 1-25. doi: 10.1177/0886260521992158

- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4), 408-420. doi: 10.1080/03637750903310360
- Hayes, A. F. (2018). Partial, conditional, and moderated moderated mediation: Quantification, inference, and interpretation. *Communication Monographs*, 85(1), 4-40. doi: 10.1080/03637751.2017.1352100
- Hayes, A. F., & Preacher, K. J. (2013). Conditional process modeling: Using structural equation modeling to examine contingent causal processes. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 219-266). IAP Information Age Publishing.
- Hoefnagels, C., & Zwikker, M. (2001). The bystander dilemma and child abuse: Extending the Latane and Darley model to domestic violence. *Journal of Applied Social Psychology*, 31(6), 1158-1183. doi: 10.1111/j.1559-1816.2001.tb02668.x
- Holbert, R. L., & Stephenson, M. T. (2007). Commentary on the uses and misuses of structural equation modeling in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of advanced data analysis methods for communication research* (pp. 185-218). Sage.
- Holfeld, B. (2014). Perceptions and attributions of bystanders to cyber bullying. *Computers in Human Behavior*, 38, 1-7. doi:10.1016/j.chb.2014.05.012
- Holloway, I., & Todres, L. (2003). The status of method: Flexibility, consistency, and coherence. *Qualitative Research*, 3(3), 345-357. doi: 10.1177/1468794103033004
- Hong, Y. J., & Lee, K. (2022). Adolescent bystanders' moral emotions in cyberbullying. *School Psychology International*, 43(3), 271-295. doi: 10.1177/01430343221088186

- Hootsuite & We Are Social. (2022). *Digital 2022: Another year of bumper growth*.
<https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. doi: 10.1080/10705519909540118
- Hudson, J. M., & Bruckman, A. S. (2004). The bystander effect: A lens for understanding patterns of participation. *The Journal of the Learning Sciences*, 13(2), 165-195. doi: 10.1207/s15327809jls1302_2
- Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, 58(4), 621-633. doi: 10.1080/08838151.2014.966365
- Hwang, H., Kim, Y., & Kim, Y. (2016). Influence of discussion incivility on deliberation: An examination of the mediating role of moral indignation. *Communication Research*, 45(2), 213-240. doi: 10.1177/0093650215616861
- I Am Here International.com. (n.d.). <https://iamhereinternational.com/>
- Jedicke, P. (2020, October 27). *As hate speech online increases, so has the resistance – DW* – 10/27/2020. dw.com. <https://www.dw.com/en/as-hate-speech-online-increases-so-has-the-resistance/a-55411555>

- Jenkins, L. N., & Nickerson, A. B. (2017). Bullying participant roles and gender as predictors of bystander intervention. *Aggressive Behavior*, 43(3), 281-290. doi: 10.1002/ab.21688
- Jenkins, L. N., Fredrick, S. S., & King, J. A. (2023). Testing the bystander intervention model in cyberbullying across students of color and white students. *International Journal of Bullying Prevention*, 1-10. doi: 10.1007/s42380-023-00162-0
- Jeyagobi, S., Munusamy, S., Kamaluddin, M. R., Ahmad Badayai, A. R., & Kumar, J. (2022). Factors influencing negative cyber-bystander behavior: A systematic literature review. *Frontiers in Public Health*, 10, 3445. doi: 10.3389/fpubh.2022.965017
- John, A., Glendenning, A. C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., Lloyd, K., & Hawton, K. (2018). Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *Journal of Medical Internet Research*, 20(4). doi: 10.2196/jmir.9044
- Jones, L. M., Mitchell, K. J., & Finkelhor, D. (2013). Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010). *Psychology of Violence*, 3(1), 53-69. doi: 10.1037/a0030309
- Jones, L. M., Mitchell, K. J., & Turner, H. A. (2015). Victim reports of bystander reactions to in-person and online peer harassment: A national survey of adolescents. *Journal of Youth and Adolescence*, 44(12), 2308-2320. doi: 10.1007/s10964-015-0342-9
- Joyce, L., Craig, W., & Johnson, M. (2015). *Young Canadians' experiences with electronic bullying*. MediaSmarts. <http://mediasmarts.ca/sites/mediasmarts/files/publication-report/full/young-canadians-electronic-bullying.pdf>

- Jubany, O. (2015). Backgrounds, experiences, and responses to online hate speech: An ethnographic multi-sited analysis. *Proceedings of the 2nd Annual International Conference on Social Science and Contemporary Humanity Development*. doi: 10.2991/sschd-16.2016.143
- Kaplan, M. F. (1989). Task, situational, and personal determinants of influence processes in group decision making. In E. J. Lawler (Ed.), *Advances in group processes* (Vol. 6, pp. 87-105). JAI Press.
- Kazerooni, F., Taylor, S. H., Bazarova, N. N., & Whitlock, J. (2018). Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication*, 23(3), 146-162. doi: 10.1093/jcmc/zmy005
- Kenny, D. A. (2015). *Measuring model fit*. <https://davidakenny.net/cm/fit.htm>
- Kim, Y., & Gonzales, A. L. (2022). When we tolerate online incivility: Dual-Process effects of argument strength and heuristic cues in uncivil user comments. *Computers in Human Behavior*, 131, 107235. doi: 10.1016/j.chb.2022.107235
- King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research* (pp. 257-270). doi: 10.4135/9781446280119
- Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Koehler, C., & Weber, M. (2018). “Do I really need to help?!” Perceived severity of cyberbullying, victim blaming, and bystanders’ willingness to help the victim.

- Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 12(4). doi: 10.5817/CP2018-4-4
- Koelsch, L. E., Brown, A. L., & Boisen, L. (2012). Bystander perceptions: Implications for university sexual assault prevention programs. *Violence and Victims*, 27(4), 563-579. doi: 10.1891/0886-6708.27.4.563
- König, A., Gollwitzer, M., & Steffgen, G. (2010). Cyberbullying as an act of revenge? *Journal of Psychologists and Counsellors in Schools*, 20(2), 210-224. doi: 10.1375/ajgc.20.2.210
- Konnikova, M. (2015, October 21). *How the Internet has changed bullying*. New Yorker. <https://www.newyorker.com/science/maria-konnikova/how-the-internet-has-changed-bullying>.
- Konrath, S. H., O'Brien, E. H., & Hsing, C. (2011). Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review*, 15(2), 180-198. doi: 10.1177/1088868310377395
- Kozubal, M., Szuster, A., & Barlińska, J. (2019). Cyberbystanders, affective empathy and social norms. *Studia Psychologica*, 61(2), 120-131. doi: 10.21909/sp.2019.02.777
- Krauss, A., Jouriles, E. N., Yule, K., Grych, J. H., Sargent, K. S., & Banyard, V. L. (2017). Adverse consequences to assisting victims of campus violence: Initial investigations among college students. *Journal of Interpersonal Violence*, 36(3-4), 1607-1624. doi: 10.1177/0886260517746944
- Kruger, A. C., Harris, J., Zabek, F., Grant, J. C., & Meyers, J. (2022). Black adolescents' perceptions of transgressor, victim, and bystander: Thinking through a sexual harassment dilemma. *Journal of Social Issues*. doi: 10.1111/josi.12518

