Behind the Counter: Exploring the Motivations and Barriers of Online Counterspeech Writing

KAIKE PING*
Computer Science, Virginia Tech, Blacksburg, Virginia, United States, pkk@vt.edu

ANISHA KUMAR*
Computer Science, Virginia Tech, Blacksburg, Virginia, United States, anishak@vt.edu

XIAOHAN DING
Computer Science, Virginia Tech, Blacksburg, Virginia, United States, xiaohan@vt.edu

EUGENIA H. RHO#
Computer Science, Virginia Tech, Blacksburg, Virginia, United States, eugenia@vt.edu

Current research mainly explores the attributes and impact of online counterspeech, leaving a gap in understanding of who engages in online counterspeech or what motivates or deters users from participating. To investigate this, we surveyed 458 English-speaking U.S. participants, analyzing key motivations and barriers underlying online counterspeech engagement. We presented each participant with three hate speech examples from a set of 900, spanning race, gender, religion, sexual orientation, and disability, and requested counterspeech responses. Subsequent questions assessed their satisfaction, perceived difficulty, and the effectiveness of their counterspeech. Our findings show that having been a target of online hate is a key driver of frequent online counterspeech engagement. People differ in their motivations and barriers towards engaging in online counterspeech across different demographic groups. Younger individuals, women, those with higher education levels, and regular witnesses to online hate are more reluctant to engage in online counterspeech due to concerns around public exposure, retaliation, and third-party harassment. Varying motivation and barriers in counterspeech engagement also shape how individuals view their own self-authored counterspeech and the difficulty experienced writing it. Additionally, our work explores people's willingness to use AI technologies like ChatGPT for counterspeech writing. Through this work we introduce a multi-item scale for understanding counterspeech motivation and barriers and a more nuanced understanding of the factors shaping online counterspeech engagement.

CCS CONCEPTS • Human-centered computing • Collaborative and social computing • Empirical studies in collaborative and social computing

Additional Keywords and Phrases: Behavior Change, Social Media/Online Communities, Empirical study that tells us about people, Method, Qualitative Methods, Quantitative Methods, Survey

1 INTRODUCTION

In today's digital age, social media platforms serve as key spaces for public discourse [46, 58, 131, 181]. While these platforms enable swift dissemination of ideas, they also serve as cultivators for hate speech [31, 113], cyberbullying [7], and harassment [30, 133]. The effectiveness of mitigating online hate through moderation by human moderators and automated systems can vary [68, 72]. Deletion or banning users can sometimes disperse rather than dispel hateful speech [31], or potentially conflict with First Amendment rights in the United States

* These authors contributed equally to this work.

Corresponding author.

ACM 1557-7325/2025/6-ART
https://doi.org/10.1145/3745769

[75]. For example, the practice of deplatforming users for sharing hateful views can simply push offenders to less regulated online spaces [85]. While tech companies continue to combat online hate through traditional moderation methods, limitations in these approaches have led scholars to examine the potential for user-driven counterspeech [22, 54, 134, 149].

Counterspeech is defined as direct responses to derogatory or harmful content, intended to undermine or refute hateful messages [142, 146]. Since 2016, tech companies like Meta (Facebook), Google (YouTube), and Twitter have partnered with NGOs around the world to foster counterspeech initiatives against online hate [99, 150]. Such focus on user-driven counterspeech efforts highlights the significance of individual and community roles in regulating online spaces [153]. For instance, the transformation of Megan Phelps-Roper, a former member of the extremist Westboro Baptist Church, stands as a testament to the profound impact that counterspeech can have [32]. Phelps-Roper was a 23-year-old legal assistant who regularly posted on Twitter on behalf of the Westboro Baptist Church, which is widely considered as a hate group [175]. In response to her hateful tweets against Jews, David Abitbol, a 50-year-old Jerusalem-based web developer decided to directly engage with her on the platform. Instead of mirroring her hostility or mocking her, Abitbol responded with humor, empathy, and questions, aiming to humanize those she vilified. His constructive counter engagement not only challenged Phelps-Roper's antisemitic views, but also led to a complete reversal of her stance [121]. What started out as a mere tweet in response to a message rooted in hatred gradually undermined and dismantled Phelps-Roper's convictions. This specific case affirms the powerful role that counterspeech can play to enact positive change against online hate. However, successfully engaging in online counterspeech can pose various challenges for individuals. People may vary in their individual motivations and barriers that affect their decision to engage in counterspeech in the first place. Understanding the factors that drive users like Abitbol, and the barriers preventing broader participation in online counterspeech, remains a significant research gap [26, 57, 112].

While counterspeech has shown promise in combating online hate, the increasing prevalence [113] and sophistication [73] of hateful content on social media platforms have made engaging in counterspeech increasingly overwhelming for individual users [35]. As the demand for counterspeech grows, understanding why people engage in or refrain from it becomes crucial for guiding potential supportive measures. Recent studies have begun to explore the potential of artificial intelligence (AI) in assisting counterspeech efforts [101, 137, 147, 166]. However, as Benesch et al. (2016) argue, successful counterspeech is deeply rooted in human empathy, cultural understanding, and personal motivation [15]. Mun et al. (2024) further emphasize that any technological intervention, including AI, should be designed with a thorough understanding of the human aspects of counterspeech [119]. Therefore, before considering how to support counterspeech through AI, it is essential to examine the human factors that drive or hinder engagement in this practice to ensure that future technological interventions address real user needs and overcome genuine barriers to participation.

1.1 Motivation of Research Questions

This study aims to comprehensively understand online counterspeech against hate speech by examining the motivations and barriers to engagement, users' experiences in writing counterspeech, and potential AI support for this practice. We explore how motivation and barrier factors individually contribute to the landscape of online counterspeech engagement and the willingness to adopt AI tools for counterspeech assistance, while also investigating their overall effects on the counterspeech writing experience.

Within this framework, we first examine the current state of counterspeech research and practice. To date, prior research has primarily focused on understanding the content of online counterspeech [55, 64, 116, 150] and its impact on the broader social media ecosystem [81]. However, there is a dearth of research on what motivates or deters users from participating in it. Our work fills this gap by comprehensively analyzing the motivations for and barriers to engaging in online counterspeech, drawing from existing literature and theoretical frameworks. Through a survey, we explore how these identified factors influence actual user behavior. Specifically, we examine how underlying motivations and barriers influence the reported frequency of engaging in online counterspeech (**RQ1**).

Demographic factors, such as age [92, 139, 173], gender [92, 139, 169], and race [72, 76] are associated with how people interact online. In the context of counterspeech, research has shown that a counter-speaker's race

[120] can influence how others perceive the effectiveness of the counter-speaker's attempt to counter a hateful post [120, 154]. While such studies are a promising start, there remains a lack of knowledge on how broader demographic variables shape people's underlying motivations and barriers that influence people's engagement in online counterspeech. Our work addresses this gap by comprehensively examining how demographic factors affect people's motivations and barriers to engaging in online counterspeech across a wide variety of topics **(RQ2)**.

