Behind the Counter: Exploring the Motivations and Perceived Effectiveness of Online Counterspeech Writing and the Potential for AI-Mediated Assistance

Anisha Kumar

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

Dr. Eugenia H. Rho, Chair

Sang Won Lee

Mohammed Seyam

December 8, 2023

Blacksburg, Virginia

Keywords: Human-centered computing, collaborative and social computing, empirical studies in collaborative and social computing

Behind the Counter: Exploring the Motivations and Perceived Effectiveness of Online Counterspeech Writing and the Potential for AI-Mediated Assistance

Anisha Kumar

(ABSTRACT)

In today's digital age, social media platforms have become powerful tools for communication, enabling users to express their opinions while also exposing them to various forms of hateful speech and content. While prior research has often focused on the efficacy of online counterspeech, little is known about peoples' motivations for engaging in it. Based on a survey of 458 U.S. participants, we develop and validate a multi-item scale for understanding counterspeech motivations, revealing that differing motivations impact counterspeech engagement between those that do and not find counterspeech to be an effective mechanism for counteracting online hate. Additionally, our analysis explores peoples' perceived effectiveness of their self-written counterspeech to hateful posts, influenced by individual motivations to engage in counterspeech and demographic factors. Finally, we examine peoples' willingness to employ AI assistance, such as ChatGPT, in their counterspeech writing efforts. Our research provides insight into the factors that influence peoples' online counterspeech activity and perceptions, including the potential role of AI assistance in countering online hate.

Behind the Counter: Exploring the Motivations and Perceived Effectiveness of Online Counterspeech Writing and the Potential for AI-Mediated Assistance

Anisha Kumar

(GENERAL AUDIENCE ABSTRACT)

In today's digital age, social media platforms have become powerful tools for communication, enabling users to express their opinions while also exposing them to various forms of hateful speech and content. In addition to content moderation, counterspeech, or direct responses aimed at undermining hateful speech, is a tool that is being explored by organizations to counteract online hate, as it has been shown to prevent "platform hopping" while also promoting free speech. While prior research has primarily focused on the effectiveness of various types of counterspeech, little is known about what motivates people to engage in it. Based on a survey of 458 U.S. participants, we develop and validate a multi-item scale for understanding counterspeech motivations, revealing that differing motivations impact counterspeech engagement between those that do and not find counterspeech to be an effective mechanism for counteracting online hate. Additionally, our analysis explores peoples' perceived effectiveness of their counterspeech, influenced by individual motivations to engage in counterspeech and demographic factors. Finally, we examine peoples' willingness to employ AI assistance, such as ChatGPT, in their counterspeech writing efforts. Our research provides insight into the factors that influence peoples' online counterspeech activity and perceptions, including the potential role of AI assistance in countering online hate.

Dedication

To my parents, Sunil and Archana Kumar, and my brother Arnav Kumar, for your lifelong love and support that has allowed me to pursue this masters degree.

To my partner, Kevin Gautham Radja, for your love and support throughout my masters program.

To my advisor, Eugenia H. Rho, for guiding me throughout my degree, always having confidence in me, and helping me understand what it means to be a good researcher.

Acknowledgments

First, I would like to thank my advisor, Eugenia H. Rho for her guidance throughout my masters program. More specifically, for having confidence in me and always pushing me to strive for the best. Second, I'd like to thank my amazing SAIL Lab mates (Kaike Ping, Xiaohan Ding, Uma Gunturi, Buse Carik, Lance Wilhelm, Xiaozheng Wang, Rohan Leekha, Taufiq Daryanto, and Sophia Stil), especially Kaike Ping and Xiaohan Ding for their integral contributions to this work. I truly could not be more grateful to have worked with such a supportive and kind group. I will never forget the interesting conversations and fun lab parties we enjoyed together. Third, I would like to thank my close friends, Kelsey Newcomb and Uma Gunturi, for always being there for me. I would like to thank Kelsey for always hopping on a call when I needed to talk, and Uma for always being willing to hang out, grab food, go on a walk, or discuss research ideas for the last two years. Fourth, I would like to specially thank my partner, Kevin Gautham Radja, for his constant love and support both in and out of work. Finally, I would like to specially thank my parents, Sunil and Archana Kumar, for the sacrifices they made throughout their life that have allowed me the opportunities I have today, as well as my brother, Arnav Kumar, for his constant love and support.

Contents

List of Tables

Chapter 1

Introduction

In today's age of widespread digital connection, social media sites are key places for public conversation [31, 36, 97]. Although these platforms allow for the rapid spread of thoughts, they also act as breeding grounds for the proliferation of hate speech [20, 60], instances of cyberbullying [10], and various forms of harassment [19, 71]. While content moderation remains a predominant approach employed by tech companies to mitigate online hate speech [88], its effectiveness has been a topic of ongoing debate [43], particularly in how it can at times disperse rather than dispel hateful speech, and how it potentially conflicts with peoples' First amendment rights [46]. Such actions, broadly referred to as "deplatforming," often cause affected users to move to other, more permissive platforms [49].

In light of these challenges, there has been a marked shift towards alternative approaches to combat online hate, such as counterspeech [33]. Counterspeech is defined as direct responses to derogatory or harmful content, intended to undermine or refute the hateful message [76, 78]. For example, Meta has partnered with various NGOs across the globe since 2016 to establish global counterspeech initiatives ranging from Search Redirect, an initiative that aims to redirect users who search hateful or violent terms towards resources, education, and outreach groups to help, to the Online Civil Courage Initiative (OCCI) which is engaged with over 100 anti-hate groups globally.

Given the rise in counterspeech as a viable option in tackling online hate speech, previous work has qualitatively and quantitatively examined the effectiveness of various types of coun-

terspeech strategies such as empathy, humor, and warning of consequences and its impact on the broader social media ecosystem [26, 34, 41, 46, 48, 63, 82]. In the case of Megan Phelphs-Roper, who broke away from the Westboro Baptist Church in 2012 after being influenced by the criticism that she received on Twitter, an empathetic counterspeech strategy proved to be effective. Megan Phelphs-Roper's grandfather founded the Westboro Baptist Church, a church known for its homophobic teachings and anti-semitic views. In 2009, she began to use Twitter to help spread the teachings of the church and was unsurprisingly met with a lot of criticism on her posts. David Atibol, a middle-aged Jewish man, learned from past experience that the best way to engage with hateful people was to relate to them on a human level and his goal was to humanize Jewish people to members of the Westboro Baptist Church. Each person that Megan engaged with scraped away at her views little by little until she decided to leave the church in 2012, an act that would put her on par with gay people and Jewish people in the eyes of her fellow church members [21]. While Atibul had a personal interest in engaging in counterspeech, incentivizing those who do not remains an open challenge [61]. Mathew et al suggested that companies gamify incentive mechanisms to encourage people to engage in counterspeech, such as hierarchical badges (similar to stack overflow) [17, 35, 61]. While designing incentive mechanisms like this to promote counterspeech efforts could be a promising approach, understanding the motivations behind why people engage in such online behavior is essential for their effective implementation. Understanding these factors is crucial for not only creating more supportive and inclusive online spaces [7, 14, 57, 74, 81, 89], but also for empowering individuals to contribute positively to online discourse [38, 52, 59].

Meanwhile, Artificial Intelligence (AI) technologies based on large language models (LLMs) are increasingly being integrated into social media platforms. Companies like Nextdoor and Quora are experimenting with AI-powered features designed to help users craft posts or

engage in dialogues that contribute positively to community engagement [3, 6]. Given this emerging trend, we need to understand whether people are open to using AI to assist in their counterspeech activities. Moreover, if they are, how they perceive the impact of their own counterspeech in order to effectively design AI tools that can augment their own counterspeech in user friendly ways. Prior scholarship in NLP has long-recognized the role of AI in generating counterspeech [4, 30, 80, 99]. Nevertheless, while companies and international and nongovernmental organizations (I/NGOs) are encouraging counterspeech, NLP researchers recognize the emotional and mental toll that writing counterspeech can take on those who engage in it [15]. Thus, researchers in this domain have been exploring the task of automatically generating counterspeech as an important application of NLP for social good [25]. Nevertheless, we have a limited understanding of what influences user adoption of AI technologies for such purposes. Thus, we pose the following research questions:

- **RQ1:** *What motivations influence how often people engage in counterspeech on social media?*
- **RQ2:** *What motivations influence peoples' perceived effectiveness of their counterspeech?*
- **RQ3:** *What factors influence peoples' willingness to use AI assistance in writing counterspeech on social media?*
    a) **Motivations:** *What motivations influence peoples' willingness to use AI assistance in writing counterspeech on social media?*
    b) **Themes: Motivations and Reservations:** *What themes characterize peoples' motivations and reservations for using AI assistance in writing counterspeech on social media?*

We carried out a pre-registered survey with 458 English-speaking U.S. participants to explore their motivations for engaging in online counterspeech, the frequency of their counterspeech activities on social media, their perceived effectiveness of their self-written counterspeech to hateful posts, and their openness to using AI for assistance. Participants were presented with three examples of hate speech, chosen at random from a topically diverse set of 900, and were asked to write a counterspeech for each. Follow-up questions were then used to gauge their perceptions and experiences regarding the counterspeech they wrote.

Our research indicates that the frequency of counterspeech engagement is closely linked to individuals' beliefs in its effectiveness; those who view it as effective are often driven by a desire to challenge hate and promote inclusivity, whereas those skeptical of its impact tend to participate in support of others (RQ1). Second, demographic variables significantly shape engagement in counterspeech (RQ2); women tend to perceive their counterspeech as less effective, whereas minority groups report greater effectiveness. Third, we found that prior use of ChatGPT positively influenced individuals' willingness to use AI assistance in writing online counterspeech (RQ3). Moreover, participants with prior experience using ChatGPT are less willing to use AI to support and defend themselves, but more willing to use AI to signal inclusion.

Contributions: We contribute to Human-Computer Interation (HCI) research by examining the factors that drive people to engage in counterspeech on social media. We created and confirmed a multi-item scale to measure the motivations for engaging in online counterspeech, showing its impact on peoples' frequency of writing online counterspeech as well as their views regarding the effectiveness of their self-written counterspeech. This scale offers a framework for future exploration of online counterspeech behavior. Furthermore, we delve into the demographic and personal factors that drive people to engage in online counterspeech, moving beyond the existing focus on strategy and content [34, 41, 63, 82] to include

the backgrounds and personal experiences of those countering hate speech. By offering insights into how the motivations for engaging in counterspeech differ among various social groups, our findings inform the development of counterspeech tools that are tailored to the needs of diverse users [62]. Our analysis informs the creation of more personalized counterspeech tools for varied user groups [62]. Finally, we investigate the willingness to use AI to assist in writing counterspeech, a pertinent topic as AI's role in online moderation grows. These findings aid tech firms and scholars in understanding how AI might assist users in counteracting hate speech on digital platforms.