- Kwon, K. H., & Cho, D. (2016). Swearing effects on citizen-to-citizen commenting online. *Social Science Computer Review*, 35(1), 84-102. doi: 10.1177/0894439315602664
- Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, 10(3), 215-221. doi: 10.1037/h0026570
- Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Appleton-Century-Crofts.
- Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2), 308-324. doi: 10.1037/0033-2909.89.2.308
- Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology*, 5(2), 189-202. doi: 10.1016/0022-1031(69)90046-8
- Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin*, 36(8), 1115-1129. doi: 10.1177/0146167210376391
- Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). 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(4), 555-579. doi: 10.5771/2192-4007-2018-4-555
- Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self-categorization and bystander non-intervention: Two experimental studies. *Journal of Applied Social Psychology*, 32(7), 1452-1463. doi:10.1111/j.1559-1816.2002.tb01446.x

- Levine, M., & Manning, R. (2013). Social identity, group processes, and helping in emergencies. *European Review of Social Psychology*, 24(1), 225-251. doi: 10.1080/10463283.2014.892318
- Liu, C.-H., Yin, X.-R., & Huang, P.-S. (2021). Cyberbullying: Effect of emergency perception on the helping tendencies of bystanders. *Telematics and Informatics*, 62. doi: 10.1016/j.tele.2021.101627
- Lodge, J., & Frydenberg, E. (2005). The role of peer bystanders in school bullying: Positive steps toward promoting peaceful schools. *Theory into Practice*, 44, 329-336.
- Macaulay, P. J., Boulton, M. J., & Betts, L. R. (2019). Comparing early adolescents' positive bystander responses to cyberbullying and traditional bullying: The impact of severity and gender. *Journal of Technology in Behavioral Science*, 4(3), 253-261. doi: 10.1007/s41347-018-0082-2
- Macaulay, P. J., Betts, L. R., Stiller, J., & Kellezi, B. (2021). 'The more public it is, the more severe it is': Teachers' perceptions on the roles of publicity and severity in cyberbullying. *Research Papers in Education*, 36(6), 726-753. doi: 10.1080/02671522.2020.1767183
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593-614. doi: 10.1146/annurev.psych.58.110405.085542
- Macklem, G. L., & Macklem, G. L. (2003). The role of bystanders in bullying episodes. *Bullying and Teasing: Social Power in Children's Groups*, 85-92. https://link.springer.com/chapter/10.1007/978-1-4757-3797-4_8

- Magellan Health. (2012, May 9). *Beyond the bully and the bullied: Bullying also impacts the mental health of the bystanders*. <https://ir.magellanhealth.com/news-releases/news-release-details/beyond-bully-and-bullied-bullying-also-impacts-mental-health>
- Marchal, N. (2020). The polarizing potential of intergroup affect in online political discussions: Evidence from Reddit r/politics. *SSRN*. doi: 10.2139/ssrn.3671497
- Marwick, A. E. (2021). Morally motivated networked harassment as normative reinforcement. *Social Media + Society*, 7(2), 1-13. doi: 10.1177/205630512110213788
- McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research*, 46(3), 390-421. doi: 10.1177/0049124115605339
- McPhate, M. (2016, April 18). *Teenager is accused of live-streaming a friend's rape on Periscope*. The New York Times. <https://www.nytimes.com/2016/04/19/us/periscope-rape-case-columbus-ohio-video-livestreaming.html>
- Menesini, E., Sanchez, V., Fonzi, A., Ortega, R., Costabile, A., & Lo Feudo, G. (2003). Moral emotions and bullying: A cross-national comparison of differences between bullies, victims and outsiders. *Aggressive Behavior*, 29(6), 515-530. doi: 10.1002/ab.10060
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage.
- Moschella, E. A., Bennett, S., & Banyard, V. L. (2016). Beyond the situational model: Bystander action consequences to intervening in situations involving sexual

- violence. *Journal of Interpersonal Violence*, 33(20), 3211-3231. doi: 10.1177/0886260516635319
- Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, 67(4), 586-609. doi: 10.1111/jcom.12312
- Mulder, R., Pouwelse, M., Lodewijkx, H., Bos, A. E. R., & van Dam, K. (2016). Predictors of antisocial and prosocial behaviour of bystanders in workplace mobbing. *Journal of Community & Applied Social Psychology*, 26(3), 207-220. doi: 10.1002/casp.2244
- Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, 20(2), 777-795. doi: 10.1177/1461444816670923
- Narula, S. K. (2017, April 10). *United Airlines violently dragged a passenger off an overbooked plane—then apologized for “having to re-accommodate” him*. Quartz. <https://qz.com/954405/video-a-united-airlines-passenger-is-violently-dragged-off-an-overbooked-plane>
- Ng, E. W., & Detenber, B. H. (2005). The impact of synchronicity and civility in online political discussions on perceptions and intentions to participate. *Journal of Computer-Mediated Communication*, 10(3). doi: 10.1111/j.1083-6101.2005.tb00252.x
- NICD (National Institute for Civil Discourse at the University of Arizona). (n.d.). *Initiative to revive civility*. <https://crowdfund.arizona.edu/project/7305>

- Nickerson, A. B., Aloe, A. M., Livingston, J. A., & Feeley, T. H. (2014). Measurement of the bystander intervention model for bullying and sexual harassment. *Journal of Adolescence*, 37(4), 391-400. doi: 10.1016/j.adolescence.2014.03.003
- Nickerson, A. B., Manges, M. E., Bellavia, G. M., Livingston, J. A., Jenkins, L. N., & Feeley, T. H. (2022). Bystander intervention in bullying and sexual harassment: Role of personal and perceived peer norms. *International Journal of Bullying Prevention*, 5, 135-150. doi: 10.1007/s42380-022-00125-x
- Nithyanand, R., Schaffner, B., & Gill, P. (2017, August 14). *Measuring offensive speech in online political discourse*. FOCI '17.
<https://www.usenix.org/conference/foci17/workshop-program/presentation/nithyanand>
- Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P., & Zaki, J. (2016). Prosocial conformity: Prosocial norms generalize across behavior and empathy. *Personality and Social Psychology Bulletin*, 42(8), 1045-1062. doi: 10.1177/0146167216649932
- Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society*, 18(8), 1491-1507. doi: 10.1177/1461444814563519
- Obermaier, M., Schmuck, D., & Saleem, M. (2021). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, 25(9), 2339-2358. doi: 10.1177/14614448211017527