Furthermore, how people feel about their counterspeech [23], the difficulty experienced when writing it [83], and their perception of its effectiveness [87], can influence their willingness to respond to online hate. For example, users' **satisfaction** with their online counterspeech increases when they feel supported by other users in their efforts to challenge hateful actors [23]. For others, the process of writing counterspeech can be daunting even with community support. Simply, the sheer **difficulty** in crafting a counter-message can potentially deter users' willingness to respond to a perpetrator [83]. Similarly, the perceived **effectiveness** of one's counterspeech may affect user's willingness to engage in online counterspeech. The belief that one's words have the power to stop or lessen the offenders' harmful actions can motivate a proactive stance against online hate [87]. Therefore, these factors are also crucial for designing effective interventions to support and encourage engagement. To explore these dynamics, our survey includes a writing task where participants compose counterspeech responses within the survey, and we capture their reflections on their writing experiences. Specifically, in this work, we examine how motivations and barriers influence users' experience in responding to hateful posts, namely their perceived **satisfaction** with their counterspeech, the level of **difficulty** they experience in writing it, and their perception of its **effectiveness** (**RQ3**).

Social media platforms are increasingly adopting artificial intelligence (AI) technologies, like large language models (LLMs), to foster more respectful online interactions [132]. Platforms such as Nextdoor and Quora are now using AI to prompt users to revise potentially provocative or policy-violating posts. For instance, Nextdoor has introduced an OpenAI-powered feature that suggests edits to users' posts to prevent inflammatory language [1, 123]. Although these AI tools aim to cultivate safer online discussions, how users perceive them and their impact on interactions with other users remains unclear. To develop effective AI support tools, it is essential to first understand the motivations and barriers that influence users' engagement with counterspeech. Mun et al. (2024) have demonstrated the importance of understanding user barriers when designing AI tools for online interactions [119]. Thus, users' pre-existing attitudes towards counterspeech and their reasons for engaging or avoiding confrontation with hateful actors online may affect their readiness to embrace AI assistance. Expanding on this approach, our study examines not only barriers but also motivations, and how they relate to users' openness to AI assistance in counterspeech (**RQ4**).

In summary, we ask the following research questions:

RQ1. What motivations and barriers influence the frequency of people's engagement in online counterspeech?

RQ2. How do demographic variables shape people's motivations and barriers in online counterspeech engagement?

RQ3. How do people's motivations and barriers in online counterspeech engagement influence:

- *a. how satisfied they are with their counterspeech?*
- *b. how difficult they find it to write counterspeech?*
- *c. how they perceive the effectiveness of their counterspeech?*

RQ4. How are people's motivations and barriers in online counterspeech engagement associated with their willingness to use AI assistance?

To answer these questions, we conducted a pre-registered survey (N = 458) across English-speaking participants in the United States. Our survey examines key motivations and barriers that underlie why people do or do not engage in online counterspeech, their frequency of writing counterspeech on social media, and their willingness to use AI to help them write counterspeech. In addition, in our survey we showed participants three randomly selected hate posts from a pool of 900 online hate speech posts across five topics: race, gender,

religion, sexual orientation, and disability. Participants were asked to respond to each of the three hate posts with a counterspeech. We then asked follow-up questions to understand their perceptions and experience of writing counterspeech (satisfaction, difficulty, and perceived effectiveness of one's counterspeech) in response to the hateful posts.

1.2 Overview of Research Findings and Contributions

Our findings show a significant relationship between exposure to online hate and the likelihood of engaging in online counterspeech. Individuals who have personally encountered online hate are often prompted to use counterspeech as a means to directly confront offenders or as an emotional outlet. In contrast, those less frequently exposed to online hate are more likely to engage in counterspeech to signal inclusion to others, but a key barrier for them is the skill gap in writing counterspeech (RQ1). Demographic factors and social media experiences significantly shape one's motivations and barriers to engage in counterspeech (RQ2). For instance, younger individuals, women, those with higher education levels, and regular witnesses to online hate report more concerns about public exposure, retaliation from the perpetrator, and additional harassment. Notably, even with these concerns, such individuals still find their self-authored counterspeech more effective and satisfying (RQ3).

Those who feel emotionally burdened when writing counterspeech, or who question their ability to write it effectively, often find their self-authored counterspeech less satisfying, more challenging to write, and less effective. This lack of confidence in their ability to craft effective counterspeech is further linked to a higher likelihood of turning to AI for assistance in crafting counterspeech (RQ4). Our results differentiate between users with versus without prior experience using AI tools like ChatGPT. Users with prior experience using AI tools are more inclined to use AI assistance in writing counterspeech to signal inclusivity, but less likely to do so for self-defense. On the other hand, those new to AI tools are more open to using AI for counterspeech writing, especially if they fear retaliation from perpetrators but still wish to confront hate. However, this concern does not translate into using AI-mediated counterspeech writing to defend those who are close to them.

We contribute to HCI research by offering a comprehensive understanding of the various factors that shape why people do or do not engage in online counterspeech. We developed and validated a multi-item measure for motivations and barriers for engaging in online counterspeech, demonstrating its significant influence on both the experience of writing counterspeech and peoples' perceptions towards their self-authored counterspeech. These validated measures allow a structured approach for researchers to examine online counterspeech dynamics for future studies. Second, we extended the scholarly discourse on the demographic and experiential factors that motivate or deter people from engaging in online counterspeech. While prior scholarship has mostly focused on counterspeech strategy and content [55, 64, 116, 150], our work contributes to the knowledge of how social, demographic, and personal experiences impact people's counterspeech writing experiences as well as their motivations and barriers to engage in it. By offering insights into how the motivations and barriers for engaging in counterspeech differ among various social groups, our findings can inform the development of counterspeech tools that are tailored to the needs of diverse users [115]. Finally, we contribute to the understanding of what influences people's openness to use AI assistance for writing online counterspeech. This contribution is particularly timely as the role of AI in moderating online communities becomes increasingly prominent. Such insights can help tech companies and researchers to better envision the potential applications and limitations of AI in assisting users countering online hate.

2 RELATED WORK

2.1 The Role of Counterspeech in Mitigating Online Hate Speech

Counterspeech operates through various mechanisms in mitigating online hate [28]. It can act as a social sanction, increasing the social cost of those who disseminate hate speech and thus discouraging its spread [55, 64, 150]. It can also counter harmful narratives by presenting alternative viewpoints [37, 116]. The efficacy of counterspeech is widely debated [11, 116, 150]. Schieb and Preuss (2016) supports the effectiveness of counterspeech, demonstrating that it can lead to the deletion of hateful posts and even elicit apologies from

hateful actors [150]. Their findings suggest that this effectiveness is amplified when counter-speakers outnumber those spreading hate speech, especially if the online community holds moderate views [150]. By contrast, Miškolci et al. (2018) raises questions about the direct impact of counterspeech, that it does not necessarily deter hateful actors from posting hateful content [116]. However, scholars note that a single counterspeech authored by one user often gains visibility among a wider online audience, thereby serving as a catalyst that inspires onlookers to initiate their own counter-responses [116, 127]. The rhetorical style and tone of counterspeech matters too. For example, counterspeech that adopts an empathetic tone has been shown to be particularly effective in leading to offenders deleting their racist and xenophobic tweets [75]. In addition to tone [75, 120, 140, 148], other counterspeech strategies include fact-sharing [23, 126], open denunciation [148, 164], and posing counter-questions [140, 148, 159].

While these studies contribute to the knowledge of counterspeech characteristics and their effectiveness, they often overlook the social and demographic backgrounds of those who use it. Few studies have examined how a counter-speaker's race and online presence might influence their impact of counterspeech [120, 154]. For instance, Munger et al. experimented on Twitter using bots designed to appear as either black or white individuals with varying levels of online status, as indicated by their follower counts [120]. Their findings revealed that counterspeech from a high-status white man bot led to a significant reduction in the use of racist slurs by the original hate speech authors. Despite these initial insights, little is known about the motivations and barriers that influence why or how often people engage in online counterspeech. Our work seeks to fill this gap.