Chapter 2

Review of Literature

2.1 Understanding Potential Motivations in Writing Online Counterspeech

For counterspeech to act as a remedy to hateful speech, there need to be enough people willing to participate in it. Present research does not effectively examine why individuals participate in counterspeech, making it essential to thoroughly investigate user motivations to assess counterspeech's efficacy as a tool to counteract online hate.

Previous studies on bystander behavior in cyberbullying [89], online harassment [12], and digital civic engagement [8, 52, 72] provide a basis for examining why people engage in counterspeech. Research shows that the reasons people step in to stop cyberbullying are similar to those motivating them to oppose online hate [54, 67, 77]. Hate speech targets individuals based on attributes like race, gender, and sexuality [47, 51, 83], which is distinct from cyberbullying's more general focus on individual harassment not linked to social identity [93]. Moreover, a single incident of hate speech can have broad repercussions, as digital platforms amplify its reach, while cyberbullying typically involves persistent harassment over time [86, 91]. In both cases, users must decide whether to step in when they witness harmful or hateful content online [13, 32, 33, 67]. Additionally, counterspeech is a form of online civic intervention. Online civic intervention (OCI) refers to the efforts made by regular internet

users to counteract disruptive online behavior with the goal of restoring civil and rational public conversation. Prior work highlights individuals' attitudes and values as significant predictors of their likelihood of intervening online. Nevertheless, while previous research has tried to understand why people engage in counterspeech, this research often fails to paint an accurate picture by examining different forms of OCI in conjunction [72]. In this study, our focus narrows to counterspeech as a particular form of Online Civic Intervention (OCI), specifically examining individual characteristics and perceptions related to counterspeech. In order to do this, we develop a comprehensive understanding of the motivations associated with online counterspeech. Our study synthesizes insights from prior research in online harassment, bystander motivations in cyberbullying, as well as peoples' motivations for engaging in OCI. By drawing from such scholarship, we devise a set of survey variables that ope rationalize our investigation into what drives users to engage in counterspeech, as discussed next.

M1. Supporting Kin: Previous research has shown that the more emotionally or socially connected a bystander is to a victim of cyberbullying, the more likely they are to step in and help [13, 32]. Those with close relationships to friends and family are more active in opposing online hate compared to individuals with weaker social ties [28]. Additionally, comments that generalize and attack entire social groups are believed to create less empathy and connection with the victims than attacks directed at specific people [2, 66]. Finally, given that cyberbullying is more common in social environments such as work and school [77], we acknowledge that the probability of a member of someone's kin being a victim of cyberbullying as opposed to hate speech is higher. Nevertheless, we include this variable in order to understand how the closeness of a bystander to a victim influences bystanders' willingness to intervene in the context of online hate.

M2. Supporting Others: The idea that people help others, not just specific individuals

or groups, is linked to theories of social responsibility and collective efficacy [77]. Collective efficacy is the belief that one can make a difference for the greater good and impact the community [1]. Research has shown that individuals are more likely to act in a prosocial manner when they feel a moral duty to the wider community [29]. Studies on online behavior suggest that this sense of collective efficacy encourages people to participate more in altruistic actions [98]. Additionally, those who feel connected to an online community are more inclined to defend others against online harassment [27]. Moreover, support for solidarity citizenship norms, or the idea that good citizens should care for others, moderates the effect of individuals' exposure to hateful comments on their willingness to engage in online civic intervention [52]. These findings suggest that people engage in counterspeech not only for personal or familial protection but also for the benefit of others.

M3. Supporting Self: Counterspeech can be a very personal response, particularly when individuals feel personally attacked or hurt [75]. However, the reasons for self-defense can be complex. According to Guo and Johnson, people often downplay the effect of hate speech on themselves compared to its effect on others [45]. This underestimation might influence their willingness to use counterspeech for self-defense, as they might not fully recognize the harm directed at them. Conversely, experiencing online harm or targeted attacks can motivate individuals to confront online hate [87]. Given these factors, we consider "Supporting Self" as a factor in our study to explore why individuals might or might not defend themselves against online hate speech.

M4. Confronting Hate: The motivation to address hateful or harmful behavior is critical in encouraging bystanders to step in, especially in situations involving online hate speech [41] and cyberbullying [13]. Bystanders tend to act more when they see the comments as more hateful and threatening [9, 54, 72, 95]. This is supported by findings that show the more threatening the harassment appears, the more likely bystanders are to challenge it [54]. Therefore, we consider the motivation to confront such behavior or individuals as a factor

in engaging in online counterspeech.

M5. Educating Ignorance: Previous studies have shown that ignorance is a key reason behind the spread of online hate speech, with a lack of knowledge often resulting in a limited understanding of others in society [22]. As a result, several non-profit and educational groups [**?**], along with researchers, have promoted education as a method to combat online hate, preferring it over other measures like user bans or censorship [26, 92]. Supporting this, Buerger et al. discovered that many people who engage in counterspeech do so with the aim of educating those who express hate, explaining why their comments are inappropriate [15].

M6. Signaling Inclusion: The decision to participate in counterspeech is frequently driven by a wish to demonstrate social inclusion, especially in online communities [42]. In this regard, empathy plays an important role: studies have found that those with greater empathy are more likely to oppose online hate speech to defend the victim [77, 89] and to promote a sense of belonging and unity in the community [47, 56, 89]. Additionally, previous research has shown that people who prioritize individualizing moral foundations — valuing justice, rights, and equality — over binding moral foundations, which focus on group cohesion, respect for authority, and sacredness or purity, tend to be more actively involved in Online Civic Intervention (OCI) [72]. Therefore, we are incorporating "Signaling Inclusion" as a motivational factor in our research.

M7. Issue Focus: The drive to participate in counterspeech is often driven by topics or issues that are of personal importance to an individual. Studies indicate that bystanders are more inclined to intervene in situations involving social groups or matters they care about [68]. For instance, research has found that people are more prone to engage in Online Civic Intervention (OCI) in response to hate speech targeting women rather than comments about social welfare recipients [52]. This implies that the urge to counteract hate speech increases when it targets subjects or issues that individuals hold dear.

M8. Venting Emotions: Previous studies indicate that online incivility often triggers

emotion-focused coping strategies, like venting [58, 79]. Feelings of anger or outrage often prompt people to counteract hate speech as a way to release these emotions. Carlo et al.'s work also backs this idea, showing that emotional volatility is linked to using such emotion-driven coping tactics, often resulting in aggressive counterspeech [16]. Based on these insights, we are examining "Venting Emotions" as a possible motivational factor for participating in online counterspeech.

Motivation Variables (M1-M8): In the survey, we presented the motivation variables to participants as statements M1-M8 as shown in Table 2.1.

Participants were asked to indicate the extent to which each factor motivated them to write counterspeech on social media (How much do the following factors motivate you to write a counterspeech on social media?) with response options being 1 (None at all), 2 (A little), 3 (A Moderate Amount), 4 (A lot), 5 (A great deal).

Table 2.1: Motivation Variables and Questionnaire Items

No	**Motivation Variables**	**Questionnaire Items**
M1	Supporting Kin	When I feel the need to stand up for people I care about (e.g., family, close friends)
M2	Supporting Others	When I feel the need to stand up for people in general
M3	Supporting Self	When I feel the need to stand up for myself
M4	Confronting Hate	When I want to confront a hateful person or behavior
M5	Educating Ignorance	When I want to educate an ignorant person
M6	Signaling Inclusion	To signal that I stand for inclusion
M7	Issue Focus	When it concerns issues or topics I care about
M8	Venting Emotions	When I want to blow off steam

2.2 The Role of Counterspeech in Mitigating Online Hate Speech

In recent years, the role of counterspeech as a mechanism to counteract expressions of online hate has garnered substantial scholarly attention [40, 41, 63, 82]. However, prior studies have generated mixed findings that encapsulate the multifaceted nature of this area of scholarship. While some studies empirically demonstrate the effectiveness of online counterspeech [40, 41, 82], others question its impact [63], highlighting the complex nature of this topic. For instance, Miškolci et al. (2018) observed that direct responses to hate speech authors was ineffective in preventing them from further posting hateful content [63]. However, the authors discovered an auxiliary benefit, with counterspeech serving as an effective tool to reach a broader audience and stimulate further counterspeech. By contrast, Schieb and Preuss (2016)'s work support the efficacy of counterspeech. Through their experiment involving a computational simulation model, the authors demonstrate that counterspeech can indeed change the hateful speaker's behavior, such as leading to the deletion of hateful posts or eliciting apologies. Further, they found that the effectiveness of a counterspeech message was amplified when counter-speakers outnumbered the individuals sharing hateful messages. Intriguingly, a small group of counter-speakers could be still impactful, as long as the other users within the online community held relatively moderate rather than extreme views [82]. Hangartner et al.'s study further nuances our understanding of counterspeech by identifying the potency of empathy-based rhetoric. In their controlled experiment involving thousands of English-speaking Twitter users who had engaged in xenophobic or racist hate speech, those who were addressed with empathy-based counterspeech were more likely to delete their hateful posts than those who received humorous or warning-based counterspeech [46]. Such prior studies enrich our understanding of the role and impact of counterspeech in

mitigating online hate speech, highlighting broader implications for the trajectory of future online public discourse. While previous research provides important insights into the kind of speech that can sway hateful speakers, such studies primarily focus on the content or strategies of counterspeech rather than those who author it, often overlooking the influence of social identities and perceptions of counter-speakers. A few studies [65, 84] have started exploring these factors. For example, researchers have shown that a counter-speaker's race and level of influence can influence how the counterspeech is taken by the audience [65, 84]. In their experiment using bots identified as black or white and high and low status to rebuke hateful speakers on Twitter, Munger et al found that hateful speakers who were sanctioned by a high-follower white male significantly reduced their use of a racist slur [65]. While this information is useful in determining additional factors that can influence the efficacy of counterspeech, there is no research to date that examines the psychological factors that drive people to engage in counterspeech or how they feel about their self-written counterspeech . Hence, our work seeks to fill this gap.

2.3 The Role of AI in Online Counterspeech Engagement

The computational research in the domain of AI and counterspeech primarily examines the technical challenges in identifying [23, 41, 48], creating [53, 73, 80, 90], and assessing [99] [38, 48, 66] counterspeech through the use of Large Language Models (LLMs) [24]. Particularly, research in Natural Language Processing (NLP) tends to concentrate on enhancing the production of counterresponses to hate speech that closely mimic human interaction, employing different criteria to gauge the quality of AI-crafted counterspeech, including aspects like informativeness [23], politeness [80], and grammatical diversity [99]. Nevertheless, amidst this technological progress as well as the surge in the use of AI technology based on

large language models by social media platforms, there is a noticeable lack of exploration into the human factors that influence peoples' use of AI in assisting their counterspeech activities. To understand what role AI can play in assisting users in engaging in counterspeech, it is first crucial to understand peoples' comfort level when it comes to using AI to help them write counterspeech in addition to their experience of crafting counterspeech. Our work aims to fill this gap by examining the factors that influence peoples' willingness to use artificial intelligence tools to help them write online counterspeech, as well as the human factors that can aid the design of AI assisted counterspeech writing tools. By delving into the nuanced interplay between user experience and AI assistance, our work seeks to understand how artificial intelligence can not only expedite but also enhance the quality and impact of counterspeech in online discourse.