- Obermaier, M., Schmid, U. K., & Rieger, D. (2023). Too civil to care? How online hate speech against different social groups affects bystander intervention. *European Journal of Criminology*, 20(3), 817-833. doi: 10.1177/14773708231156328
- O'Connell, P., Pepler, D., & Craig, W. (1999). Peer involvement in bullying: Insights and challenges for intervention. *Journal of Adolescence*, 22(4), 437-452. doi: 10.1006/jado.1999.0238
- Orue, I., Fernández-González, L., Machimbarrena, J. M., González-Cabrera, J., & Calvete, E. (2023). Bidirectional relationships between cyberbystanders' roles, cyberbullying perpetration, and justification of violence. *Youth & Society*, 55(4), 611-629. doi: 10.1177/0044118X211053356
- Ouvrein, G., De Backer, C. J. S., & Vandebosch, H. (2018). Joining the clash or refusing to bash? Bystanders reactions to online celebrity bashing. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 12(4), 1-17. doi: 10.5817/CP2018-4-5
- Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, 20(9), 3400-3419. doi: 10.1177/1461444817749516
- Pabian, S., & Vandebosch, H. (2016). An investigation of short-term longitudinal associations between social anxiety and victimization and perpetration of traditional bullying and cyberbullying. *Journal of Youth and Adolescence*, 45(2), 328-339. doi: 10.1007/s10964-015-0259-3
- Pabian, S., Vandebosch, H., Poels, K., Van Cleemput, K., & Bastiaenssens, S. (2016). Exposure to cyberbullying as a bystander: An investigation of desensitization effects

- among early adolescents. *Computers in Human Behavior*, 62, 480-487. doi: 10.1016/j.chb.2016.04.022
- Palasinski, M. (2012). The roles of monitoring and cyberbystanders in reducing sexual abuse. *Computers in Human Behavior*, 28(6), 2014-2022. doi: 10.1016/j.chb.2012.05.020
- Panumaporn, J., Hongsanguansri, S., Atsariyasing, W., & Kiatrungrit, K. (2020). Bystanders' behaviours and associated factors in cyberbullying. *General Psychiatry*, 33(3). doi: 10.1136/gpsych-2019-100187
- Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2), 259-283. doi: 10.1177/1461444804041444
- Patterson, L. J., Allan, A., & Cross, D. (2016). Adolescent bystanders' perspectives of aggression in the online versus school environments. *Journal of Adolescence*, 49(1), 60-67. doi: 10.1016/j.adolescence.2016.02.003
- Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60(2), 307-317. doi: 10.1037/0022-3514.60.2.307
- Pepler, D., Mishna, F., Doucet, J., & Lameiro, M. (2021). Witnesses in cyberbullying: Roles and dilemmas. *Children & Schools*, 43(1), 45-53. doi: 10.1093/cs/cdaa027
- Petrosyan, A. (2022). *Internet usage penetration in the United States from 2019 to 2028* [Dataset]. Statista. <https://www.statista.com/statistics/590800/internet-usage-reach-usa/>

Pew Research Center. (2017a, February 23). 2. *How the public reacted on Facebook*.

<http://www.peoplepress.org/2017/02/23/how-the-public-reacted-on-facebook>

Pew Research Center (2017b, July 11). *Online Harassment 2017*.

<http://www.pewinternet.org/2017/07/11/online-harassment-2017>

Pew Research Center. (2019, June 19). *Public highly critical of state of political discourse in*

the U.S. <https://www.pewresearch.org/politics/2019/06/19/public-highly-critical-of-state-of-political-discourse-in-the-u-s/>

Pew Research Center. (2021, January 13). *The state of online harassment*.

https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/01/PI_2021.01.13_Online-Harassment_FINAL-1.pdf

Pew Research Center (2022a, May 9). *About a third of Asian Americans say they have*

changed their daily routine due to concerns over threats, attacks.

<https://www.pewresearch.org/fact-tank/2022/05/09/about-a-third-of-asian-americans-say-they-have-changed-their-daily-routine-due-to-concerns-over-threats-attacks/>

Pew Research Center (2022b, August 10). *Teens, social media and technology 2022*.

<https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/>

Pew Research Center (2022c, December 14). *Americans differ by party, ideology over the*

impact of social media on U.S. democracy. <https://www.pewresearch.org/fact-tank/2022/12/14/americans-differ-by-party-ideology-over-the-impact-of-social-media-on-u-s-democracy/>

- Polanco-Levicán, K., & Salvo-Garrido, S. (2021). Bystander roles in cyberbullying: A mini-review of who, how many, and why. *Frontiers in Psychology, 12*. doi: 10.3389/fpsyg.2021.676787
- Polanin, J. R., Espelage, D. L., Grotz, J. K., Ingram, K., Michaelson, L., Spinney, E., Valido, A., Sheikh, A. E., Torgal, C., & Robinson, L. (2022). A systematic review and meta-analysis of interventions to decrease cyberbullying perpetration and victimization. *Prevention Science, 23*(3), 439-454. doi: 10.1007/s11121-021-01259-y
- Polder-Verikie, S. E. (2012). Online responsibility: Bad samaritanism and the influence of internet mediation. *Science and Engineering Ethics, 18*(1), 117-142. doi: 10.1007/s11948-010-9253-z
- Pouwels, J. L., Lansu, T. A., & Cillessen, A. H. (2017). Adolescents' explicit and implicit evaluations of hypothetical and actual peers with different bullying participant roles. *Journal of Experimental Child Psychology, 159*, 219-241. doi: 10.1016/j.jecp.2017.02.008
- Pöyhönen, V., Juvonen, J., & Salmivalli, C. (2012). Standing up for the victim, siding with the bully or standing by? Bystander responses in bullying situations. *Social Development, 21*(4), 722-741. doi: 10.1111/j.1467-9507.2012.00662.x
- Pozzoli, T., & Gini, G. (2013). Why do bystanders of bullying help or not? A multidimensional model. *The Journal of Early Adolescence, 33*(3), 315-340. doi: 10.1177/0272431612440172
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments & Computers, 36*(4), 717-731. doi: 10.3758/BF03206553

- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1), 185-227. doi: 10.1080/00273170701341316
- Preacher, K. J., & Yaremych, H. E. (2022). Model selection in structural Equation Modeling. In R. H. Hoyle (Ed.) *Handbook of structural equation modeling* (2nd Ed) pp. 206-222. Guilford.
- Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. *Communication Theory*, 16(1), 47-74. doi: 10.1111/j.1468-2885.2006.00005.x
- Pronk, R. E., & Zimmer-Gembeck, M. J. (2010). It's 'mean', but what does it mean to adolescents? Relational aggression described by victims, aggressors, and their peers. *Journal of Adolescent Research*, 25(2), 175-204. doi: 10.1177/0743558409350504
- Quirk, R., & Campbell, M. (2015). On standby? A comparison of online and offline witnesses to bullying and their bystander behaviour. *Educational Psychology*, 35(4), 430-448. doi: 10.1080/01443410.2014.893556
- Redmiles, E. M., Kross, S., & Mazurek, M. L. (2019). How well do my results generalize? Comparing security and privacy survey results from MTurk, web, and telephone samples. *2019 IEEE Symposium on Security and Privacy (SP)*, 1326-1343. doi: 10.1109/sp.2019.00014
- Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet*. 13(3), 433-451. doi: 10.1002/poi3.257

- Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, 88(2), 307-321. doi: 10.1037/0033-2909.88.2.307
- Roland, E., & Idsøe, T. (2001). Aggression and bullying. *Aggressive Behavior*, 27(6), 446-462. doi: 10.1002/ab.1029
- Romera, E. M., Ortega-Ruiz, R., Runions, K., & Camacho, A. (2021). Bullying perpetration, moral disengagement and need for popularity: Examining reciprocal associations in adolescence. *Journal of Youth and Adolescence*, 50(10), 2021-2035. doi: 10.1007/s10964-021-01482-4
- Rosen, S., Mickler, S. E., & Collins, J. E., II. (1987). Reactions of would-be helpers whose offer of help is spurned. *Journal of Personality and Social Psychology*, 53(2), 288-297. doi: 10.1037/0022-3514.53.2.288
- Ross, J., Irani, I., Silberman, M. Six, Zaldivar, A., & Tomlinson, B. (2010, April 10-15). Who are the crowdworkers?: Shifting demographics in Amazon Mechanical Turk. *Proceedings of the 28th International Conference on Human Factors in Computing Systems*. doi: 10.1145/1753846.1753873
- Rudy, R., & Linz, D. G. (2008). Desensitization. *The International Encyclopedia of Communication*. doi: 10.1002/9781405186407.wbiecd015
- Runions, K., & Bak, M. (2015). Online moral disengagement, cyberbullying, and cyber-aggression. *Cyberpsychology, Behavior and Social Networking*, 18(7), 400-405. doi: 10.1089/cyber.2014.0670
- Sabatini, F., & Reggiani, T. (2019, March 26). *Locking horns: Online incivility is seen as the new normal, and it inhibits social trust*. The London School of Economics and

- Political Science. <https://blogs.lse.ac.uk/businessreview/2019/03/26/locking-horns-online-incivility-is-seen-as-the-new-normal-and-it-inhibits-social-trust/>
- Salem, A. A., Al-Huwailah, A. H., Abdelsattar, M., Al-Hamdan, N. A., Derar, E., Alazmi, S., Abu Al-Diyar, M., & Griffiths, M. D. (2023). Empathic skills training as a means of reducing cyberbullying among adolescents: An empirical evaluation. *International Journal of Environmental Research and Public Health*, 20(3), 1846. doi: 10.3390/ijerph20031846
- Salmivalli, C. (2014). Participant roles in bullying: How can peer bystanders be utilized in interventions?. *Theory into Practice*, 53(4), 286-292. doi: 10.1080/00405841.2014.947222
- Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1998). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*, 22(1), 1-15. doi:10.1002/(sici)1098-2337(1996)22:1<1::aid-ab1>3.0.co;2-t
- Salmivalli, C., Kaukiainen, A., Kaistaniemi, L., & Lagerspetz, K. (1999). Self-evaluated self-esteem, peer-evaluated self-esteem, and defensive egotism as predictors of adolescents' participation in bullying situations. *Personality and Social Psychology Bulletin*, 25(10), 1268-1278. doi: 10.1177/0146167299258008
- Salmivalli, C., Voeten, M., & Poskiparta, E. (2011). Bystanders matter: Associations between reinforcing, defending, and the frequency of bullying in classrooms. *Journal of Clinical Child & Adolescent Psychology*, 40(5), 668-676. doi: 10.1080/15374416.2011.597090

- Sarmiento, A., Herrera-López, M., & Zych, I. (2019). Is cyberbullying a group process? Online and offline bystanders of cyberbullying act as defenders, reinforcers and outsiders. *Computers in Human Behavior*, 99, 328-334. doi: 10.1016/j.chb.2019.05.037
- Schieb, C., & Preuss, M. (2016, June). *Governing hate speech by means of counterspeech on Facebook* [Conference presentation]. 66th ICA Annual Conference, Fukuoka, Japan.
- Schroeder, D. A., Penner, L. A., Dovidio, J. F., & Piliavin, J. A. (1995). *The psychology of helping and altruism: Problems and puzzles*. McGraw-Hill.
- Schwartz, S. H. (1977). Normative influences on altruism. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 221-279). Academic Press.
- SELMA. (2019). *Hacking online hate: Building an evidence base for educators*.
<https://hackinghate.eu/assets/documents/hacking-online-hate-research-report-1.pdf>
- Shatz, I. (2017). Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, 35(4), 537-549. doi: 10.1177/0894439316650163
- Shelton, M., Lo, K., & Nardi, B. (2015, March). Online media forums as separate social lives: A qualitative study of disclosure within and beyond Reddit. *iConference 2015 Proceedings*.
- Shen, Y., Yuan, L., & Xiong, X. (2022). Empathy and cyberbystander behavior: The role of moral disengagement. *Current Psychology*, 42, 18070-18079. doi: 10.1007/s12144-022-03015-z
- Shotland, R. L., & Huston, T. L. (1979). Emergencies: What are they and do they influence bystanders to intervene?. *Journal of Personality and Social Psychology*, 37(10), 1822-1834. doi: 10.1037/0022-3514.37.10.1822

- Siciliano, R. (2014, June 3). *Teens' online behavior can get them in trouble*. McAfee.
<https://www.mcafee.com/blogs/consumer/family-safety/teens-and-screens/>
- Smith, P. K. (2012). Cyberbullying and cyber aggression. In S. R. Jimerson, A. B. Nickerson, M. J. Mayer, & M. J. Furlong (Eds.), *Handbook of school violence and school safety: International research and practice* (pp. 93-103). Routledge.
- Snow, D. A., Zurcher, L. A., & Peters, R. (1981). Victory celebrations as theater: A dramaturgical approach to crowd behavior. *Symbolic Interaction*, 4(1), 21-42. doi: 10.1525/si.1981.4.1.21
- Solomontos-Kountouri, O., & Strohmeier, D. (2021). The need to belong as motive for (cyber) bullying and aggressive behavior among immigrant adolescents in Cyprus. *New Directions for Child and Adolescent Development*, 2021(177), 159-178. doi: 10.1002/cad.20418
- Song, J., & Oh, I. (2017). Investigation of the bystander effect in school bullying: Comparison of experiential, psychological and situational factors. *School Psychology International*, 38(3), 319-336. doi: 10.1177/0143034317699997
- Song, J., & Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. *Computers in Human Behavior*, 78, 273-282. doi: 10.1016/j.chb.2017.10.008
- Spadafora, N., Marini, Z. A., & Volk, A. A. (2020). Should I defend or should I go? An adaptive, qualitative examination of the personal costs and benefits associated with bullying intervention. *Canadian Journal of School Psychology*, 35(1), 23-40. doi: 10.1177/0829573518793752