2.2 Understanding User Motivations and Barriers in Online Counterspeech Engagement

The success of counterspeech as a remedy to hate speech lies in individuals' readiness to act [23, 24]. Yet, current research falls short in examining *why* people decide to engage in counterspeech or opt to stay bystanders. In comparison, bystander motivation and behavior in cyberbullying are well-researched [165], offering valuable insights that inform this current study.

Researchers highlight strong parallels between bystander reactions to cyberbullying and those faced by people encountering online hate [103, 127, 145]. Hate speech attacks individuals on the basis of social identifiers such as race, gender, and sexual orientation [78, 94, 152]. This differs[1] from cyberbullying, which is characterized by derogating or threatening individuals without necessarily disparaging their social identity [170]. Despite these differences, the challenges and barriers to countering online hate are similar to those in cyberbullying contexts [54]—both involve online users witnessing harmful or hateful behavior [21] and deciding whether to intervene [50]. Likewise, reasons that might deter an individual from defending a bullied peer—fear of exposure, retaliation, or the emotional toll—are similar to the hesitations one might feel when confronting online hate speech [128]. Hence, to develop a comprehensive and nuanced understanding of what drives or dissuades people from engaging in online counterspeech, our study synthesizes insights from prior research in bystander motivation in cyberbullying. Drawing on this body of work, we develop asset of survey variables to delve into the motivations and barriers potentially influencing online counterspeech engagement in the following section.

2.2.1 Motivations for Engaging in Online Counterspeech.

M1. Supporting Kin: Studies show that bystanders are more proactive in countering cyberbullying when they have close emotional or social ties with the victim [21, 50], with a similar trend seen in those countering online hate due to strong connections with friends and family [42].

M2. Supporting Others: The motivation to support others in general, as opposed to specific groups or individuals, can be traced back to theories of social responsibility and collective efficacy [145]. Collective efficacy refers to the belief that one's actions can contribute to the greater good, influencing community outcomes [69]. Studies have shown that people are more likely to engage in prosocial behavior when they perceive a moral obligation toward a broader community [45]. In a study examining prosocial online behavior, researchers demonstrate that collective efficacy drives individuals to engage more frequently in altruistic

[1]While a single incident of online hate speech can result in repeated victimization of targets as utterances can have widespread reach on digital platforms, cyberbullying is generally defined to require sustained, long-term exposure on the victim [161, 167].

activities [182]. Similarly, those who report feeling close to an online community are more likely to defend someone being targeted by online harassment [42, 43]. Such findings imply that individuals may engage in counterspeech not just to protect themselves or their kin, but for others as well.

M3. Supporting Self: Engaging in counterspeech can be a deeply personal act [141], especially when individuals feel directly targeted or harmed. However, motivations for self-defense are often nuanced. Guo and Johnson's study shows that users often underestimate the impact of online hate on themselves compared to its impact on others [71]. This perception could potentially influence users' motivation to engage in counterspeech for self-defense, as they may unknowingly downplay the harm directed towards them. Research also shows that personal experiences of online harm or targeted attacks also influence individuals' decisions to counter online hate [162]. Considering these complexities, we include "Supporting Self" as a variable for understanding motivations for counterspeech as it aims to capture the reason that might influence individuals to standup for themselves against online hate speech.

M4. Confronting Hate: The urgency to confront hateful or harmful behavior plays a critical role in motivating bystanders to intervene, both in the contexts of online hate speech [65] and cyberbullying [21]. For instance, bystanders are more likely to intervene when the bullying behavior is perceived as more severe [4, 53]. Similarly, the likelihood of bystanders challenging online harassment directly correlates with how menacing they perceive the harassment to be [103]. Hence, we include motivation to confront hateful behavior or people as a variable for engaging in online counterspeech.

M5. Educating Ignorance: Researchers have identified ignorance as one of the many factors contributing to the spread of online hate speech, as lack of awareness can lead people to adopt a narrow-minded view of others in society [33]. Hence, various non-profit and educational organizations [22] have advocated education as a strategy to counter online hate over banning users or online censorship [37, 168]. In line with this, Buerger et al. found that counter-speakers are often motivated to educate perpetrators of online hate as to why their message is harmful [24].

M6. Signaling Inclusion: Willingness to engage in counterspeech can often be influenced by a desire to signal inclusion, particularly within online communities [65]. Empathy emerges as a key factor in this context: research shows that individuals with higher levels of empathy are not only more inclined to stand up against online hate speech to protect the victim [145, 165], but also to signal a sense of inclusion and community cohesion [79, 108, 165].

M7. Issue Focus: Research shows that bystanders are more likely to intervene when the subject matter directly concerns social groups or issues that are important to them [127]. For instance, studies have found that bystanders were more willing to confront misogynist hate speech as compared to homophobic hate speech [127]. This suggests that motivation to engage in counterspeech may depend on issues or topics users are particularly passionate about.

Table 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

M8. Venting Emotions: Studies show that exposure to online incivility often trigger emotion-focused coping strategies, such as venting [111, 143]. Emotional responses, such as anger or frustration, may provoke individuals to "blow off steam" by countering the hate speech they encounter. Carlo et al. further supports this notion, indicating that emotional instability positively correlates with the adoption of emotion-focused coping strategies, which can manifest as aggressive counter responses [25]. Given these findings, we consider "Venting Emotions" as a potential motivation variable for engaging 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 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 ranging from 1 (None at all) to 5 (A great deal).

2.2.2 Barriers to Writing Counterspeech on Social Media.

B1. Fear of Public Exposure: Fear of public exposure can play a crucial role in bystander inaction, particularly when intervening would mean revealing oneself to a larger online audience [21]. Studies on online harassment have shown that the larger the audience size, the less inclined bystanders are to take action [21, 77, 103, 109, 125]. Similarly, the public nature of online platforms may deter individuals from engaging in online counterspeech due to fear of public exposure.

B2. Fear of Perpetrator Retaliation: Similarly, fear of retaliation from a harasser can significantly influence bystander motivations [13]. A study by Balakrishnan in 2018 found that 40% of bystanders chose not to intervene in instances of cyberbullying due to fears of retaliation [12].

B3. Fear of Third-Party Harassment: In addition to retaliation from the perpetrator, users also frequently express concerns about harassment from third parties [89, 145], as confronting online hate can also influence the likelihood of becoming a target of hate speech from others [42]. This phenomenon is supported by Ernst et al.'s 2017 study, which revealed that counterspeech in YouTube comments often attracted additional hateful remarks from third parties [60].

B4. Time Concern: Perceived time investment can deter users from engaging in online conflicts [16, 59]. This is reflected in commonly expressed views like "arguing on Facebook is a waste of time" [157]. Hence, it is plausible that concerns about time commitment could discourage users from engaging in online counterspeech, even if they are otherwise inclined to do so.

B5. Emotional Burden: While positive emotions like empathy have been found to increase bystander intervention in cyberbullying [63, 124], negative emotional burden could deter such actions [155]. Buerger's work on online activists highlights the emotional toll that counter-speaking can exert, especially when users voluntarily undertake these activities [23]. As a result, for some individuals the emotional cost of online counterspeech can outweigh the perceived benefits, leading to inaction.