Chapter 3

Methods

We conducted a pre-registered survey (N = 458) across English-speaking participants in the U.S. to examine the key motivations that drive people to engage in online counterspeech, how often they write counterspeech on social media, their perceived effectiveness of their counterspeech, and their willingness to use AI to help them write counterspeech. The survey showed participants three different examples of hate speech randomly selected from a topically diverse pool of 900 hateful posts, and asked them to respond by writing a counterspeech in response to it. Following this, participants were asked questions to assess their perceived effectiveness of their self-written counterspeech, the frequency of their counterspeech activities, their willingness to use AI to assist them in writing counterspeech, their social media activity, and demographic characteristics.

3.1 Selecting Hateful Posts

To curate a balanced and representative sample of hate speech for our survey, we sourced hateful posts from three prominent online hate datasets: the ETHOS dataset [64], the Multi-Target Counter Narrative Dataset [39], and the Multilingual and Multi-Aspect Hate Speech Analysis (MLMA) collection [69]. We randomly selected hateful posts across five commonly occurring topics from this combined corpus: gender, religion, disability, sexual orientation, and race. To avoid over or under representation of a specific topic, we balanced our dataset by manually examining all hate posts to ensure each one was topially relevant. This resulted

in a total of 900 hateful posts across the five topics: race (183), gender (183), religion (182), sexual orientation (182), and disability (170).

3.2 Survey Design and Variables

The survey was designed using Qualtrics and consisted of (a) a consent form (b) relevant background information about hateful speech and counterspeech, (c) 3 hateful posts and questions pertaining to them, (d) questions about past online hate speech experience, frequency of writing counterspeech online, and motivations for writing online counterspeech (e) questions about prior use of ChatGPT, perceived usefulness of ChatGPT, as well as willingness to use such AI tools to aid in counterspeech writing, and finally (f) demographic and social media use questions. The consent form informed participants that they were being invited to a study to evaluate the efficacy of counterspeech to hateful posts on social media, as well as informing them of the potential psychological risks due to the offensive nature of hateful speech. Then, participants were provided with definitions of hateful speech, counterspeech, as well as examples of effective counterspeech. Following this, participants were shown three unique hateful posts randomly selected from the set of 900 hate posts described in 3.1. For each hateful post, participants were prompted with "Imagine you are a user of an online group on social media. Another user (perpetrator) in the group posted the following. Do you consider this post to be hateful?" If they answered Yes, participants were prompted to write a counterspeech to the hateful post shown. The survey asked, "Please write a counterspeech to this post. The goal is to further reduce hateful behavior from the perpetrator." Participants were then asked to rate their perceived effectiveness of each counterspeech they wrote using a five-point Likert scale (Not effective at all, Slightly effective, Moderately effective, Very effective, Extremely effective). Participants were asked to write a counterspeech response so that they could assess the effectiveness of their self-

written counterspeech. Finally, participants answered questions related to their motivations for writing online counterspeech, frequency of writing online counterspeech, and willingness to use ChatGPT to write counterspeech on social media. Table 3.1 lists all variables included in our survey.

3.3 Recruitment

Participants were recruited via Prolific, limited to U.S.-based, English-speaking adults with approval ratings above 95%. All participants were warned about potentially harmful content in the survey. Of the initial 536 respondents, we excluded those who failed attention checks or failed to complete the survey, resulting in a final sample of 458 participants. The demographic details of participants can be found in Table 1 of the Appendix. The average survey completion time was 15 minutes with a compensation rate of $12/hour.

Table 3.1: Survey Variables and Associated Measures

Variables	Response Range
Motivations (IV)	*Response range listed in in Section 2.1*
M1: Supporting kin	
M2: Supporting public	
M3: Supporting self	
M4: Confronting hate	
M5: Educating ignorance	
M6: Signaling inclusion	
M7: Issue focus	
M8: Venting emotions	
Social Media Behavior and Experience (Control)	
Social media commenting frequency	1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)
Use Real Name on Social Media	1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)
Prior experience of online hate speech target	1 (No), 2 (Yes)
Frequency of encountering online hate speech	1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)
Perceived general efficacy of counterspeech	1 (Not effective at all), 2 (Slightly effective), 3 (Moderately effective), 4 (Very effective), 5 (Extremely effective)
Frequency of writing counterspeech	1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)
Prior Use and Perception of ChatGPT (Control)	
Prior Use of ChatGPT	1 (Yes), 2 (No)
Perceived Usefulness of ChatGPT	1 (Not at all Useful), 2 (Slightly Useful), 3 (Moderately Useful), 4 (Very Useful), 5 (Extremely Useful)
Demographics (Control)	
Age	*Free response*
Gender	Man, Woman, Other
Ethnicity	Asian, Black, Hispanic, Middle Eastern, Native American, Pacific Islander, White
Education Level	Less than High School, High school graduate, Some college, 2-year degree, 4-year degree, Professional degree, Doctorate
Sexual Orientation	Heterosexual, Homosexual, Other, Prefer Not to Say
Political View	Very Conservative, Conservative, Moderate, Liberal, Very Liberal
Dependent Variables	
RQ1: Frequency of writing counterspeech	1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)
RQ2: Perceived effectiveness of self-written counterspeech	1 (Not effective at all), 2 (Slightly effective), 3 (Moderately effective), 4 (Very effective), 5 (Extremely effective)
RQ3: Willingness to use ChatGPT to write counterspeech	1 (Definitely Not), 2 (Probably Not), 3 (Might or Might not), 4 (Probably yes), 5 (Definitely yes)

3.4 Analysis

RQ1 What Motivations Influence How Often People Engage in Counterspeech on Social Media? To address RQ1, we performed a linear regression analysis to examine the factors that influence peoples' frequency of writing counterspeech on social media. The dependent variable was participants' self-reported frequency of writing counterspeech, which

was measured on a five-point Likert scale (Never, Rarely, Sometimes, Often, Frequently) in response to the question "How often do you write counterspeech online?" The independent variables were the 8 motivation variables, prior experience being a target of online hate speech, perceived general efficacy of counterspeech *(How effective do you think counterspeech is in reducing online hate?)*, as well as control variables relating to social media behavior and experience, as well as demographics. To detect multicollinearity, we also calculated the variance inflation factor (VIF) for each independent variable, with a VIF value greater than 5 indicating a serious multicollinearity problem [70].

RQ2: What Motivations Influence Peoples' Perceived Effectiveness of Their Counterspeech? To address RQ2, we performed a mixed linear regression analysis to examine the motivations that influence peoples' perceived effectiveness of their self-written counterspeech. This model allowed us to account for both fixed effects and random effects, making it suitable for modeling the repeated measures in our data (3 counterspeech per participant). The dependent variable was participants' perceived effectiveness of their self-written counterspeech in response to the 3 hateful posts that that they saw. This was measured on a five-point Likert (Not effective at all, Slightly effective, Moderately effective, Very effective, Extremely effective) scale in response to the question *"How effective do you think your counterspeech would be in preventing the perpetrator from engaging in further hateful behavior?"* The fixed effects were the 8 motivation variables, prior experience being a target of online hate speech, perceived general efficacy of counterspeech *(How effective do you think counterspeech is in reducing online hate?)*, frequency of writing counterspeech, control variables relating to social media behavior and experience, as well as demographics. To account for repeated measures in our data, we introduced random effects, considering participants.

RQ3: What Factors Influence Peoples' Willingness to Use AI Assistance in Writing Counterspeech on Social Media? To address RQ3, we conducted a linear regression model to examine the relationship between peoples' motivations for engaging in counterspeech on social media and their willingness to use AI technology, such as ChatGPT, for this purpose. The independent variables consisted of the eight motivation items, prior experience being a target of hateful speech, perceived general efficacy of counterspeech, frequency of writing counterspeech, with demographic factors and as social media use and experience variables as control variables. The dependent variable was captured via a five-point Likert scale (Definitely not, Probably not, Might or Might not, Probably yes, Definitely yes) in response to the question, "If you were writing counterspeech on social media, would you use artificial intelligence technology like ChatGPT to assist you?" Given that some participants had no prior use of AI technologies like ChatGPT, we conducted a subgroup analysis for only participants with prior experience using ChatGPT (N=296). For those who have used ChatGPT before, an additional control variable for perceived usefulness of ChatGPT was included.

Chapter 4

Findings

4.1 What Motivations Influence How Often People Engage in Counterspeech on Social Media?

Our linear regression results in Table 4.1 show seven key factors to be significantly associated with how frequently individuals engage in counterspeech on social media. One's social media commenting frequency understandably emerged as the most influential predictor, with those that comment more frequently on social media engaging in counterspeech more often (b = .211, β = .208, p =.000). The second strongest predictor was one's past experience as a target of hateful speech, with individuals who have been a victim of online hateful speech engaging in counterspeech more often than those without such experience (b = .354, β = .171, p =.000). With respect to peoples' motivations, those that are more motivated to write online counterspeech in order to stand up for and support others (other than kin) (b = .104, β = .134, p = .043), confront a hateful person or behavior (b = .117, β = .152, p = .015), and vent their emotions (b = .078, β = .092., p = .027) write counterspeech more frequently. Finally, our results show that those who perceive counterspeech to be more effective also write counterspeech more often (b = .126, β = .131, p =.002).

Table 4.1: Linear Regression Results for Counterspeech Writing Frequency on Social Media (N=458)

Variables	B	β	Std.Error	t value	p value	VIF
M1: Supporting kin	-0.012	-0.015	0.044	-0.265	0.791	2.377
M2: Supporting others	0.104	0.134	0.051	**2.032***	**0.043***	3.276
M3: Supporting self	-0.018	-0.023	0.039	-0.444	0.657	2.036
M4: Confronting hate	0.117	0.152	0.048	**2.444***	**0.015***	2.924
M5: Educating ignorance	0.021	0.029	0.045	0.473	0.636	2.825
M6: Signaling inclusion	0.036	0.051	0.038	0.945	0.345	2.166
M7: Issue focus	0.037	0.045	0.049	0.754	0.452	2.721
M8: Venting emotions	0.078	0.092	0.035	**2.220***	**0.027***	1.309
Past experience of online hate speech target	0.354	0.171	0.080	**4.430*****	**0.000*****	1.122
Perceived general efficacy of counterspeech	0.126	0.131	0.041	**3.084****	**0.002****	1.361
Social media commenting frequency	0.211	0.208	0.041	**5.186*****	**0.000*****	1.214
Use of real name on social media	0.034	0.057	0.023	1.505	0.133	1.101
Frequency of encountering online hate speech	0.078	0.075	0.040	1.955	0.051	1.111
Age	0.001	0.019	0.003	0.493	0.623	1.180
Gender	-0.015	-0.007	0.077	-0.199	0.843	1.079
Ethnicity	-0.078	-0.039	0.079	-0.988	0.324	1.156
Education level	0.003	0.002	0.054	0.053	0.958	1.111
Sexual orientation	-0.131	-0.053	0.101	-1.306	0.192	1.257
Political views	0.073	0.084	0.036	**2.035***	**0.042***	1.295

Adjusted R-squared = 0.3963; $F(19, 438) = 16.79$, p $< .001$.