- Steer, O. L., Betts, L. R., Baguley, T., & Binder, J. F. (2020). "I feel like everyone does it"- Adolescents' perceptions and awareness of the association between humour, banter, and cyberbullying. *Computers in Human Behavior*, 108. 106297. doi: 10.1016/j.chb.2020.106297
- Stopbullying.gov. (n.d.). *Bystanders are essential to bullying prevention and intervention*. https://www.stopbullying.gov/resources/research-resources/bystanders-are-essential#footnote3_3xqhahd
- Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.
- Sun, L. H., & Fichman, P. (2020). The collective trolling lifecycle. *Journal of the Association for Information Science and Technology*, 71(7), 770-783. doi: 10.1002/asi.24296
- Taylor, S. E., Peplau, L. A., & Sears, D. O. (2006). *Social psychology* (12th ed.). Pearson Education.
- Terry, R. (2020, October 29). *How to be an active bystander when you see casual racism*. The New York Times. <https://www.nytimes.com/2020/10/29/smarter-living/how-to-be-an-active-bystander-when-you-see-casual-racism.html?searchResultPosition=1>
- Thompson, D. F. (2014). Responsibility for failures of government: The problem of many hands. *The American Review of Public Administration*, 44(3), 259-273. doi: 10.1177/0275074014524013
- Thornberg, R. (2015). School bullying as a collective action: Stigma processes and identity struggling. *Children & Society*, 29(4), 310-320. doi: 10.1111/chso.12058

- Thornberg, R., & Jungert, T. (2013). Bystander behavior in bullying situations: Basic moral sensitivity, moral disengagement and defender self-efficacy. *Journal of Adolescence*, 36(3), 475-483. doi: 10.1016/j.adolescence.2013.02.003
- Troop-Gordon, W., Frosch, C. A., Totura, C. M. W., Bailey, A. N., Jackson, J. D., & Dvorak, R. D. (2019). Predicting the development of pro-bullying bystander behavior: A short-term longitudinal analysis. *Journal of School Psychology*, 77, 77-89. doi: 10.1016/j.jsp.2019.10.004
- Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.
- Turpin-Petrosino, C. (2002). Hateful sirens... who hears their song? An examination of student attitudes toward hate groups and affiliation potential. *Journal of Social Issues*, 58(2), 281-301. doi: 10.1111/1540-4560.00261
- United Nations. (n.d.). *International day for countering hate Speech, 18 June*.
<https://www.un.org/en/observances/countering-hate-speech>
- Valkenburg, P. M. (2017). Understanding self-effects in social media. *Human Communication Research*, 43(4), 477-490. doi: 10.1111/hcre.12113
- Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine “helping,” “joining in,” and “doing nothing” when witnessing cyberbullying. *Aggressive Behavior*, 40(5), 383-396. doi: 10.1002/ab.21534

- van der Ploeg, R., Kretschmer, T., Salmivalli, C., & Veenstra, R. (2017). Defending victims: What does it take to intervene in bullying and how is it rewarded by peers? *Journal of School Psychology, 65*, 1-10. doi: 10.1016/j.jsp.2017.06.002
- Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International Journal of Environmental Research and Public Health, 15*(9), 2030. doi: 10.3390/ijerph15092030
- Wachs, S., Wright, M. F., Sittichai, R., Singh, R., Biswal, R., Kim, E. M., Yang, S., Gámez-Guadix, M., Almendros, C., Flora, K., Daskalou, V., & 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*(20), 3992. doi: 10.3390/ijerph16203992
- Wang, S. (2021). Standing up or standing by: Bystander intervention in cyberbullying on social media. *New Media & Society, 23*(6), 1379-1397. doi: 10.1177/1461444820902541
- Wang, S., & Kim, K. J. (2021). Effects of victimization experience, gender, and empathic distress on bystanders' intervening behavior in cyberbullying. *The Social Science Journal, 1*-10. doi: 10.1080/03623319.2020.1861826
- Warren, B. J. (2011). Two sides of the coin: The bully and the bullied. *Journal of Psychosocial Nursing and Mental Health Services, 49*(10), 22-29. doi: 10.3928/02793695-20110830-01

- Watson, B. R., Peng, Z., & Lewis, S. C. (2019). Who will intervene to save news comments? Deviance and social control in communities of news commenters. *New Media & Society*, 21(8), 1840-1858. doi: 10.1177/1461444819828328
- Weber Shandwick (2017). *Civility in America VII: The state of civility*.
https://www.webershandwick.com/uploads/news/files/Civility_in_America_the_State_of_Civility.pdf
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Eds.), *Handbook of structural equation modeling* (pp. 209-231). The Guilford Press.
- Williams, A. J., Banks, C. S., & Blake, J. J. (2018). High school bystanders motivation and response during bias-based bullying. *Psychology in the Schools*, 55(10), 1259-1273. doi: 10.1002/pits.22186
- WITH US Center for Bystander Intervention at Cal Poly. (2020). *National college student bystander intervention survey summary report*.
<https://www.withus.org/bystanderdatareports>
- Wong, R. Y., Cheung, C. M., Xiao, B., & Thatcher, J. B. (2021). Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research*, 32(2), 561-581. doi: 10.1287/isre.2020.0983
- Wright, P. (2020, December 16). *Abuse and bullying is everywhere online. You can help deal with it without making yourself a target*. ABC Everyday.
<https://www.abc.net.au/everyday/how-to-be-a-better-bystander-against-online-abuse-bullying/11482342>

- Xiao, B. S., & Wong, Y. M. (2013). Cyber-bullying among university students: An empirical investigation from the social cognitive perspective. *International Journal of Business and Information*, 8(1), 34-69.
- Xu, J. (2015). Online weiguan in web 2.0 China: Historical origins, characteristics, platforms and consequences. In G. Yang (Eds.), *China's contested Internet* (pp. 257-282). NIAS Press of the University of Copenhagen. doi: 10536/DRO/DU:30091601
- Zhang, S., & Leidner, D. (2018). From improper to acceptable: How perpetrators neutralize workplace bullying behaviors in the cyber world. *Information & Management*, 55(7), 850-865. doi: 10.1016/j.im.2018.03.012
- Zhao, Y., Chu, X., & Rong, K. (2023). Cyberbullying experience and bystander behavior in cyberbullying incidents: The serial mediating roles of perceived incident severity and empathy. *Computers in Human Behavior*, 138, 107484. doi: 10.1016/j.chb.2022.107484
- Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, 22(5), 731-751. doi: 10.1177/1461444819870130

Appendix A

Interview Participant Recruitment Post

You are invited to join an interview study about online public discussion. USD \$15 will be offered as a participation reward. We are looking for interviewees who meet the following relevant qualifications to our study.