B6. Skill Gap: According to social cognitive theory, self-efficacy plays a crucial role in bystander decisions [4, 53, 176]. In cyberbullying, bystanders are more likely to intervene when they feel capable and have the necessary resources to help [36]. However, bystanders are less likely to act when they think that other bystanders are more competent than themselves [98]. Likewise, users who feel less equipped to write an effective counterspeech might feel more reluctant to do so.

B7. Engagement Unqualified: Freis et al. and Rudnicki et al., highlight that bystanders may refrain from intervening if they feel that it is not their place to interject [63, 145]. Bystanders are also less likely to act in ambiguous situations [98]. Similarly, Piliavin et al.'s work shows that bystanders who witness only the aftermath of a harassment are less likely to intervene than those who see the entire situation unfold, potentially due to feeling less qualified to intervene due to a lack of contextual awareness [135].

B8. Engagement Reluctance: A general reluctance to engage in social media discourse can extend to counterspeech, with some individuals more hesitant to enter challenging or confrontational online conversations [48, 139].

B9. Engagement Ineffective: Wong et al.'s work shows that the perceived effectiveness of an intervention significantly influences bystanders' willingness to intervene in cases of online harassment [4, 53, 176]. While online counterspeech has been demonstrated to offer support to victims and encourage further counter-responses [55, 64, 150], there remains skepticism about its ability to genuinely alter the perpetrator's attitudes

or behaviors [11, 116, 150]. Hence, the perception that one's counterspeech may be ineffective could deter one from engaging in such activities.

Barrier Variables (B1-B9): Similar to the motivation variables, we presented the barrier variables to survey participants as statements B1-B9 (Table 2), and asked them to answer the following question: "*How much do the following factors deter you from writing a counterspeech on social media?*" - with response options ranging from 1 (None at all) to 5 (A great deal).

Table 2: Barrier Variables and Questionnaire Items

No	Barrier Variables	Questionnaire Items
B1	Fear of Public Exposure	I fear being publicly exposed
B2	Fear of Perpetrator Retaliation	I'm afraid of retaliation from the perpetrator
B3	Fear of Third-Party Harassment	I'm afraid that I will be harassed by people (other than the perpetrator)
B4	Time Concern	I don't want to spend time on this
B5	Emotional Burden	Writing a counterspeech is emotionally burdensome
B6	Skill Gap	I don't know how to write an effective counterspeech
B7	Engagement Unqualified	I feel that it's not my place to engage in counterspeech
B8	Engagement Reluctance	I don't like to engage in social media conversations
B9	Engagement Ineffective	I feel that my counterspeech would not make a difference

2.3 The Role of Artificial Intelligence (AI) in Online Counterspeech Engagement

The role of AI, particularly LLMs, in counterspeech research has traditionally focused on detecting hateful speech [34, 65, 81], and generating [101, 137, 147, 166], or evaluating [51, 62, 91, 183] counterspeech. LLMs excel at processing vast amounts of data quickly, alleviating the emotional toll on moderators and users who would otherwise have to detect, respond to, or report hateful online content manually [40, 132]. However, these models are nonetheless limited in their capacity to discern subtle nuances [17, 122] or cultural contexts [88] in hate speech, or distinguish between implicit and explicit forms of hate speech [97, 133]. These constraints often lead to detection failures, resulting in false positives or negatives [72], thus calling the need for better human-AI collaboration in mitigating online hate [96, 97].

More recently, researchers have focused on using LLMs to generate human-like counter-responses to hate speech, using various metrics for measuring the quality of AI-generated counterspeech, such as informativeness [34], politeness [147], and grammatical diversity [183]. However, the complex nature of hate speech, including subjective perceptions of hate [112] highlights the need for human-AI collaboration, not only in detecting hate speech, but also responding to it [42]. Currently, there is a notable lack of research on how people perceive the role of AI assistance with online counterspeech writing. Understanding how users feel about their own counterspeech or what aspects of writing a counterspeech are difficult for them, can better inform the design of AI tools for such purposes. Our research fills this gap by exploring how people's motivations and barriers influence their willingness to use AI assistance when writing online counterspeech. By doing so, our work aims to contribute to a more comprehensive understanding of AI's potential role in online counterspeech engagement, and by large, in combating online hate.

3 METHODS

We conducted a pre-registered survey[2] (N = 458) across English-speaking participants in the U.S. to examine key motivations and barriers that underlie why people do or do not engage in online counterspeech, their reported frequency of writing counterspeech on social media, and their willingness to use AI to help them write

[2] The survey was preregistered on Open Science Framework (OSF): https://osf.io/rzmg3/?view_only=6b2fd3a3d42b4b25a37f014612fac18a

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 counterspeech in response to each of the three hate posts.

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 [117], the Multi-Target Counter Narrative Dataset [61], and the Multilingual and Multi-Aspect Hate Speech Analysis (MLMA) collection [130]. We randomly selected hateful posts across five commonly occurring topics from the 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 instances of hate speech posts to ensure topical relevance. 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) three hateful posts and questions pertaining to them, (d) questions about past online hate speech experience, frequency of writing counterspeech online, and motivations as well as barriers to writing online counterspeech, (e) questions about prior use of ChatGPT, perceived usefulness of ChatGPT, as well as willingness of using such AI tools to aid in counterspeech writing, and finally (f) demographic and social media use questions. We illustrate the survey flow in Figure 1. All survey questions are presented in the appendix.

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. Participants then completed an eligibility check, requiring them to be over 18 years old, use English, reside in the U.S., and have social media experience. Then, participants were provided with definitions of hateful speech, counterspeech, as well as examples of effective counterspeech. The attention check included questions that assessed direct recall of information explicitly stated in the background materials about hate speech and counterspeech. These questions required no interpretation or inference; participants who read the materials would be able to identify the correct answers based solely on the content presented, thus ensuring they had carefully read the introductory materials.

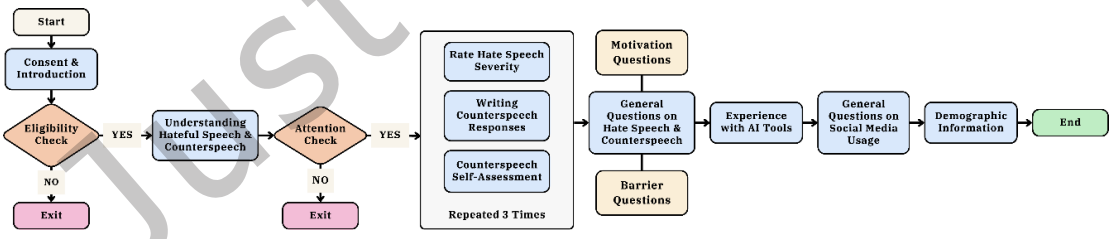

Figure 1: Flowchart of the Survey Process

Following this, participants were shown three unique hateful posts randomly selected from the set of 900 hate posts described in 3.1. The random selection occurred individually for each participant within the survey, so that each participant saw a different set of three posts. 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 also asked to rate the hatefulness of each post using a four-point scale, with the question, "*How hateful do you find this post?*" Response options ranged from (1) A little to (4) A great deal.

Participants were then prompted to respond 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 **satisfaction**, perceived **difficulty**, and perceived **effectiveness** of each

counterspeech they wrote using a five-point Likert scale. While prior research does not provide specific measures for these variables, given their significance in relation to our study, we developed corresponding response measures. Specifically, we asked, "*How satisfied are you with the counterspeech that you've written?*" (satisfaction), "*How difficult was it to write this counterspeech?*", and "*How effective do you think your counterspeech would be in preventing the perpetrator from engaging in further hateful behavior?*" with response options ranging from 1- 5 (extremely dissatisfied-extremely satisfied; extremely difficulty-extremely easy; not effective at all-extremely effective).