Variability in Counterspeech Engagement Among Individuals With High and Low Perceived General Efficacy of Counterspeech. Given that our results show that people who perceive counterspeech to be more effective tend to engage in counterspeech more often, we conducted a subgroup analysis to understand how counterspeech motivations differed between the two groups (individuals who do/do not perceive counterspeech to be an effective method to combat online hate). We conducted two regression models: those who do not find counterspeech to be effective, N=270 (Likert scale values 1-2 (Not effective at all, Slightly effective)), and those that do, N=188 (Likert scale values 3-5 (Moderately effective, Very effective, Extremely effective). Results are shown in Table 4.2. All VIF values were below 3.3.

The subgroup analysis shows that individuals who do not believe counterspeech to be effec-

Table 4.2: Subgroup Linear Regression Results for Counterspeech Writing Frequency on Social Media)

Variables	**Not Effective (N=270)**				**Effective (N=188)**			
	B	***β***	**t value**	**p value**	***B***	***β***	**t value**	**p value**
M1: Supporting kin	0.017	0.024	0.312	0.755	-0.119	-0.133	-1.405	0.162
M2: Supporting others	0.136	0.180	**2.203***	**0.029***	0.052	0.060	0.544	0.587
M3: Supporting self	-0.014	-0.019	-0.287	0.774	-0.033	-0.038	-0.434	0.665
M4: Confronting hate	0.103	0.143	1.829	0.069	0.207	0.231	**2.165***	**0.032***
M5: Educating ignorance	0.069	0.101	1.254	0.211	0.013	0.015	0.153	0.878
M6: Signaling inclusion	-0.057	-0.078	-1.145	0.253	0.144	0.193	**2.292***	**0.023***
M7: Issue focus	0.004	0.005	0.059	0.953	0.092	0.099	1.078	0.283
M8: Venting emotions	0.055	0.064	1.104	0.271	0.105	0.134	**1.994***	**0.048***
Past experience of online hate speech target	0.266	0.139	**2.640****	**0.009****	0.396	0.189	**2.813****	**0.005****
Social media commenting frequency	0.240	0.258	**4.716*****	**0.000*****	0.156	0.143	**2.189***	**0.030***
Use of real name on social media	0.057	0.104	1.939	0.054	-0.002	-0.003	-0.042	0.967
Frequency of encountering online hate speech	0.083	0.086	1.653	0.100	0.047	0.045	0.661	0.510
Age	0.000	0.007	0.123	0.902	0.001	0.013	0.190	0.849
Gender	0.021	0.011	0.211	0.833	-0.006	-0.003	-0.043	0.965
Ethnicity	-0.071	-0.037	-0.698	0.486	-0.110	-0.053	-0.817	0.415
Education level	-0.005	-0.003	-0.067	0.946	0.013	0.009	0.137	0.891
Sexual orientation	-0.246	-0.107	-1.883	0.061	0.116	0.047	0.683	0.496
Political views	0.106	0.127	**2.166***	**0.031***	0.050	0.060	0.853	0.395

Adjusted R-squared = 0.3485; $F(18, 251) = 8.995$, p $< .001$
Adjusted R-squared = 0.3268; $F(18, 169) = 6.044$, p $< .001$

tive in counteracting online hate write more counterspeech if they are motivated to support others (b = .136, β = .180, p =.029). On the other hand, those who believe counterspeech is effective and have a desire to confront hate, signal inclusion, or vent their emotions write counterspeech more often (b= .207, β = .231, p = .032, b=.144, β = .193, p = .023, b=.105, β = .134, p = .048). This analysis highlights key differences in motivations for engaging in counterspeech for those that do and do not believe in the efficacy of counterspeech.

Finally, those that do not believe counterspeech to be effective and identify as more liberal engage in counterspeech more often.

4.2 What Motivations Influence Peoples' Perceived Effectiveness of Their Counterspeech?

The random effect in this model was the UserID of each participant.

The fixed effects Table 4.3 shows six key factors to be significantly associated with how effective people perceive their counterspeech to be. The strongest predictor of individuals' perceived effectiveness of their counterspeech is how generally effective they find counterspeech to be in combating online hate (b = .499, p=0.000). People who think that counterspeech can change the perspectives of those posting hateful content find their self-written counterspeech to be more effective than those who do not believe that counterspeech can be an effective tool to combat online hate. Following this, gender and ethnicity were the second and third strongest predictors, respectively, with women finding their counterspeech to be less effective than men (b = -0.365, p = 0.000) and minority races finding their counterspeech to be more effective than majority races (b = 0.193, p = 0.013).

With respect to peoples' motivations, those who are more motivated to write online counterspeech in order to support others (other than kin) find their counterspeech to less effective (b = -0.132, p=0.009) while those that are motivated to write counterspeech to support themselves find their counterspeech to more effective (b = 0.114, p=0.003).

Table 4.3: Mixed Linear Regression Results for Perceived Effectiveness of Counterspeech(N=1261)

Variables	***B***	**Std. Error**	**df**	**t value**	**p value**
M1: Supporting kin	0.025	0.043	436.371	0.571	0.569
M2: Supporting others	-0.132	0.050	444.135	**-2.616****	**0.009****
M3: Supporting self	0.114	0.038	436.975	**2.968****	**0.003****
M4: Confronting hate	0.038	0.047	428.208	0.825	0.410
M5: Educating ignorance	0.059	0.044	434.904	1.349	0.178
M6: Signaling inclusion	0.024	0.037	432.097	0.642	0.522
M7: Issue focus	-0.020	0.048	431.661	-0.426	0.670
M8: Venting emotions	-0.008	0.034	434.236	-0.224	0.823
Past experience of online hate speech target	-0.060	0.079	428.428	-0.751	0.453
Frequency of writing counterspeech	0.028	0.047	432.048	0.598	0.550
Perceived general efficacy of counterspeech	0.499	0.040	430.076	**12.416*****	**0.000*****
Social media commenting frequency	0.047	0.041	430.383	1.162	0.246
Frequency of encountering online hate speech	0.029	0.039	433.961	0.731	0.465
Use of real name on social media	0.065	0.022	430.995	**2.933****	**0.004****
Age	-0.002	0.003	432.381	-0.834	0.405
Gender	-0.365	0.075	433.290	**-4.881*****	**0.000*****
Ethnicity	0.193	0.077	429.608	**2.501***	**0.013***
Education level	-0.033	0.053	437.234	-0.625	0.532
Sexual orientation	0.070	0.097	425.092	0.720	0.472
Political views	-0.063	0.035	432.365	-1.797	0.073

4.3 What Factors Influence Peoples' Willingness to Use AI Assistance in Writing Counterspeech on Social Media?

Our linear regression results in Table 4.4 show three key factors to be significantly associated with peoples' willingness to use AI to help them write counterspeech. Our results indicate that the strongest predictor of one's willingness to use AI tools such as ChatGPT to help them write counterspeech is their prior use of ChatGPT, with those with prior experience with ChatGPT being more willing to use AI to help them write counterspeech (b = 0.425, β = .178, p =.000). The second strongest predictor was peoples' motivation to signal

inclusion (b=0.135, β = .168, p=0.013). In other words, those who are more motivated to signal inclusion are more willing to use AI to help them write counterspeech. Finally, the third strongest predictor was an individual's past experience as a target of hateful speech. Our results indicate that those that have been past targets of hateful speech depicted less willingness to use AI to help them write counterspeech (b = -0.341, β = −.146 p =.003).

Table 4.4: Linear Regression Results for Willingness to Use AI Assistance for Writing Online Counterspeech (N=458)

Variables	***B***	***β***	**Std. Error**	**t value**	**p value**	**VIF**
M1: Supporting kin	-0.023	-0.026	0.062	-0.377	0.706	2.377
M2: Supporting others	-0.001	-0.001	0.073	-0.017	0.986	3.307
M3: Supporting self	-0.045	-0.052	0.056	-0.805	0.421	2.038
M4: Confronting hate	0.001	0.002	0.068	0.022	0.983	2.967
M5: Educating ignorance	-0.001	-0.001	0.063	-0.012	0.990	2.830
M6: Signaling inclusion	0.135	0.168	0.054	**2.502***	**0.013***	2.181
M7: Issue focus	0.048	0.052	0.069	0.687	0.492	2.725
M8: Venting emotions	-0.017	-0.018	0.050	-0.346	0.730	1.323
Past experience of online hate speech target	-0.341	-0.146	0.115	**-2.967****	**0.003****	1.174
Prior use of ChatGPT	0.425	0.178	0.113	**3.765*****	**0.000*****	1.079
Frequency of writing counterspeech	-0.120	-0.106	0.067	-1.777	0.076	1.736
Perceived general efficacy of counterspeech	0.037	0.034	0.058	0.630	0.529	1.403
Use of real name on social media	0.014	0.021	0.032	0.444	0.657	1.116
Social media commenting frequency	0.107	0.093	0.059	1.811	0.071	1.291
Frequency of encountering online hate speech	-0.022	-0.019	0.056	-0.398	0.691	1.122
Age	-0.001	-0.012	0.004	-0.245	0.807	1.184
Gender	0.120	0.053	0.109	1.106	0.269	1.093
Ethnicity	0.129	0.056	0.112	1.148	0.252	1.171
Education level	0.102	0.064	0.076	1.338	0.182	1.111
Sexual orientation	0.084	0.030	0.142	0.590	0.555	1.265
Political views	-0.094	-0.096	0.051	-1.847	0.065	1.311

Adjusted R-squared = 0.05727; $F(21, 436) = 2.322$, p < .001.

People Who Have Used ChatGPT Since an individual's prior use of ChatGPT was the strongest predictor of their willingness to use AI tools such as ChatGPT, we conducted a second regression model with only those with prior experience using ChatGPT (N=296). Results are shown in Table 4.5. For those that have used ChatGPT, the strongest predictor

of their willingness to use AI assistance to write counterspeech is their perceived usefulness of ChatGPT (b = 0.348, β = .323, p =0.000). The second strongest predictor for this group of people is the motivation to signal inclusion. People who have used ChatGPT before and are motivated to signal inclusion are more willing to use AI to help them write counterspeech (b = 0.182, β = .222, p =0.006). Additionally, those who write counterspeech more frequently are actually less willing to use AI to help them write counterspeech (b = -.193, $\beta = -.168$, p =0.020). Finally, those who are more motivated to write counterspeech to support themselves are less willing to use AI tools such as ChatGPT to assist them (b = -0.178, $\beta = -.200$, p =0.012).