This study is for people:

- who are U.S. residents over the age of 18 – who have their own email address and can use email for a written interview
- who have experience reading user comments in online discussion platforms (e.g., Reddit, Twitter, Facebook, blogs, news websites, community websites, etc.)
- who have experience of reacting positively to toxic comments made by others, by supporting, defending, reinforcing, or assisting those comments

Note that we define what we mean by "toxic" comments as below:

What are "toxic" comments?

People discuss political and societal topics online and often engage in heated and conflicting discussions. In doing so, people make toxic comments that contain aggressive, impolite, and even hateful speech, especially when they express disagreement, doubt, criticism, refusal, and negative emotions against a person or group holding a particular political tendency (e.g., liberal, conservative) or social identity (e.g., gender, race/ethnicity, religion, disability, sexual orientation). Here are examples:

- **Name-calling:** Mean-spirited or insulting speech against a person or group (e.g., "You Chinese morons must go back to your country. No more Asian virus!" "Millennials are retards! That's why they can't do anything right.")
- **Profanity:** Crude, noxious, or vulgar speech against a person or group (e.g., "Stupid Liberals, their brains are full of shit!" "Yeah fuck Crenshaw I'll vote 3rd party before that neocon")
- **Pejorative speech:** Disparaging speech about the way in which a person or group communicates (e.g., "Damn, I am sick and tired of blacks throwing tantrums..." "Keep talking about feminism and I'll really give you something to cry about")
- **Antagonistic stereotype:** Speech that overgeneralizes the superiority or inferiority of a person or group (e.g., "Women aren't good for anything besides cooking and cleaning." "I'm not letting transgenders into my bathroom! What if they try to attack me?!" "Why are you so broke? I thought Jews were supposed to be good with money.")
- **Political threat:** Speech that threatens the democratic rights of a person or group (e.g., "Do autistic people really deserve to vote? Do they even know what it is?!" "How dare lazy millennials talk about labor rights?! I don't share their opinion at all!")
- **Violent threat:** Speech that announces aggressive action against a person or group (e.g., "These immigrants should be stoned and executed publicly." "Wouldn't the world be a better place if we just killed the Democrats?")
- **Sexual harassment:** Speech that makes obscene and erotic remarks against a person or group (e.g., "This stuck-up bitch has a terrible face but a damn nice body!" "Did she sleep with her boss? That's the only way she could have received that promotion...")

What are "positive reactions" to toxic comments?

We see people react positively to toxic comments by engaging in the following behaviors:

- Clicking the like button on a toxic comment
- Posting a comment or reply to agree with, support, or defend a toxic comment
- Posting a comment or reply to elaborate further on a toxic comment
- Responding with laughter to a toxic comment (e.g., sharing a laughing emoji)

- Sharing a toxic comment with more people
- Any other behavior that reinforces or assists a toxic comment

Pre-interview Screening Survey Questionnaire

Online Incivility Exposure Experience	Q. Have you ever seen people make toxic comments against a person or group (not you) online, where people discuss political and social topics (e.g., Twitter, Facebook, Reddit, blogs, news websites, community websites)? How often do you see such toxic comments during a typical week? 0 = <i>never</i> : 5 = <i>very often</i>
Bystander Counter- intervention Experience	Q. Have you ever reacted positively to the toxic comments you read, by engaging in any of the following behaviors? (Please select all that apply) ☐ Clicking the like button on a toxic comment ☐ Posting a comment or reply to agree with, support, or defend a toxic comment ☐ Posting a comment or reply to elaborate further on a toxic comment ☐ Responding with laughter to a toxic comment (e.g., sharing a laughing emoji) ☐ Sharing a toxic comment with more people ☐ Any other behavior that reinforces or assists a toxic comment (Please specify) Q. How often have you engaged in each behavior when you read toxic comments? 0 = <i>never</i> : 5 = <i>very often</i> Q. You reported that you have reacted positively to toxic comments you have read (which target someone else), by engaging in behaviors of supporting, defending, reinforcing, or assisting that comment(s). Can you recall and briefly describe (1) what the comment(s) said and (2) how you responded to it/them, in 3-5 sentences?
Online Incivility Perpetration Experience	Q. Personally, have you ever made your own toxic comment(s) against a person or group? If yes, how often do you make your own toxic comments during your time in online discussions? 0 = <i>never</i> : 5 = <i>very often</i>
Online Incivility Victimization Experience	Q. Personally, have you ever received a toxic comment(s) from others that targeted you? If yes, how often do you receive toxic comments from others during your time in online discussions? 0 = <i>never</i> : 5 = <i>very often</i>
Sociodemographic Information	Q. How old are you? Please enter your age in years (e.g., "23", "45"). [] Q. How would you describe your gender? 1 = <i>male</i> 2 = <i>female</i> 3 = <i>other</i> 4 = <i>prefer not to answer</i> Q. What is your political affiliation? 1 = <i>liberal/democrat</i> 2 = <i>conservative/republican</i> 3 = <i>moderate</i> 4 = <i>independent</i>

5 = *other*
6 = *no particular political orientation*
7 = *prefer not to answer*

Post-interview Survey Questionnaire

Online Discussion Platform	Q. Which digital platforms do you use to view or join online public discussions about political or social issues? (Please select all that apply) ☐ Twitter ☐ Facebook ☐ Reddit – please specify ☐ News website (comment sections) – please specify ☐ Online community – please specify ☐ Instagram ☐ TikTok ☐ YouTube ☐ Snapchat ☐ Quora ☐ Blog ☐ Other platforms (please specify) 1 = <i>Yes, I use this platform</i> 0 = <i>No, I do not use this platform</i>
Online Comment Reading/Posting	Q. Today people discuss political or social topics online by using various digital platforms like Twitter, Facebook, Reddit, blogs, community bulletin boards, the comment sections of news websites, etc. During your visit to these online discussions... Do you read others' comments? If yes, how often during a typical week? 0 = <i>never</i> : 5 = <i>very often</i> Do you post your own comments? If yes, how often during a typical week? 0 = <i>never</i> : 5 = <i>very often</i>
Sociodemographic Information	Q. What is your ethnic background? 1 = <i>White / Caucasian</i> 2 = <i>African-American</i> 3 = <i>Native-American</i> 4 = <i>Asian-Eastern</i> 5 = <i>Asian-Indian</i> 6 = <i>Hispanic</i> 7 = <i>mixed race</i> 8 = <i>other (please specify)</i> 9 = <i>prefer not to answer</i> Q. What is the highest level of education you have completed? 1 = <i>less than high school diploma or GED</i> 2 = <i>high school diploma or GED</i> 3 = <i>some college education</i> 4 = <i>bachelor's degree</i> 5 = <i>graduate or professional degree</i> 6 = <i>other</i> 7 = <i>prefer not to answer</i>

Appendix B

Survey Participant Recruitment Post

You are invited to a research study about toxic comments in online public discussions. This study consists of two parts: an eligibility screener and a series of surveys. If you are interested in this study, please participate in this screener to share your experiences related to toxic comments. Once you are deemed eligible for this study, you will receive an invitation to one to three surveys over the next few weeks. Note that you will receive \$2-6 depending on how many surveys you participate in.