Finally, participants answered questions related to motivations and barriers to writing online counterspeech, frequency of writing online counterspeech, and willingness to use ChatGPT to write counterspeech on social media. We did not ask users to use AI in writing their counterspeech in the survey. However, our study serves as an initial examination of whether users are willing to use AI for counterspeech, and the factors that shape this willingness. Table 3 lists all variables included in our survey. We asked participants' opinions in the rest of the survey using the conventional 5-point Likert scale, a standard in social science research [44], except for binary response questions (such as prior usage of ChatGPT) and demographic inquiries.

Table 3: Survey Variables

Independent Variables	**Control Variables**	**Dependent Variables**
Barriers	**C1. Demographics**	**RQ1**
B1: Fear of public exposure	Age	Frequency of writing online counterspeech
B2: Fear of perpetrator retaliation	Gender	
B3: Fear of third-party harassment	Ethnicity	**RQ2**
B4: Time concern	Education level	Satisfaction
B5: Emotional burden	Sexual orientation	Difficulty
B6: Skill gap	Political View	Effectiveness
B7: Engagement unqualified	**C2. Social Media Behavior & Experience**	**RQ3**
B8: Engagement reluctance	Social media commenting frequency	Willingness to use ChatGPT to write counterspeech
B9: Engagement ineffective	Use of real name on social media	
Motivations	Prior experience of online hate speech target	
M1: Supporting kin	Frequency of encountering online hate speech	
M2: Supporting public	**C3. Prior Use & Perception of ChatGPT**	
M3: Supporting self	Prior use of ChatGPT	
M4: Confronting hate	Perceived usefulness of ChatGPT	
M5: Educating ignorance		
M6: Signaling inclusion		
M7: Issue focus		
M8: Venting emotions		

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. The average survey completion time was 15 minutes with a compensation rate of $12/hour. Our initial target sample size was 580 participants, aiming to achieve a Goodness-of-Fit Index (GFI) of 0.99, equivalent to a Root Mean Square Error of Approximation (RMSEA) of 0.02, with at least 95% power at the standard 0.05 alpha error probability. Although an RMSEA below 0.08 is generally considered acceptable [86, 156], we selected this stringent RMSEA target of 0.02 to ensure robustness.

We conducted an interim analysis after collecting 450 responses. After excluding those who failed attention checks or failed to complete the survey, we had 376 valid responses, which achieved the minimum required RMSEA (<0.08). To ensure we maintained a better level of model fit with a slightly larger sample, we continued data collection to 536 responses. Of the initial 536 respondents, we excluded those who failed attention checks (n=72) or failed to complete the survey (n=6), resulting in a final sample of 458 participants. To confirm the adequacy of this sample size, we conducted a post hoc power analysis. With our final sample size of 458 and an actual RMSEA value of 0.06, the analysis revealed a power greater than 0.99, indicating enough statistical power to detect the effects of interest in our study.

3.4 Analysis

RQ1 What motivations and barriers influence the frequency of people's engagement in online counterspeech?

To address RQ1, we performed an ordinal logistic regression model 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 in response to the question "*How often do you write counterspeech online?*" The independent variables were the 9 barrier and 8 motivation variables, prior experience being a target of online hate speech, 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 [129]. Due to significant interaction effects, we conducted a subgroup analysis to further explore differences in motivations and barriers between targeted and non-targeted individuals. This involved running two separate ordinal logistic regression models: one for participants who had been targets of online hate speech and another for those who had not.

RQ2 How do demographic variables shape people's motivations and barriers in online counterspeech engagement?

To answer RQ2, we conducted a structural equation model (SEM) to investigate the effects of key demographic factors and social media behavior on peoples' motivations and barriers in engaging in online counterspeech [95]. We performed an exploratory factor analysis (EFA) using principal axis factoring with oblimin rotation to group the items related to the barriers and motivators into latent variables [100]. The detailed steps for determining the number of factors and selecting the final factors are provided in the Appendix. This analysis resulted in a five-factor structure that was subsequently used in our structural equation model (SEM). We performed the SEM using diagonally weighted least squares (DWLS) to estimate the path coefficients, as our data consisted of ordinal categorical variables (Likert scale) [144]. We assessed the model fit using chi-square, comparative fit index (CFI), RMSEA, and standardized root mean square residual (SRMR) [86, 156]. We report the standardized coefficients, standard errors, p-values, and R-squared for each endogenous variable in the model.

RQ3 How do people's motivations and barriers in online counterspeech engagement influence: (a) how satisfied they are with their counterspeech, (b) how difficult they find it to write counterspeech, and (c) how they perceive the effectiveness of their counterspeech?

In our study, we adapted the SEM to include both two-item and one-item variables for addressing RQ2 and RQ3, because this approach allows us to model more latent variables, enhancing theoretical sophistication and statistical control [80]. Following Bollen's (1989) guidelines [19] and Daniel's (2021) recommendations [110], our model employed DWLS estimators, ensuring methodological soundness for two-item variables.

Additionally, following the approach by Hayduk et al. (2012) [80], we used a one-item variable, "time", to capture its unique influence on engagement in counterspeech. The use of one-item and two-item variables for latent constructs is not uncommon in the field, as demonstrated by Wright [179, 180], Blalock [93], Duncan [56], and Heise [82]. This methodological choice aimed to account for the complexity of counterspeech engagement while maintaining model parsimony.

To address RQ3, we used the same SEM approach as in RQ2, but included peoples' perceived satisfaction, difficulty, and effectiveness towards their self-written counterspeech as key dependent variables. We chose to use SEM because it allows us to consider multiple dependent variables simultaneously. We examined how these dependent variables were related to the barrier and motivation variables. We followed the same approach for model assessment and reporting as in RQ2. For each participant, the results for peoples' perceived satisfaction, difficulty, and effectiveness were averaged across the three writing tasks to provide a comprehensive measure of their overall experience with counterspeech writing.

RQ4 How are people's motivations and barriers in online counterspeech engagement associated with their willingness to use AI assistance?

To address RQ4, we conducted an ordinal logistic regression model to examine the relationship between peoples' barriers and motivations for engaging in counterspeech on social media and their willingness to use AI technology, such as ChatGPT, for this purpose. Similar to RQ1, the independent variables consisted of the nine barriers and eight motivation items, while demographic factors as well as social media use and experience variables were used as control variables. The dependent variable was captured via a five-point Likert scale in response to the question, "*If you were writing counterspeech on social media, would you use artificial intelligence technology like ChatGPT to assist you?"* Tests for interaction effects showed significant relationships between prior ChatGPT experience and several predictor variables. To understand how counterspeech motivations/barriers relate to willingness to adopt AI assistance, and given that some participants had no prior experience with ChatGPT, we conducted two subgroup analyses: one for participants with prior experience using ChatGPT (N=296) and another for those without (N=162). Here we treatd prior ChatGPT experience as the primary distinguishing factor, while target status was included as a control variable. For those who have used ChatGPT before, an additional control variable for perceived usefulness was included.