Table 4.5: Linear Regression Results for Willingness to Use AI Assistance for Writing Online Counterspeech Among Prior Users of ChatGPT (N=296)

Variables	***B***	***β***	**Std.Error**	**t value**	**p value**	**VIF**
M1: Supporting kin	0.081	0.089	0.076	1.079	0.282	2.474
M2: Supporting others	0.049	0.055	0.094	0.521	0.602	4.008
M3: Supporting self	-0.178	-0.200	0.071	**-2.527***	**0.012***	2.283
M4: Confronting hate	-0.060	-0.067	0.084	-0.711	0.478	3.188
M5: Educating ignorance	0.009	0.010	0.077	0.114	0.909	2.956
M6: Signaling inclusion	0.182	0.222	0.066	**2.768****	**0.006****	2.332
M7: Issue focus	0.011	0.012	0.084	0.133	0.895	2.815
M8: Venting emotions	-0.037	-0.039	0.057	-0.641	0.522	1.325
Past experience of online hate speech target	-0.265	-0.112	0.138	-1.915	0.057	1.248
Frequency of writing counterspeech	-0.193	-0.168	0.082	**-2.338***	**0.020***	1.885
Perceived general efficacy of counterspeech	0.024	0.022	0.073	0.328	0.743	1.568
Perceived usefulness of ChatGPT	0.348	0.323	0.062	**5.637*****	**0.000*****	1.189
Use of real name on social media	0.018	0.026	0.038	0.477	0.634	1.113
Social media commenting frequency	0.176	0.149	0.073	**2.398***	**0.017***	1.399
Frequency of encountering online hate speech	0.020	0.018	0.067	0.305	0.761	1.197
Age	0.003	0.035	0.005	0.613	0.540	1.201
Gender	0.137	0.059	0.129	1.059	0.290	1.111
Ethnicity	0.067	0.029	0.134	0.500	0.618	1.187
Education level	0.135	0.083	0.090	1.498	0.135	1.121
Sexual orientation	0.057	0.021	0.163	0.351	0.726	1.299
Political views	-0.166	-0.167	0.060	**-2.761****	**0.006****	1.323

Adjusted R-squared = 0.1877; $F(21, 274) = 4.245$, p $< .001$.

Qualitative Analysis: Motivations and Reservations for Using AI for Counterspeech Writing Our qualitative analysis of participants' open responses revealed three themes supporting the openness to using AI for writing counterspeech and three themes opposing its use. Tables 4.6 and 4.7 show the main themes that emerged, along with illustrative examples and the proportion of responses that fell into each theme. We had three raters who independently coded the participants' open responses into the themes that were identified. The overall Cohen's kappa coefficient for our analysis was 0.854, with a 95% confidence interval of 0.817 to 0.891, indicating a very good level of agreement among the raters. Because some user statements contained more than one theme, we coded them into multiple categories; thus, the total percentages of the themes exceed 100%. We discuss each theme in detail below.

Efficiency and Convenience: Participants emphasized how using AI can "save time and effort compared to writing a response from scratch", thereby making the writing process quicker for them. Some participants also highlighted that AI could help them more easily come up with ideas that they could elaborate on themselves as well as provide them with useful strategies for writing effective counterspeech that they can use later on.

Less Emotional Burden: Many participants felt that AI tools like ChatGPT could help alleviate many of the negative emotions, such as anger and frustration, that often inevitably arise when writing counterspeech. For example, one participant noted: "it can save the stress and irritation of responding to an ignorant person".

Access to Larger Knowledge Base & Better Articulation: Many participants also underscored the ability of AI to not only help them express themselves more clearly, but also provide supporting evidence for arguments. For example, participants stated that AI tools like ChatGPT can help them "find the right words and vocabulary to express [their] thoughts more clearly," as well as "provide data and facts to make [their] argument stronger".

Table 4.6: Reasons for Using AI to Write Online Counterspeech (38.5%)

Themes	Illustrative Quotes	%
Efficiency and Convenience	• I think it's better and faster at putting together coherent sentences that get my point across than myself. • It can save time and effort compared to writing a response from scratch. • It can give me ideas quickly and I can elaborate with my own perspective of the facts. • I think it would save me time and energy, maybe it could help me to better learn the skill so I could use it more.	24.7%
Less Emotional Burden	• It would take all of the emotional work out of it for me. • It saves you the stress and irritation of having to respond to an ignorant person. • I feel like ChatGPT would be able to refute it with facts and logic in better ways than me, because I feel like counterspeech is an emotional burden on me and I get overwhelmed. • I would probably have the AI help, partially because I just don't have the energy for that sort of thing anymore.	11.6%
Access to Larger Knowledge Base and Better Articulation	• It can help me find the right words and vocabulary to express my thoughts more clearly and eloquently. • It has access to a huge breadth of knowledge that I don't, so it can provide data and facts to make my argument stronger. • ChatGPT would be able to assist me with my argument in order to make my counterspeech more effective. • I would use it to help get my statement across in a much clearer way. Also to help me make sure that the information I am writing about is correct. • ChatGPT has a broad database full of statistics and information, and I feel as though it would create the most effective counterspeech because of that. It has nearly all of the information in the world within it, it would certainly make an argument more efficient than I probably could.	2.2%

Authenticity and Ethical Concerns: The most common reservations for using AI assistance in writing counterspeech are authenticity and ethical concerns. Some participants expressed guilt, with one participant stating that "Using ChatGPT to make counterspeech and then posting it as if it were [their] own is lying and unethical at best.". Moreover, others expressed worries that using AI for assistance prevents them from having ownership of their words, with a participant saying that they would want their counterspeech "to be in their own words and thoughts".

Lack of Emotional, Human, or Personal Touch: Many participants raised doubt about AI's ability to mimic human emotions such as empathy, with a participant saying that they believe AI "can use logic but not empathy to write counterspeech." Additionally, participants also mentioned AI's lack of ability to capture their life experiences or "fully express what [they] want to express".

Lack of Familiarity or Trust in AI: Many participants also seem to have a general distrust of AI technology, with one participant stating they "don't think ChatGPT and AI in general is quite the 'do it all' answer everyone acts like it is". Others cite their lack of familiarity with AI tools as the primary reason for their distrust. Moreover, many also recognize that AI may not be "100% accurate or correct".

Table 4.7: Reservations Against Using AI to Write Online Counterspeech (71.7%)

Themes	**Illustrative Quotes**	%
Authenticity and Ethical Concerns	• If I were being graded for a counterspeech, it would be cheating to have anyone or anything write it for me. • It's not my voice. It's not my perspective. Personally I'd be ashamed to utilize Artificial Intelligence for counterspeech. • Because then it wouldn't even be MY counterspeech. Why would I use AI to write MY opinion? It's stupid. • If my statement were to be judged by others, I would want the statement to be in my own words using my own thoughts. • Using ChatGPT to make counterspeech and then posting it as if it were my own is lying and unethical at best. • I would want to build the skills to effectively and reliably write such speech myself.	33.0%
Lack of Emotional, Human, or Personal Touch	• This needs human sentiment with human feelings behind them. ChatGPT AI may get there but it's not there yet. • I think I could write it better because I can use personal experiences and my emotions to hopefully make the perpetrator really think about it. • It can use logic, but not empathy, to write counterspeech. • I would rather tailor my response to be exactly what I'm thinking. I'm not sure it could fully express what I want to express, and it may lack nuance • It doesn't come from the heart.	26.0%
Lack of Familiarity or Trust in AI	• I'm not familiar with it, hence my trust level in its performance is low. • I don't think ChatGPT has enough understanding of how internet commenting dynamics work. • ChatGPT would be able to assist me with my argument in order to make my counterspeech more effective. • I don't think ChatGPT and AI in general is quite the "do it all" answer everyone acts like it is. • It's not always 100% accurate or correct and could cause issues if you post it as counterspeech and it turns out to be incorrect. • I trust my own words more than a robot's.	12.7%

Chapter 5

Discussion

5.1 Factors that Drive Online Counterspeech

5.1.1 Prior Victimization

Studies in online bystander intervention show that prior victimization is a key predictor of bystander action [77]. Moreover, Costello et al. (2016) found that past victims of online hate are more than 3 times as likely to defend fellow victims [27]. Our RQ1 results confirm these insights, demonstrating having been a target of online hate speech in the past to be the second strongest predictor that drives people to frequently engage in counterspeech on social media.

5.1.2 Perceived Efficacy of Counterspeech

Existing literature has explored how individuals respond to online hate speech, distinguishing between high and low threshold Online Civic Interventions (OCI)—that is, direct versus indirect engagement—and between those who remain silent and those who actively intervene [8, 72]. Our study pivots from examining these behavioral patterns to analyzing how personal attitudes affect the frequency with which people engage in counterspeech. Specifically, we examine whether the belief in counterspeech as an effective tool to mitigate online hate

predicts how often people choose to engage in it. Thus, given that our results reveal the perceived efficacy of counterspeech to be a significant predictor of peoples' frequency of engaging in counterspeech, we conducted a subgroup analysis by differentiating between those who do and do not find counterspeech to be effective in minimizing online hate. Prior work has shown that the effectiveness of an intervention significantly influences bystanders' willingness to intervene in cases of online harassment [9, 96]. We provide further nuance to prior scholarship by showing how motivations in fact, significantly vary between those that find counterspeech to be effective as opposed to those who do not. Our results show that those who find counterspeech to be effective frequently engage in counterspeech when they are motivated to confront hate and signal inclusion while those that do not believe counterspeech is effective frequently engage when they are motivated to support others. In other words, those who feel that counterspeech is generally ineffective in minimizing hate speech may be more motivated to engage in order to stand up for others experiencing hate rather than confront a hateful person that they believe may be unwilling to see their perspective. Prior work highlights that those who support solidarity citizenship norms, or a belief that they have a responsibility to care for others, as well as those who value fairness, justice, and equality tend to engage in more frequent OCI [52, 72]. Our work provides a more granular look at what motivates people to engage in counterspeech, a specific type of OCI.

In RQ2, we explore how peoples' motivations to write counterspeech influence their perceived effectiveness of their counterspeech. Our results reveal that people who are motivated to write counterspeech to support themselves report higher perceived levels of effectiveness of their counterspeech, while those who are motivated to support others report lower perceived levels of effectiveness. According to Guo and Johnson, people often downplay the effect of hate speech on themselves compared to its effect on others [45]. One possible explanation

for this could be that people may have a lower threshold for what counts as an effective counterspeech for themselves as opposed to a counterspeech written to defend someone else. Finally, our results show that women report lower levels of perceived effectiveness in their counterspeech as opposed to men. Recent research suggests how women who are targeted by online harassment become more cautious in expressing their opinions publicly, as they tend to normalize harassment, self-censor, or withdraw from online spaces to avoid further harm [18]. Moreover, Wilhelm and Jackoel observed that hate speech authored by females is flagged more frequently than hate speech authored by males[95]. Thus, many women may feel less comfortable expressing themselves in public spaces, leading them to have lower confidence in the effectiveness of their responses.