This study is for people:

- who sometimes view or join online public discussions about political or social issues on digital platforms (e.g., Twitter, Facebook, Reddit, blogs, communities, news sites)
- who have had experience seeing and reacting to toxic comments made by others that target a person or group

Note that we define "toxic" comments as below.

What are "toxic" comments?

People discuss political and societal topics online and often engage in heated and conflicting discussions. In doing so, people make toxic comments that contain aggressive, impolite, and even hateful speech, especially when they express disagreement, doubt, criticism, refusal, and negative emotions against a person or group holding a particular political tendency (e.g., liberal, conservative) or social identity (e.g., gender, race/ethnicity, religion, disability, sexual orientation). Here are examples:

- **Name-calling:** Mean-spirited or insulting speech against a person or group (e.g., "You Chinese morons must go back to your country. No more Asian virus!" "Millennials are retards! That's why they can't do anything right.")
- **Profanity:** Crude, noxious, or vulgar speech against a person or group (e.g., "Stupid Liberals, their brains are full of shit!" "Yeah fuck Crenshaw I'll vote 3rd party before that neocon")
- **Pejorative speech:** Disparaging speech about the way in which a person or group communicates (e.g., "Damn, I am sick and tired of blacks throwing tantrums..." "Keep talking about feminism and I'll really give you something to cry about")
- **Antagonistic stereotype:** Speech that overgeneralizes the superiority or inferiority of a person or group (e.g., "Women aren't good for anything besides cooking and cleaning." "I'm not letting transgenders into my bathroom! What if they try to attack me?!" "Why are you so broke? I thought Jews were supposed to be good with money.")
- **Political threat:** Speech that threatens the democratic rights of a person or group (e.g., "Do autistic people really deserve to vote? Do they even know what it is?!" "How dare lazy millennials talk about labor rights?! I don't share their opinion at all")
- **Violent threat:** Speech that announces aggressive action against a person or group (e.g., "These immigrants should be stoned and executed publicly." "Wouldn't the world be a better place if we just killed the Democrats?")
- **Sexual harassment:** Speech that makes obscene and erotic remarks against a person or group (e.g., "This stuck-up bitch has a terrible face but a damn nice body!" "Did she sleep with her boss? That's the only way she could have received that promotion...")

Screening Survey Questionnaire

Online Discussion Activity	Q. How often do you usually view or join online public discussions about political or social issues in a typical week, by reading others' comments and/or posting your own comments? 1 = <i>never</i> : 7 = <i>everyday</i> Q. Which digital platforms do you use to view or join online public discussions about political or social issues? (Please select all that apply) ☐ Twitter ☐ Facebook ☐ Reddit – please specify ☐ News website (comment sections) – please specify ☐ Online community – please specify ☐ Instagram ☐ TikTok ☐ YouTube ☐ Snapchat ☐ Quora ☐ Blog ☐ Other platforms (please specify) 1 = <i>Yes, I use this platform</i> 0 = <i>No, I do not use this platform</i>
Online Incivility Exposure	Q. In a typical week, how often do you see toxic comments made by others that target a person or group (not you), in online public discussions about political or social issues? 1 = <i>never</i> : 7 = <i>always</i>
Online Bystander Behavior	<i>Bystander Counter-intervention</i> Q. When you see toxic comments made by others (against a person or group, not you), have you ever reacted favorably to those comments, despite their toxic language and tone? If yes, how often? <i>Note. Favorable reactions to toxic comments include the following behaviors that support, reinforce, assist, or defend these comments.</i> – clicking the like/upvote button on toxic comments – posting my own comment/reply to support, reinforce, assist, or defend toxic comments – posting emojis, gifs, or pictures to support, reinforce, assist, or defend toxic comments – sharing toxic comments with others who might also react favorably to them – other behaviors that support, reinforce, assist, or defend toxic comments 1 = <i>never</i> : 7 = <i>always</i> <i>Bystander Intervention</i> Q. When you see toxic comments made by others (against a person or group, not you), have you ever reacted unfavorably to those comments, due to their toxic language and tone? If yes, how often? <i>Note. Unfavorable reactions to toxic comments include the following behaviors that oppose, confront, correct, or stop those comments.</i> – clicking the dislike/downvote button on toxic comments – reporting or flagging toxic comments to signal platform moderators – posting my own comment/reply to oppose, confront, correct, or stop toxic

	comments – posting emojis, gifs, or pictures to oppose, confront, correct, or stop toxic comments – other behaviors that oppose, confront, correct, or stop toxic comments 1 = <i>never</i> : 7 = <i>always</i> <i>Bystander Non-intervention</i> Q. When you see toxic comments made by others (against a person or group, not you), have you ever remained silent? If yes, how often? <i>Note. Remaining silent to toxic comments includes the following behaviors that do not involve any direct/indirect and supportive/oppositional reaction to those comments</i> – reading toxic comments but ignoring them – looking at others' responses to toxic comments – continuing to scroll past toxic comments – leaving the discussion or platform where toxic comments exist – other behaviors that do not involve any reaction to toxic comments 1 = <i>never</i> : 7 = <i>always</i>
Online Incivility Victimization	Q. During your time in online discussions, have you ever personally received toxic comments from someone—for example, comments that are rude or mean, embarrass or make fun of you, or hurt your feeling? If yes, how often? 1 = <i>never</i> : 4 = <i>several times</i>
Online Incivility Perpetration	Q. During your time in online discussions, have you ever initiated toxic comments toward someone else—for example, comments that are rude or mean, embarrass or make fun of them, or hurt their feeling? If yes, how often? 1 = <i>never</i> : 4 = <i>several times</i>
Social Desirability Bias	Q. Please indicate the extent to which you agree or disagree with each of the following statements. - I have not always been honest with myself ® - I never regret my decisions - I am a completely rational person - I am very confident of my judgments - I sometimes tell lies if I have to ® - I never cover up my mistakes - There have been occasions when I have taken advantage of someone ® - I have said something bad about people behind his or her back ® - It's hard for me to shut off a disturbing thought ® 1 = <i>strongly disagree</i> : 7 = <i>strongly agree</i>
Sociodemographic Information	Q. How old are you? Please enter your age in years (e.g., "23", "45"). [] Q. How would you describe your gender? 1 = <i>male</i> 2 = <i>female</i> 3 = <i>other</i> 4 = <i>prefer not to answer</i> Q. What is your political affiliation?