Qualitative analysis of open responses: Finally, to answer RQ4 with more depth, we analyzed the participants' open-ended responses to the following, "*Why would or wouldn't you use artificial intelligence technology like ChatGPT to help you write a counterspeech?"* Using an inductive open-coding approach [163] we first coded the open responses, allowing codes and themes to emerge from the data. We then conducted axial coding to organize and refine the codes and themes to understand how they connected to each other [163]. We then used memoing to make sense of the emerging codes and connections between codes and themes. Throughout this process, three authors discussed emerging themes and connections between codes and themes, and used Cohen's kappa to evaluate the inter-rater agreement for the codes across the authors [171].

4 RESULTS

A total of 458 participants (50.6% woman, mean age: 40.3±13.3) completed the survey. More demographic details are in Table A6, Appendix. Skewness and kurtosis values for study variables are presented in Table A7, Appendix. On average, the length of counter speech authored by the participants was 41.6 words.

4.1 What motivations and barriers influence the frequency of people's engagement in online counterspeech (RQ1)?

Figure 2 presents the motivations, barriers, and frequency of writing counterspeech on social media based on the survey. Panel A shows that supporting kin received the highest percentage, with 61% of respondents rating it as at least "a lot." Issue focus and supporting self followed closely, each with at least 50% of respondents choosing "a lot" or higher. In contrast, signaling inclusion had the lowest percentage, with only 31% of respondents considering it to be at least "a lot." Panel B indicates that engagement ineffectiveness was the most commonly reported barrier, with 45% of respondents rating it as at least "a lot." Time concern was the second most reported barrier, with 34% of respondents selecting at least "a lot," while skill gap had the lowest endorsement, with only 14% of respondents considering it to be at least "a lot." Panel C illustrates the frequency

of writing counterspeech on social media: 3% of respondents reported doing so "frequently," 7% "often," 30% "sometimes," 33% "rarely," and 27% "never." This distribution shows that at least 60% of participants rarely or never engage in counterspeech, only 10% reported doing so frequently or often.

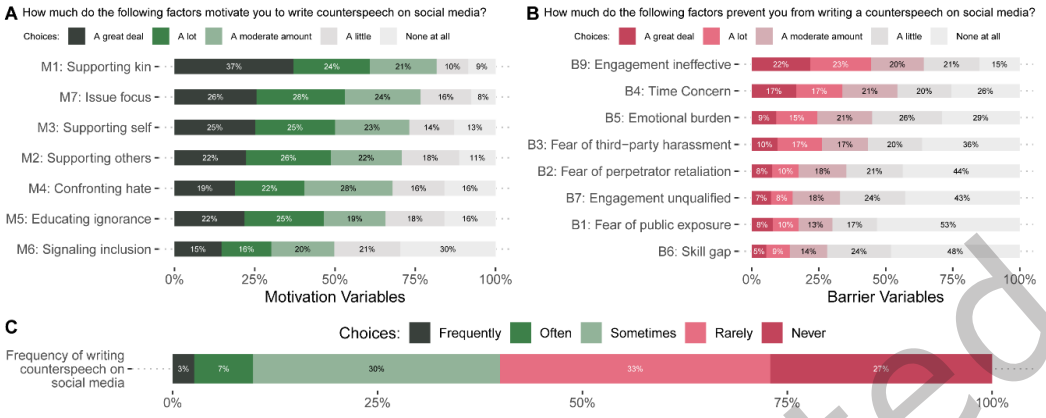

Figure 2: Motivations, Barriers, and Frequency of Writing Counterspeech on Social Media. Panels A and B show motivating and preventive factors for writing counterspeech on social media, respectively, sorted by the cumulative percentage of responses from "A great deal" to "A moderate amount". Panel C displays the frequency of engaging in counterspeech on social media.

Our ordinal logistic regression results from the full sample analysis (Table A8, Appendix) show that having been a target of online hate speech emerged as a strong predictor that drives people to frequently engage in online counter speech ($\beta = 1.190$, $OR = 3.288$, $p < .001$). Individuals who have been a victim of online hate speech in the past engaged in counterspeech on social media significantly more often than those without such experience. Given the significant interaction effects between target status and other predictors (Table A9, Appendix), and that people who have been targets of online hate speech (targets) in the past tend to engage in counterspeech on social media more frequently than those without such experience (non-targets), we conducted a subgroup analysis to understand how counterspeech motivations and barriers differed between the two groups. We conducted two ordinal logistic regression models: non-target group (N=276) and target group (N=182). Results from the subgroup analysis are shown in Table 4. The McFadden R^2 values were 0.422 for the non-target group and 0.414 for the target group. All VIF values were below 2.00.

Targets of Online Hate Speech: Targets engage in counterspeech more often if they have stronger desires to confront hateful persons or behavior ($\beta = 3.238$, $OR = 25.490$, $p = .004$), or to emotionally vent ($\beta = 1.485$, $OR = 4.416$, $p = .011$). However, the stronger the perception that counterspeech is ineffective in general ($\beta = -1.253$, $OR = 0.286$, $p = .043$), the less often do targets engage in counterspeech. Additionally, among targets, political views showed a significant positive association with counterspeech frequency ($\beta = 2.232$, $OR = 9.319$, $p = .010$).

Non-Targets of Online Hate Speech: By contrast, non-targets engage in counterspeech more often when they have a strong desire to signal inclusion to others through counterspeech ($\beta = 1.204$, $OR = 3.334$, $p = .038$). Furthermore, non-targets are less likely to participate in counterspeech if they generally prefer to avoid social media conversations ($\beta = -2.879$, $OR = 0.056$, $p < .001$) or if they perceive a skill gap ($\beta = -1.544$, $OR = 0.214$, $p = .045$), which is also insignificant for targets. Non-targets also engage in more counterspeech if they frequently encounter online hate ($\beta = 1.882$, $OR = 6.567$, $p = .010$), which is insignificant for targets.

In summary, the RQ1 results demonstrate that people who have been targets of online hate speech (targets) tend to engage in counterspeech on social media more frequently than those who do not have such experience (non-targets). This higher engagement stems from a desire to emotionally vent and to confront hateful persons or behaviors. On the other hand, for those who have never been a target of online hate speech, signaling inclusion is one of the main drivers for engaging in online counterspeech. Regarding barriers, targets are less likely to engage in counterspeech when they perceive it as ineffective, while non-targets' engagement is

primarily hindered by their general reluctance to participate in social media conversations and perceived skill gaps.

Table 4: Factors Affecting Counterspeech Writing Frequency in Targets and Non-Targets of Online Hate Speech