5.2 Development of a New Survey Scale for Examining Counterspeech Motivations

While previous research in the field of cyberbullying has developed various scales to assess bystander motivations [85, 89], most of these scales lack generalizability to the context of online counterspeech due to key differences highlighted in prior research [77]. Furthermore, such studies are often based on children and adolescents [5, 11], while our study focuses on adults. To the best of our knowledge, our work is the first to provide a comprehensive set of survey scales for understanding counterspeech motivations. Future researchers can thus use these variable items to examine relationships with other salient variables that are yet to be examined.

5.3 The Role of AI Assistance in Counterspeech Writing: AI-Mediated Counterspeech

AI-mediated communication refers to interpersonal interactions enhanced or even generated by algorithms, aiming at specific communicative or relational outcomes [50]. Trust, a cornerstone of human relationships, is crucial for collaborative behaviors like depending on and sharing information. Prior research, such as [50, 55], has extensively explored people's perceptions of the trustworthiness of AI-generated messages. Our qualitative findings provide further nuance to prior research by revealing a tension between individuals' motivations and hesitations in using AI for crafting online counterspeech. While the utilitarian advantages motivate adoption, concerns mainly revolve around issues of trust. Participants commonly expressed distrust in AI's ability to accurately convey emotions and the credibility of presented information. Building on research showing decreased trust with higher AI agency [55, 61], we propose the term "AI-mediated counterspeech" as an alternative to "AI-generated counterspeech."

5.4 Design Implications

5.4.1 Empowering Users by Customizing for Authenticity

A notable finding in our study is that people who have used ChatGPT before are less inclined to use AI to defend themselves. Our qualitative analysis suggests this reluctance stems from a perceived absence of the human element in AI-generated content. For example, one participant states that "[They] are not sure if [AI] could fully express what they want to express, and that it may lack nuance." In addition, another participant states that "[They]

could write it better because [they] can use personal experiences and [their] emotions to hopefully make the perpetrator really think about it." In order to cater to the needs of these users, AI-mediated counterspeech writing tools could provide users with a dropdown menu in which they can select the dominant emotion(s) that they want to convey in their counterspeech. In addition, the tool could provide a text box in which users can add any personal narratives or facts that they would want the AI to incorporate into the AI-generated counterspeech.

5.4.2 Towards Better Understanding of Human-AI Collaboration in Co-Writing Online Counterspeech

Research shows that when users work together with AI toward a common goal, they may treat AI as a collaborative partner instead of a tool [44, 94]. Likewise, our qualitative analysis in response to RQ3b reveals that participants' inclination to utilize AI for counterspeech writing often hinges on their expectations of the AI's role and the extent of AI involvement in the writing process. For instance, many participants expressed a preference for AI assistance in brainstorming ideas rather than having AI autonomously generate the counterspeech. For users with such preferences, Language Model (LLM)-powered AI systems can be designed to facilitate brainstorming sessions. This might involve users inputting words or phrases as fragments of their thoughts or sharing personalized experiences, with the system providing feedback and suggestions based on its training and understanding of constructive counterspeech.

Chapter 6

Conclusions

Our research investigates the various motivations that drive people to engage in counterspeech online. We have created and confirmed the reliability of a comprehensive scale that measures these motivations, revealing its strong impact on both how often indivudials write counterspeech and their views of the impact of their own counterspeech. Our study also sheds light on demographic differences in how people perceive the effectiveness of their counterspeech, which is an area that has not been extensively explored before. Additionally, this is the first study to examine what influences people's willingness to use AI tools to assist them in writing counterspeech.

Our results suggest a close connection between how often people engage in counterspeech and their perceptions of its effectiveness. Those who consider counterspeech to be effective are often motivated by a desire to combat hate and promote inclusivity, while those skeptical of its impact tend to participate in support of others (RQ1). Additionally, demographic factors significantly influence peoples' perceptions of their own counterspeech (RQ2); women tend to perceive their counterspeech as less effective, whereas minority groups report greater effectiveness. Thirdly, we observed that prior usage of ChatGPT positively influences individuals' willingness to use AI assistance in composing online counterspeech (RQ3). Furthermore, participants with previous experience using ChatGPT are less inclined to use AI assistance to write counterspeech when it comes to defending and support themselves, but show greater willingness to use it to to signal inclusion.

The insights provided by our findings are crucial for technology companies and researchers to enhance their understanding of AI's emerging role in supporting users in addressing hate speech on digital platforms.

Chapter 7

Limitations

We used a survey as our data source, which has some limitations. The participants' self-reported responses may not match their actual motivations, or perceptions in a real social media context. Moreover, while Likert scales are a commmon measure used in social science research [37], people may have different interpretations of what counts as "frequent" or "effective". Additionally, given that each participant was asked to write 3 counterspeech responses to 3 hateful posts of differing (random) categories, these differing categories may influence how people perceive and respond to hatefulness. Finally, while we tried to keep demographic bias to a minimum we were restricted to the availability on Prolific.

Bibliography

[1] Collective Efficacy Beliefs:Theoretical Developments, Empirical Evidence, and Future Directions - Roger D. Goddard, Wayne K. Hoy, Anita Woolfolk Hoy, 2004, . URL https://journals.sagepub.com/doi/abs/10.3102/0013189X033003003?casa_token=WhloptsDyKUAAAAA:2aF6FBf8_bz5-k3xtgDUqRNBSbjPHbpXngdBENwo6bAM9ccQets6maabRwno-slIEMs1DhA-n10.

[2] The Effect of Narrative News Format on Empathy for Stigmatized Groups - Mary Beth Oliver, James Price Dillard, Keunmin Bae, Daniel J. Tamul, 2012, . URL https://journals.sagepub.com/doi/abs/10.1177/1077699012439020.

[3] Poe's AI chatbot: testing Quora's universal AI messaging app for ChatGPT and more - The Verge, . URL https://www.theverge.com/23674656/poe-ai-chatbot-messaging-app.

[4] Self-supervision and Controlling Techniques to Improve Counter Speech Generation | Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, . URL https://dl.acm.org/doi/abs/10.1145/3539597.3572991?casa_token=OnKlC0wU1t0AAAAA:PIpUMvtv_FVU2pXCGgdeNFipL9U-8SqCGWsdtOCmz_SsJoSS6Zyf1CkePrd-6GdSWzo9TnVJ7Aw.

[5] Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech - Magdalena Obermaier, 2022, . URL https://journals.sagepub.com/doi/full/10.1177/14614448221125417.

[6] Nextdoor Is Integrating Generative AI to Drive Engaging and Kind Conversa-

tions in the Neighborhood, May 2023. URL https://finance.yahoo.com/news/nextdoor-integrating-generative-ai-drive-103000201.html.

[7] Dane Acena and Guo Freeman. "In My Safe Space": Social Support for LGBTQ Users in Social Virtual Reality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–6, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8095-9. doi: 10.1145/3411763.3451673. URL https://dl.acm.org/doi/10.1145/3411763.3451673.

[8] Shuaa Aljasir. Effect of online civic intervention and online disinhibition on online hate speech among digital media users. *Online Journal of Communication and Media Technologies*, 13(4):e202344, October 2023. ISSN 1986-3497. doi: 10.30935/ojcmt/13478. URL https://www.ojcmt.net/article/effect-of-online-civic-intervention-and-online-disinhibition-on-online-hate-speech-

[9] Kimberley R. Allison and Kay Bussey. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, 65:183–194, June 2016. ISSN 0190-7409. doi: 10.1016/j.childyouth.2016.03.026. URL https://www.sciencedirect.com/science/article/pii/S0190740916300913.

[10] Zahra Ashktorab and Jessica Vitak. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3895–3905, New York, NY, USA, May 2016. Association for Computing Machinery. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858548. URL https://dl.acm.org/doi/10.1145/2858036.2858548.

[11] Vimala Balakrishnan. Actions, emotional reactions and cyberbullying – From the lens of bullies, victims, bully-victims and bystanders among Malaysian young adults.

Telematics and Informatics, 35(5):1190–1200, August 2018. ISSN 0736-5853. doi: 10.1016/j.tele.2018.02.002. URL https://www.sciencedirect.com/science/article/pii/S0736585317308936.

[12] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):24:1–24:19, December 2017. doi: 10.1145/3134659. URL https://dl.acm.org/doi/10.1145/3134659.

[13] Nicholas Brody and Anita L. Vangelisti. Bystander Intervention in Cyberbullying. *Communication Monographs*, 83(1):94–119, January 2016. ISSN 0363-7751. doi: 10.1080/03637751.2015.1044256. URL https://doi.org/10.1080/03637751.2015.1044256. Publisher: Routledge _eprint: https://doi.org/10.1080/03637751.2015.1044256.

[14] Amy S. Bruckman, Jennifer E. Below, Lucas Dixon, Casey Fiesler, Eric E. Gilbert, Sarah A. Gilbert, and J. Nathan Matias. Managing Deviant Behavior in Online Communities III. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages 1–4, New York, NY, USA, April 2018. Association for Computing Machinery. ISBN 978-1-4503-5621-3. doi: 10.1145/3170427.3186319. URL https://dl.acm.org/doi/10.1145/3170427.3186319.

[15] Catherine Buerger. Why They Do It: Counterspeech Theories of Change. *SSRN Electronic Journal*, 2022. ISSN 1556-5068. doi: 10.2139/ssrn.4245211. URL https://www.ssrn.com/abstract=4245211.

[16] Gustavo Carlo, Maria Vicenta Mestre, Meredith M. McGinley, Paula Samper, Ana Tur, and Deanna Sandman. The interplay of emotional instability, empathy, and coping on prosocial and aggressive behaviors. *Personality and Individual Differences*, 53(5):

675–680, October 2012. ISSN 0191-8869. doi: 10.1016/j.paid.2012.05.022. URL https://www.sciencedirect.com/science/article/pii/S0191886912002589.

[17] Huseyin Cavusoglu, Zhuolun Li, and Ke-Wei Huang. Can Gamification Motivate Voluntary Contributions? The Case of StackOverflow Q&A Community. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, CSCW'15 Companion, pages 171–174, New York, NY, USA, February 2015. Association for Computing Machinery. ISBN 978-1-4503-2946-0. doi: 10.1145/2685553.2698999. URL https://dl.acm.org/doi/10.1145/2685553.2698999.

[18] Kalyani Chadha, Linda Steiner, Jessica Vitak, and Zahra Ashktorab. Women's Responses to Online Harassment. *International Journal of Communication*, 14(0):19, January 2020. ISSN 1932-8036. URL https://ijoc.org/index.php/ijoc/article/view/11683. Number: 0.

[19] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1201–1213, New York, NY, USA, February 2016. Association for Computing Machinery. ISBN 978-1-4503-3592-8. doi: 10.1145/2818048.2819963. URL https://dl.acm.org/doi/10.1145/2818048.2819963.

[20] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, December 2017. ISSN 2573-0142. doi: 10.1145/3134666. URL https://dl.acm.org/doi/10.1145/3134666.