-
- 1 = *liberal/democrat*
 - 2 = *conservative/republican*
 - 3 = *moderate*
 - 4 = *independent*
 - 5 = *other*
 - 6 = *no particular political orientation*
 - 7 = *prefer not to answer*

Q. What is your ethnic background?

- 1 = *White/Caucasian*
- 2 = *African-American*
- 3 = *Native-American*
- 4 = *Asian-Eastern*
- 5 = *Asian-Indian*
- 6 = *Hispanic*
- 7 = *mixed race*
- 8 = *other (please specify)*
- 9 = *prefer not to answer*

Q. What is the highest level of education you have completed?

- 1 = *less than high school diploma or GED*
 - 2 = *high school diploma or GED*
 - 3 = *some college education*
 - 4 = *bachelor's degree*
 - 5 = *graduate or professional degree*
 - 6 = *other*
 - 7 = *prefer not to answer*
-

Bystander Counter-intervention (BCI) Survey Questionnaire

Online Incivility Justification	Q. When you reacted favorably to the toxic comment(s), to what extent did you think the comment(s) were justifiable? Note. For example... - You thought the target deserved the comment(s) - You thought the comment(s) helped people vent their negative emotions against the target (e.g., anger, frustration) - You thought the comment(s) helped to get a point across to others - You thought the comment(s) contributed to increasing awareness about an important issue - You thought the comment(s) was worthwhile because it revealed a truth - You thought the comment(s) was problematic because it degraded the discussion quality - You thought the comment(s) was well-intended to correct or advise the target - You thought the comment was well within the bounds of freedom of expression 1 = <i>not at all justifiable</i> : 7 = <i>extremely justifiable</i>
------------------------------------	--

Benefit–Cost Perception of Counter- intervention	Q. Recall what you expected to occur as a positive consequence of your favorable reaction to the toxic comment(s). To what extent does each of the following statements apply to you at that time? a. I expected to feel connected and supported from like-minded people b. I expected my reaction to positively contribute to the discussion c. I expected my reaction to help convince other comment-readers to reconsider or change their opinions about the target 1 = <i>not at all applicable</i> : 7 = <i>extremely applicable</i> Q. Recall what you expected to occur as a negative consequence of your favorable reaction to the toxic comment(s). To what extent does each of the following statements apply to you at that time? a. I expected to receive negative evaluations from other comment-readers [e.g., dislikes, harsh replies] and be in conflict with them b. I expected my reaction could be moderated or my account could be suspended c. I expected my reaction to help the discussion escalate into online bullying 1 = <i>not at all applicable</i> : 7 = <i>extremely applicable</i>
--	--

Counter-intervention Responsibility	Q. When you reacted favorably to the toxic comment, to what extent did you feel you needed to support, reinforce, assist, or defend the comment? Note. For example... - You felt you should show the commenter was not alone in his/her/their opinion about the target - You felt it was up to you to defend or assist the commenter against criticism from other comment-readers - You felt it was your duty to make the comment(s) popular and reach more people 1 = <i>not at all needed</i> : 7 = <i>extremely needed</i>
--	--

Counter-intervention Efficacy	Q. When you reacted favorably to the toxic comment, to what extent did you feel you were confident enough to support, reinforce, assist, or defend the comment? Note. For example... - You were sure you had the skills and knowledge necessary to defend the comment(s) - You were confident you knew what to say or do to reinforce and assist the comment(s) - You did not feel reluctant to support the comment(s) because of the potential of backlash 1 = <i>not at all confident</i> : 7 = <i>extremely confident</i>
Pro-perpetrator Attitude	Q. Recall what you thought and felt about the commenter when you reacted favorably to the toxic comment. Remember that the "commenter" is the person (not you) who made the toxic comment(s). To what extent does each of the following statements apply to your thoughts and feelings at that time? a. I felt good seeing the commenter express the same belief about the target that I have b. I felt a sense of similarity or identification with the perpetrator c. I felt grateful that the commenter said what I wanted to say d. I thought the commenter made a thought-provoking and clever argument e. I respected the commenter's bravery for speaking up 1 = <i>not at all applicable</i> : 7 = <i>extremely applicable</i>
Pro-incivility Norm	Q. Recall your observation of other comment-readers when you reacted favorably to the toxic comment(s). To what extent does each of the following statements reflect your thoughts at that time? a. I saw a lot of other comment-readers like or reply favorably to the comment(s) b. I thought other comment-readers would consider the comment(s) acceptable c. I thought other comment-readers would NOT approve of me defending the comment(s) ® 1 = <i>not at all applicable</i> : 7 = <i>extremely applicable</i>

Bystander Intervention (BI) Survey Questionnaire

Incivility Severity Perception	Q. When you reacted unfavorably to the toxic comment(s), to what extent did you think its potential harm to the target was severe? Note. For example, - You thought the comment(s) would harm the target - You thought the comment(s) went too far in blaming the target - You thought the comment(s) was probably a big deal for the target 1 = <i>not at all severe</i> : 7 = <i>extremely severe</i>
Intervention Responsibility	Q. When you reacted unfavorably to the toxic comment(s), to what extent did you feel you needed to oppose, confront, correct, or stop the comment(s)? Note. For example, - You felt it was your duty to do something against the commenter (e.g., disliking, reporting/flagging) - You felt you should help the target feel they were NOT alone - You felt it was up to you to defend or assist the target against criticism from others 1 = <i>not at all needed</i> : 7 = <i>extremely needed</i>
Intervention Efficacy	Q. When you reacted unfavorably to the toxic comment(s), to what extent did you feel you were confident enough to oppose, confront, correct, or stop the comment(s)? Note. For example... - You were sure you had the skills and knowledge necessary to defend the target - You were confident you knew what to say or do to support and assist the target - You did <i>not</i> feel reluctant to confront the commenter because of the potential of backlash 1 = <i>not at all confident</i> : 7 = <i>extremely confident</i>

Bystander Non-intervention (BNI) Survey Questionnaire

Incivility Severity Perception	Q. When you remained silent in reaction to the toxic comment(s), to what extent did you think its potential harm to the target was severe? Note. For example, - You thought the comment(s) would harm the target - You thought the comment(s) went too far in blaming the target - You thought the comment(s) was probably a big deal for the target 1 = <i>not at all severe</i> : 7 = <i>extremely severe</i>
Intervention Responsibility	Q. When you remained silent in reaction to the toxic comment(s), to what extent did you feel you needed <i>not</i> to oppose, confront, correct, or stop the comment(s)? Note. For example, - You felt it was <i>not</i> your duty to do something against the commenter (e.g., disliking, reporting/flagging) - You felt it was <i>not</i> up to you to help the target feel they were NOT alone - You felt you needed <i>not</i> to defend or assist the target against criticism from others 1 = <i>not at all needed</i> : 7 = <i>extremely needed</i>
Intervention Efficacy	Q. When you remained silent in reaction to the toxic comment(s), to what extent did you feel you were <i>not</i> confident enough to oppose, confront, correct, or stop the comment(s)? Note. For example... - You were <i>not</i> sure you had the skills and knowledge necessary to defend the target - You were <i>not</i> confident you knew what to say or do to support and assist the target - You felt reluctant to confront the commenter because of the potential of backlash 1 = <i>not at all confident</i> : 7 = <i>extremely confident</i>