		Non-target group (n=276)			Target group (n=182)		
		β	**OR**	**P**	**β**	**OR**	**P**
Motivations	M1: Supporting kin	0.873	2.393	0.193	0.497	1.643	0.577
	M2: Supporting others	1.368	3.929	0.104	0.438	1.549	0.131
	M3: Supporting self	0.752	2.120	0.198	0.371	1.450	0.606
	M4: Confronting hate	0.393	1.481	0.553	**3.238**	**25.490**	**0.004****
	M5: Educating ignorance	0.359	1.431	0.564	-1.686	0.185	0.074
	M6: Signaling inclusion	**1.204**	**3.334**	**0.038***	-0.190	0.827	0.327
	M7: Issue focus	-0.235	0.790	0.759	1.424	4.153	0.229
	M8: Venting emotions	0.192	1.211	0.760	**1.485**	**4.416**	**0.011***
Barriers	B1: Fear of public exposure	-0.458	0.632	0.469	0.879	2.409	0.254
	B2: Fear of perpetrator retaliation	0.036	1.037	0.966	-0.593	0.553	0.477
	B3: Fear of third-party harassment	0.634	1.885	0.454	-0.263	0.769	0.729
	B4: Time Concern	-0.635	0.530	0.211	-0.711	0.491	0.259
	B5: Emotional burden	-0.550	0.577	0.391	0.225	1.253	0.755
	B6: Skill gap	**-1.544**	**0.214**	**0.045***	-0.181	0.834	0.378
	B7: Engagement unqualified	-0.949	0.387	0.233	0.270	1.310	0.728
	B8: Engagement reluctance	**-2.879**	**0.056**	**0.000*****	-1.098	0.334	0.070
	B9: Engagement ineffective	0.649	1.914	0.204	**-1.253**	**0.286**	**0.043***
SNS	**Frequency of encountering online hate speech**	**1.882**	**6.567**	**0.010***	0.802	2.230	0.381
	Social media commenting frequency	0.996	2.707	0.173	1.454	4.280	0.074
	Use of real name on social media	0.544	1.723	0.133	0.198	1.219	0.104
Demographic	Age	-0.003	0.997	0.803	0.040	1.041	0.077
	Gender	0.193	1.213	0.601	0.487	1.628	0.289
	Ethnicity	-0.313	0.731	0.362	-0.058	0.944	0.891
	Education level	-0.113	0.893	0.423	0.025	1.025	0.882
	Sexual orientation	-0.634	0.531	0.238	-0.027	0.974	0.958
	Political views	-0.289	0.749	0.551	**2.232**	**9.319**	**0.010***

β = regression coefficient; OR = odds ratio (exp(β)); P = p-value. *p < .05, **p < .01, ***p < .001.

4.2 How do demographic variables shape people's motivations and barriers in online counterspeech engagement (RQ2)?

To answer RQ2, we conducted a structural equation model (SEM) to investigate the effects of key demographic factors and social media behavior on people's motivations and barriers in engaging in online counterspeech. Prior to constructing our model, we examined the underlying structure of the motivations and barriers influencing counterspeech engagement using EFA [95]. The detailed steps are provided in the Appendix. These analyses helped us categorize the individual motivation and barrier variables into broader latent variables (LV) as shown in Table 5.

The EFA results revealed a five-factor structure that characterizes the motivations and barriers related to counterspeech. We applied a robust cut-off threshold of 0.55 for the factor loadings, following [100], and included variable items with strong correlations to the latent variables to ensure that our model was parsimonious and reliable [100], eliminating B9 and M8. This process resulted in five latent variables associated with counterspeech engagement: **Fear-Driven Inhibition (LV1)**, **Time Concern (LV2)**, **Emotional and Skill Barriers (LV3)**, **Engagement Hesitation (LV4)**, and **Motivation (LV5)** – for a description of these latent variables, see Appendix.

Table 5: Factor Loadings for Barriers and Motivations Associated with Writing Counterspeech on Social Media

	Motivation and Barrier Items	LV1	LV2	LV3	LV4	LV5
		Fear-Driven Inhibition	Time Concern	Emotional & Skill Barrier	Engagement Hesitation	Motivation
Barriers	B1: Fear of public exposure	**0.694**				
	B2: Fear of perpetrator retaliation	**0.882**				
	B3: Fear of third-party harassment	**0.848**				
	B4: Time Concern		**1.000**			
	B5: Emotional burden			**0.653**		
	B6: Skill gap			**0.565**		
	B7: Engagement unqualified				**0.641**	
	B8: Engagement reluctance				**0.593**	
	B9: Engagement ineffective				0.538	
Motivations	M1: Supporting kin					**0.729**
	M2: Supporting others					**0.842**
	M3: Supporting self					**0.673**
	M4: Confronting hate					**0.815**
	M5: Educating ignorance					**0.793**
	M6: Signaling inclusion					**0.716**
	M7: Issue focus					**0.822**
	M8: Venting emotions					0.372

Note. Factor loadings are standardized. Shaded cells indicate items retained in the final model. LV = Latent Variable.

4.2.1 *Demographic Characteristics Significantly Influence Motivations and Barriers for Engaging in Online Counterspeech*

Our SEM analyses examining the effects of key demographic factors on each latent variable (LV1 – LV5) underlying counterspeech motivations and barriers indicated a strong fit. The chi-square test statistic was 501.092 with 189 degrees of freedom and p < .001. The CFI was 0.915, indicating a sufficient fit [86, 156]. The RMSEA was 0.060, which was within the acceptable range of 0.05 to 0.08 [86, 156]. See Table A14, Appendix for complete results.

Age: Younger participants experienced higher fear-driven inhibition (β = *-0.162, p = .001*), suggesting that they were more deterred from engaging in counterspeech on social media due to fears of public exposure, perpetrator retaliation, and third-party harassment, compared to their older counterparts. **Gender:** Similarly, women, compared to men also reported higher fear-driven inhibition related to potential retaliation from perpetrators, public exposure, and third-party harassment (β = *0.228, p < .001*). Furthermore, women also reported higher emotional and skill-related barriers - namely the emotional toll of engaging in online counterspeech and uncertainties on how to construct effective counter responses (β = *0.298, p < .001*). **Education:** Similarly, participants with higher education levels also reported higher levels of fear-driven inhibition (β = *0.194, p < .000*) and higher emotional burden & skill barriers (β = *0.181, p = .003*) than participants with lower education backgrounds. **Political views:** Participants who were more politically liberal were more motivated to write online counterspeech (β = *0.150, p = .001*) and reported less engagement hesitation (β = *-0.153, p = .010*) compared to their conservative counterparts.

To summarize, in RQ2, we show that motivations and barriers influencing counterspeech engagement can be grouped into five latent variables - Fear, Time Concern, Emotional & Skill Barrier, Engagement Hesitation, and Motivation. People's demographic backgrounds and social media experiences impact their willingness to engage in online counterspeech across these five latent variables. Specifically, women, highly educated users, and those frequently encountering online hate are less likely to engage due to fear and emotional/skill barriers, while older, more liberal users, and those who have been targeted by online hate exhibit greater overall motivation for counterspeech participation.

4.3 How do people's motivations and barriers in online counterspeech engagement influence: (a) how satisfied they are with their counterspeech, (b) how difficult they find it to write counterspeech, and (c) how they perceive the effectiveness of their counterspeech (RQ3)?

Figure 3 illustrates the results of our second SEM analysis evaluating the impact of the five latent variables (LV1-LV5) on the three central dependent outcomes: (a) participants' perceived **satisfaction** with their self-authored counterspeech, (b) perceived **difficulty** in writing the counterspeech, and (c) perceived **effectiveness** of their own counterspeech in mitigating the hate speech they were responding to. Detailed results are shown in Table A10 and A11, Appendix.

The resulting fit indices for our SEM model indicated a good fit. The chi-square statistic was 693.423 (p < .001, df = 264). The CFI was 0.900. The RMSEA was 0.060 (*90% CI = [0.054, 0.065]*), which was within the acceptable range of 0.05 to 0.08. The SRMR was 0.051, which was below the cut-off value of 0.08. In addition to controlling for the control variables in Table 3, we also controlled the model for how hateful the participants perceived the hate speech they were responding to. The results show that the perceived hatefulness of the speech had a significant positive effect on participants' satisfaction (β = *0.126, p = .004*) and perceived effectiveness (β = *0.132, p = .026*) of their counterspeech, but did not significantly impact the perceived difficulty (β = *-0.030, p = .615*) of writing counterspeech. Below we discuss our main results.