[21] Adrian Chen. Conversion via Twitter. *The New Yorker*, November 2015. ISSN 0028-792X. URL https://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper. Section: dept. of technology.

[22] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108–118, May 2018. ISSN 1359-1789. doi: 10.1016/j.avb.2018.05.003. URL https://www.sciencedirect.com/science/article/pii/S1359178917301064.

[23] Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. Multilingual Counter Narrative Type Classification, September 2021. URL http://arxiv.org/abs/2109.13664. arXiv:2109.13664 [cs].

[24] Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. Understanding Counterspeech for Online Harm Mitigation, July 2023. URL http://arxiv.org/abs/2307.04761. arXiv:2307.04761 [cs].

[25] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565.

[26] Raphael Cohen-Almagor. Fighting Hate and Bigotry on the Internet. *Policy & Internet*, 3(3):1–26, 2011. ISSN 1944-2866. doi: 10.2202/1944-2866.1059. URL

https://onlinelibrary.wiley.com/doi/abs/10.2202/1944-2866.1059. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2202/1944-2866.1059.

[27] Matthew Costello, James Hawdon, and Amanda Cross. Virtually Standing Up or Standing By? Correlates of Enacting Social Control Online. *International Journal of Criminology and Sociology*, 6:16–28, February 2016. ISSN 1929-4409. doi: 10.6000/1929-4409.2017.06.03. URL https://lifescienceglobal.com/pms/index.php/ijcs/article/view/4409.

[28] Matthew Costello, James Hawdon, and Thomas N. Ratliff. Confronting Online Extremism: The Effect of Self-Help, Collective Efficacy, and Guardianship on Being a Target for Hate Speech. *Social Science Computer Review*, 35(5):587–605, October 2017. ISSN 0894-4393. doi: 10.1177/0894439316666272. URL https://doi.org/10.1177/0894439316666272. Publisher: SAGE Publications Inc.

[29] Esther Cuadrado, Carmen Tabernero, and Wolfgang Steinel. Determinants of Prosocial Behavior in Included Versus Excluded Contexts. *Frontiers in Psychology*, 6, 2016. ISSN 1664-1078. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2015.02001.

[30] Niklas Felix Cypris, Severin Engelmann, Julia Sasse, Jens Grossklags, and Anna Baumert. Intervening Against Online Hate Speech: A Case for Automated Counterspeech.

[31] Lincoln Dahlberg. The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication & Society*, 4(4):615–633, January 2001. ISSN 1369-118X. doi: 10.1080/13691180110097030. URL https://doi.org/10.1080/13691180110097030. Publisher: Routledge _eprint: https://doi.org/10.1080/13691180110097030.

[32] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D. Wherry, and Natalya N. Bazarova. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Montreal QC Canada, April 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173785. URL https://dl.acm.org/doi/10.1145/3173574.3173785.

[33] Evelyn Douek. Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability, August 2020. URL https://papers.ssrn.com/abstract=3679607.

[34] Moli Dow and Ross Frenett. One to One Online Interventions A pilot CVE methodology. January 2014. URL https://www.academia.edu/16479356/One_to_One_Online_Interventions_A_pilot_CVE_methodology.

[35] David Easley and Arpita Ghosh. Incentives, Gamification, and Game Theory: An Economic Approach to Badge Design. *ACM Transactions on Economics and Computation*, 4(3):16:1–16:26, June 2016. ISSN 2167-8375. doi: 10.1145/2910575. URL https://dl.acm.org/doi/10.1145/2910575.

[36] Arthur Edwards. Bowling Together. Online Public Engagement in Policy Deliberation, by Stephen Coleman and John Gøtze. *Information Polity*, 7:247–252, December 2002. doi: 10.3233/IP-2002-0021.

[37] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2007.00367.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2007.00367.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1083-6101.2007.00367.x.

[38] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to Peer Hate: Hate Speech Instigators and Their Targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), June 2018. ISSN 2334-0770. doi: 10.1609/icwsm.v12i1.15038. URL https://ojs.aaai.org/index.php/ICWSM/article/view/15038. Number: 1.

[39] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.250. URL https://aclanthology.org/2021.acl-long.250.

[40] Ross Frenett. One to One Online Interventions.

[41] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Countering hate on social media: Large scale classification of hate and counter speech, June 2020. URL http://arxiv.org/abs/2006.01974. arXiv:2006.01974 [cs].

[42] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):3, January 2022. ISSN 2193-1127. doi: 10.1140/epjds/s13688-021-00314-6. URL https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-021-00314-6.

[43] Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros Fernandez, Sarah T. Roberts, Aram Sinnreich, and Sarah My-

ers West. Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates, May 2023. URL https://papers.ssrn.com/abstract=4459448.

[44] Jonathan Grudin. *From Tool to Partner: The Evolution of Human-Computer Interaction*. Springer Nature, May 2022. ISBN 978-3-031-02218-0. Google-Books-ID: Z4lyEAAAQBAJ.

[45] Lei Guo and Brett G. Johnson. Third-Person Effect and Hate Speech Censorship on Facebook. *Social Media + Society*, 6(2):2056305120923003, April 2020. ISSN 2056-3051. doi: 10.1177/2056305120923003. URL https://doi.org/10.1177/2056305120923003. Publisher: SAGE Publications Ltd.

[46] Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118, December 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2116310118. URL https://pnas.org/doi/full/10.1073/pnas.2116310118.

[47] James Hawdon, Atte Oksanen, and Pekka Räsänen. Exposure to Online Hate in Four Nations: A Cross-National Consideration. *Deviant Behavior*, 38(3):254–266, March 2017. ISSN 0163-9625. doi: 10.1080/01639625.2016.1196985. URL https://doi.org/10.1080/01639625.2016.1196985. Publisher: Routledge _eprint: https://doi.org/10.1080/01639625.2016.1196985.

[48] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Ku-

mar. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, Virtual Event Netherlands, November 2021. ACM. ISBN 978-1-4503-9128-3. doi: 10.1145/3487351.3488324. URL https://dl.acm.org/doi/10.1145/3487351.3488324.

[49] Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus*, 2(3):pgad035, March 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad035. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10029837/.

[50] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300469. URL https://dl.acm.org/doi/10.1145/3290605.3300469.

[51] Teo Keipi, Matti Näsi, Atte Oksanen, and Pekka Räsänen. *Online Hate and Harmful Content: Cross-National Perspectives*. Taylor & Francis, 2016. ISBN 978-1-317-24084-6 978-0-367-87696-8 978-1-315-62837-0 978-1-138-64506-6. doi: 10.4324/9781315628370. URL https://library.oapen.org/handle/20.500.12657/22350. Accepted: 2020-03-10 09:29:55.

[52] Marlene Kunst, Pablo Porten-Cheé, Martin Emmer, and Christiane Eilders. Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, 18(3): 258–273, July 2021. ISSN 1933-1681. doi: 10.1080/19331681.2020.1871149. URL

https://doi.org/10.1080/19331681.2020.1871149. Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2020.1871149.

[53] Huije Lee, Young Ju NA, Hoyun Song, Jisu Shin, and Jong C. Park. ELF22: A Context-based Counter Trolling Dataset to Combat Internet Trolls, September 2022. URL http://arxiv.org/abs/2208.00176. arXiv:2208.00176 [cs].

[54] Larissa Leonhard, Christina Rueß, Magdalena Obermaier, and Carsten Reinemann. Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Studies in Communication and Media*, 7(4): 555–579, December 2018. ISSN 2192-4007. doi: 10.5771/2192-4007-2018-4-555. URL https://www.nomos-elibrary.de/10.5771/2192-4007-2018-4-555/perceiving-threat-and-feeling-responsible-how-severity-of-hate-speech-number-of-bys page=1. Publisher: Nomos Verlagsgesellschaft mbH & Co. KG.

[55] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, New Orleans LA USA, April 2022. ACM. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3517731. URL https://dl.acm.org/doi/10.1145/3491102.3517731.

[56] Hana Machackova, Lenka Dedkova, and Katerina Mezulanikova. Brief report: The bystander effect in cyberbullying incidents. *Journal of Adolescence*, 43:96–99, August 2015. ISSN 0140-1971. doi: 10.1016/j.adolescence.2015.05.010. URL https://www.sciencedirect.com/science/article/pii/S0140197115001049.

[57] Regan L Mandryk, Julian Frommel, Nitesh Goyal, Guo Freeman, Cliff Lampe, Sarah Vieweg, and Donghee Yvette Wohn. Combating Toxicity, Harassment, and Abuse in

Online Social Spaces: A Workshop at CHI 2023. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, pages 1–7, New York, NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9422-2. doi: 10.1145/3544549.3573793. URL https://dl.acm.org/doi/10.1145/3544549.3573793.

[58] Gina Masullo Chen and Shuning Lu. 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, January 2017. ISSN 0883-8151. doi: 10.1080/08838151.2016.1273922. URL https://doi.org/10.1080/08838151.2016.1273922. Publisher: Routledge _eprint: https://doi.org/10.1080/08838151.2016.1273922.

[59] Ariadna Matamoros-Fernández and Johan Farkas. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, 22(2):205–224, February 2021. ISSN 1527-4764, 1552-8316. doi: 10.1177/1527476420982230. URL http://journals.sagepub.com/doi/10.1177/1527476420982230.

[60] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media*, 13:369–380, July 2019. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v13i01.3237. URL https://ojs.aaai.org/index.php/ICWSM/article/view/3237.

[61] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate begets Hate: A Temporal Study of Hate Speech. *Proceedings*

of the ACM on Human-Computer Interaction, 4(CSCW2):1–24, October 2020. ISSN 2573-0142. doi: 10.1145/3415163. URL https://dl.acm.org/doi/10.1145/3415163.

[62] Christian Meske and Enrico Bunde. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*, 25(2):743–773, April 2023. ISSN 1572-9419. doi: 10.1007/s10796-021-10234-5. URL https://doi.org/10.1007/s10796-021-10234-5.

[63] Jozef Miškolci, Lucia Kováčová, and Edita Rigová. Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review*, 38(2):128–146, April 2020. ISSN 0894-4393. doi: 10.1177/0894439318791786. URL https://doi.org/10.1177/0894439318791786. Publisher: SAGE Publications Inc.

[64] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. ETHOS: an Online Hate Speech Detection Dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, December 2022. ISSN 2199-4536, 2198-6053. doi: 10.1007/s40747-021-00608-2. URL http://arxiv.org/abs/2006.08328. arXiv:2006.08328 [cs, stat].

[65] Kevin Munger. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, 39(3):629–649, September 2017. ISSN 1573-6687. doi: 10.1007/s11109-016-9373-5. URL https://doi.org/10.1007/s11109-016-9373-5.

[66] Teresa K Naab, Anja Kalch, and Tino GK Meitz. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, 20(2):777–795, February 2018. ISSN 1461-4448. doi: 10.1177/1461444816670923. URL https://doi.org/10.1177/1461444816670923. Publisher: SAGE Publications.

[67] Magdalena Obermaier, Desirée Schmuck, and Muniba Saleem. 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*, page 146144482110175, August 2021. ISSN 1461-4448, 1461-7315. doi: 10.1177/14614448211017527. URL http://journals.sagepub.com/doi/10.1177/14614448211017527.

[68] Magdalena Obermaier, Ursula Kristin Schmid, and Diana Rieger. Too civil to care? How online hate speech against different social groups affects bystander intervention. *European Journal of Criminology*, 20(3):817–833, May 2023. ISSN 1477-3708. doi: 10.1177/14773708231156328. URL https://doi.org/10.1177/14773708231156328. Publisher: SAGE Publications.

[69] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1474. URL https://aclanthology.org/D19-1474.

[70] Robert M. O'brien. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5):673–690, October 2007. ISSN 1573-7845. doi: 10.1007/s11135-006-9018-6. URL https://doi.org/10.1007/s11135-006-9018-6.

[71] Jessica A. Pater, Oliver L. Haimson, Nazanin Andalibi, and Elizabeth D. Mynatt. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1185–1200, New York, NY, USA, February 2016. Association for Computing Machinery. ISBN 978-1-4503-3592-8. doi: 10.1145/2818048.2820030. URL https://dl.acm.org/doi/10.1145/2818048.2820030.

[72] Pablo Porten-Cheé, Marlene Kunst, and Martin Emmer. Online Civic Intervention: A New Form of Political Participation Under Conditions of a Disruptive Online Discourse. *International Journal of Communication*, 14(0):21, January 2020. ISSN 1932-8036. URL https://ijoc.org/index.php/ijoc/article/view/10639. Number: 0.

[73] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A Benchmark Dataset for Learning to Intervene in Online Hate Speech, September 2019. URL http://arxiv.org/abs/1909.04251. arXiv:1909.04251 [cs].

[74] Casey Randazzo and Tawfiq Ammari. "If Someone Downvoted My Posts—That'd Be the End of the World": Designing Safer Online Spaces for Trauma Survivors. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–18, New York, NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581453. URL https://dl.acm.org/doi/10.1145/3544548.3581453.

[75] Robert D. Richards and Clay Calvert. Counterspeech 2000: A New Look at the Old Remedy for Bad Speech. *Brigham Young University Law Review*, 2000:553, 2000. URL https://heinonline.org/HOL/Page?handle=hein.journals/byulr2000&id=563&div=&collection=.

[76] Diana Rieger, Josephine B. Schmitt, and Lena Frischlich. Hate and counter-voices in the Internet: Introduction to the special issue. *SCM Studies in Communication and Media*, 7(4):459–472, December 2018. ISSN 2192-4007. doi: 10.5771/2192-4007-2018-4-459. URL https://www.nomos-elibrary.de/10.5771/2192-4007-2018-4-459/hate-and-counter-voices-in-the-internet-introduction-to-the-special-issue-jahrgang-page=1. Publisher: Nomos Verlagsgesellschaft mbH & Co. KG.

[77] Konrad Rudnicki, Heidi Vandebosch, Pierre Voué, and Karolien Poels. Systematic

review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*, 42(5):527–544, April 2023. ISSN 0144-929X, 1362-3001. doi: 10.1080/0144929X.2022.2027013. URL https://www.tandfonline.com/doi/full/10.1080/0144929X.2022.2027013.

[78] Derek Ruths Ruths, Haji Mohammad Saleem Saleem, Kelly P. Dillon Dillon, Lucas Wright Wright, and Susan Benesch Benesch. Considerations for Successful Counterspeech. Technical report, Dangerous Speech Project, Washington, DC USA, October 2016. URL https://www.issuelab.org/permalink/download/34065.

[79] Leonie Rösner and Nicole C. Krämer. Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. *Social Media + Society*, 2(3):2056305116664220, July 2016. ISSN 2056-3051. doi: 10.1177/2056305116664220. URL https://doi.org/10.1177/2056305116664220. Publisher: SAGE Publications Ltd.

[80] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5157–5163, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/716. URL https://www.ijcai.org/proceedings/2022/716.

[81] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):155:1–155:27, November 2018. doi: 10.1145/3274424. URL https://dl.acm.org/doi/10.1145/3274424.

[82] Carla Schieb and Mike Preuss. Governing Hate Speech by Means of Counter Speech on Facebook.

[83] Ulrike Schwertberger and Diana Rieger. Hass und seine vielen Gesichter: Eine sozial- und kommunikationswissenschaftliche Einordnung von Hate Speech. In Sebastian Wachs, Barbara Koch-Priewe, and Andreas Zick, editors, *Hate Speech - Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen*, pages 53–77. Springer Fachmedien, Wiesbaden, 2021. ISBN 978-3-658-31793-5. doi: 10.1007/978-3-658-31793-5_4. URL https://doi.org/10.1007/978-3-658-31793-5_4.

[84] Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 111–125, New York, NY, USA, February 2017. Association for Computing Machinery. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998277. URL https://dl.acm.org/doi/10.1145/2998181.2998277.

[85] Emily Shultz, Rebecca Heilman, and Kathleen J. Hart. Cyber-bullying: An exploration of bystander behavior and motivation. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(4), December 2014. ISSN 1802-7962. doi: 10.5817/CP2014-4-3. URL https://cyberpsychology.eu/article/view/4324.

[86] Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4):376–385, 2008. ISSN 1469-7610. doi: 10.1111/j.1469-7610.2007.01846.x. URL https://

onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7610.2007.01846.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-7610.2007.01846.x.

[87] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44(2):136–146, 2018. ISSN 1098-2337. doi: 10.1002/ab.21737. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21737. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ab.21737.

[88] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445092. URL https://dl.acm.org/doi/10.1145/3411764.3445092.

[89] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, November 2019. ISSN 2573-0142. doi: 10.1145/3359220. URL https://dl.acm.org/doi/10.1145/3359220.

[90] Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study, April 2022. URL http://arxiv.org/abs/2204.01440. arXiv:2204.01440 [cs].

[91] Robert S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26(3):277–

287, May 2010. ISSN 0747-5632. doi: 10.1016/j.chb.2009.11.014. URL https://www.sciencedirect.com/science/article/pii/S074756320900185X.

[92] Choice Ubangha. Hate Speech in Cyberspace: Why Education is Better than Regulation, April 2016. URL https://papers.ssrn.com/abstract=2865053.

[93] Sebastian Wachs, Angela Mazzone, Tijana Milosevic, Michelle F. Wright, Catherine Blaya, Manuel Gámez-Guadix, and James O'Higgins Norman. Online correlates of cyberhate involvement among young people from ten European countries: An application of the Routine Activity and Problem Behaviour Theory. *Computers in Human Behavior*, 123:106872, October 2021. ISSN 0747-5632. doi: 10.1016/j.chb.2021.106872. URL https://www.sciencedirect.com/science/article/pii/S0747563221001953.

[94] Dakuo Wang, Pattie Maes, Xiangshi Ren, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. Designing AI to Work WITH or FOR People? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–5, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8095-9. doi: 10.1145/3411763.3450394. URL https://dl.acm.org/doi/10.1145/3411763.3450394.

[95] Claudia Wilhelm, Sven Joeckel, and Isabell Ziegler. Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users' Moral Orientation. *Communication Research*, 47(6):921–944, August 2020. ISSN 0093-6502. doi: 10.1177/0093650219855330. URL https://doi.org/10.1177/0093650219855330. Publisher: SAGE Publications Inc.

[96] Randy Yee Man Wong, Christy M. K. Cheung, Bo Xiao, and Jason Bennett Thatcher. Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. *Information Systems Research*, 32(2):561–581,

June 2021. ISSN 1047-7047. doi: 10.1287/isre.2020.0983. URL https://pubsonline.informs.org/doi/abs/10.1287/isre.2020.0983. Publisher: INFORMS.

[97] Scott Wright and John Street. Democracy, deliberation and design: the case of online discussion forums. *New Media & Society*, 9(5):849–869, October 2007. ISSN 1461-4448. doi: 10.1177/1461444807081230. URL https://doi.org/10.1177/1461444807081230. Publisher: SAGE Publications.

[98] Qianqian Zhang, Wenxuan Li, Mengnan Zhang, Kai Zhao, Running Liu, and Chong Ma. Internet altruistic behavior among Chinese early adolescents: Exploring differences in gender and collective efficacy using a latent growth modeling. *Current Psychology*, May 2023. ISSN 1936-4733. doi: 10.1007/s12144-023-04660-8. URL https://doi.org/10.1007/s12144-023-04660-8.

[99] Wanzheng Zhu and Suma Bhat. Generate, Prune, Select: A Pipeline for Counter-speech Generation against Online Hate Speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.12. URL https://aclanthology.org/2021.findings-acl.12.

Appendices

Appendix A

Appendix

A.1 Participant Demographics

Table A.1: Participant Demographics ($N = 458$)

Factor	Category	N
Age group	18-30	138
	31-60	293
	61-81	27
Gender	Male	226
	Female	232
	Prefer Not To Say	0
Ethnicity	**Majority**	234
	White	234
	Minority	224
	Asian	55
	Black	110
	Hispanic	53
	Other	6
Education level	Less than high school or high school graduate	65
	Some college or 2 year degree	154
	4 year degree or higher	239
Sexual orientation	Heterosexual	359
	Non-Heterosexual	99
Political views	Very conservative	28
	Conservative	91
	Moderate	121
	Liberal	134
	Very Liberal	84

A.2 Survey Questions

Motivations: How much do the following factors motivate you to write a counterspeech on social media? 1 (None at all), 2 (A little), 3 (A moderate amount), 4 (A lot), 5 (A great deal)

Social Media Commenting Frequency: How often do you comment on content that you encounter on social media? 1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)

Use Real Name: Do you use your real name on social media? 1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)

Prior Experience of Online Hate Speech: Have you been a target of hateful speech on the internet? 1 (No), 2 (Yes)

Frequency of Encountering Online Hate Speech: I encounter content that I find hateful on social media: 1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)

Perceived General Efficacy of Counterspeech: How effective do you think counterspeech is in reducing online hate speech? 1 (Not effective at all), 2 (Slightly effective), 3 (Moderately effective), 4 (Very effective), 5 (Extremely effective)

Frequency of Writing Counterspeech: How often do you write counterspeech on social media? 1 (Never), 2 (Rarely), 3 (Sometimes), 4 (Often), 5 (Always)

Perceived Effectiveness of Self-Written Counterspeech: How effective do you think your counterspeech would be in preventing the perpetrator from engaging in further hateful behavior? 1 (Not effective at all), 2 (Slightly effective), 3 (Moderately effective), 4 (Very effective), 5 (Extremely effective)

Willingness to use ChatGPT to write counterspeech: If you were writing a coun-

terspeech on social media, would you use artificial intelligence technology like ChatGPT to help you write it? 1 (Definitely Not), 2 (Probably Not), 3 (Might or Might not), 4 (Probably Yes, 5 (Definitley Yes)