Fear-Driven Inhibition (LV1): People who have higher levels of fear of writing counterspeech due to perpetrator retaliation, public exposure, and third-party harassment were significantly more satisfied ($\beta = .536$, $p = .022$) with their self-written counterspeech and perceived their counterspeech to be more effective ($\beta = .812$, $p = .019$) than those who scored lower on the fear-driven inhibition latent variable. One plausible explanation for this is that, due to heightened concerns about negative consequences, users may invest additional effort in crafting their counterspeech [158]. This extra diligence could translate into increased satisfaction and a stronger belief in the effectiveness of their counterspeech [9].

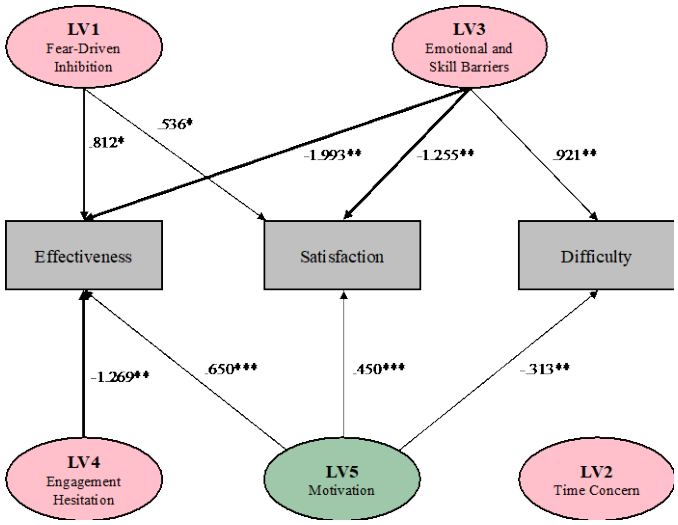

Figure 3: Standardized path coefficients of the structural equation model. We show how the barrier latent variables in red and the motivation latent variable in green affect peoples' perceived satisfaction, difficulty, and effectiveness of their counterspeech. *(* $p < .05$; ** $p < .01$; *** $p < .001$).*

Time Concern (LV2): Time Concern did not yield statistically significant effects on the outcome variables, suggesting that concerns about the time required for writing counterspeech did not notably influence participants' satisfaction, difficulty, or perceived effectiveness in this context.

Emotional and Skill Barriers (LV3): Those who scored higher on the Emotional and Skill Barriers latent factor were significantly less satisfied with their own counter speech ($\beta = -1.255$, $p = .002$), and experienced more difficulty in writing counterspeech ($\beta = .921$, $p = .007$) in response to the hateful speech they were shown in the survey. These individuals were also significantly less likely to perceive their self-written counterspeech as effective in deterring the hate speech they were responding to ($\beta = -1.993$, $p = .001$).

Engagement Hesitation (LV4): Surprisingly, individuals who scored higher on the Engagement Hesitation latent factor were significantly more likely to perceive their own counterspeech as effective ($\beta = 1.269$, $p = .008$). A plausible explanation for this seemingly paradoxical result could be that those who are hesitant to engage in counterspeech due to feeling unqualified to engage, or are reluctant to converse on social media in general, may set a higher threshold for action. In other words, they may only choose to engage when they believe they have something truly impactful to say. As a result, when they do overcome their hesitation and contribute counterspeech, it may be more thoughtfully crafted, and, thus, such individuals may perceive their counterspeech as more effective.

Motivation (LV5): Participants with stronger motivations for crafting counterspeech not only felt more satisfied with their counter responses ($\beta = .450$, $p < .001$), but also found the writing process less challenging ($\beta = -.313$, $p = .001$). Additionally, they were more confident in the effectiveness of the counterspeech they wrote in response to the hate speech shown in the survey ($\beta = .650$, $p < .001$).

In summary, RQ3 results show that motivations and barriers associated with counterspeech engagement significantly impact people's writing experience and perception of their own counterspeech. Individuals more apprehensive about writing counterspeech due to retaliation, public exposure, and third-party harassment were more likely to perceive their own writing as effective and satisfying. Conversely, people with greater emotional and skill-related barriers in writing counterspeech were less likely to perceive their counterspeech as effective and also found it more difficult to write counterspeech to the hateful posts shown in the survey. Surprisingly, individuals who scored higher on the Engagement Hesitation latent factor were significantly more likely to perceive their own counterspeech as effective.

4.4 How are people's motivations and barriers in online counterspeech engagement associated with their willingness to use AI assistance (RQ4)?

Figure 4 illustrates the willingness to use ChatGPT for writing counterspeech among participants with and without prior ChatGPT experience. Among those who have used ChatGPT (n = 162) before, 5% indicated they would "definitely" use it for counterspeech, 17% said they would "probably" use it, 26% were uncertain ("might or might not"). In contrast, among those who have never used ChatGPT (n = 296), 11% chose either "definitely" or "probably" use it.

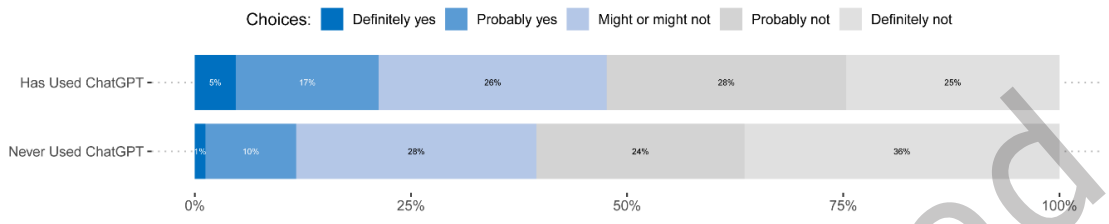

Figure 4: Willingness to Use ChatGPT for Counterspeech Writing Based on Prior Using Experience. It shows the distribution of willingness to use ChatGPT for writing counterspeech, comparing participants who have previously used ChatGPT (top bar) with those who have never used it (bottom bar).

Results from our ordinal logistic regression for RQ4 , based on the full sample analysis, demonstrate that people who have used ChatGPT in the past are significantly more willing to use AI assistance in writing online counterspeech, compared to those who have never used such tools before ($\beta = 1.031$, $OR = 2.805$, $p < .001$). For specific details, see Table A12 and Table A13, Appendix. Hence, similar to RQ1, we conducted a subgroup analysis between the two groups. Results from this subgroup analysis are shown in Table 6. The McFadden R^2 values were .329 for non-users of ChatGPT and .174 for prior users of ChatGPT. All VIF values were below 2.00.

People with prior experience using AI tools like ChatGPT: For those who have used ChatGPT in the past, feeling less capable of writing effective counterspeech (skill gap) is a significant motivator for using AI assistance ($\beta = 1.723$, $OR = 5.603$, $p < .001$). Interestingly, this group is also more willing to use ChatGPT when writing counterspeech to signal inclusion to others ($\beta = 0.834$, $OR = 2.302$, $p = .050$), but less inclined to use it when engaging in online counterspeech to stand up for themselves ($\beta = -0.952$, $OR = 0.386$, $p = .040$). Notably, the perceived usefulness of ChatGPT is a strong predictor of willingness to use AI assistance ($\beta = 3.576$, $OR = 35.737$, $p < .001$).

Table 6: Results of Subgroup Analyses for Willingness to Use AI Assistance for Writing Online Counterspeech

		Non Users of ChatGPT (n=162)			Prior Users of ChatGPT (n=296)		
		β	**OR**	**P**	**β**	**OR**	**P**
Motivations	**M1: Supporting kin**	**-3.161**	**0.042**	**0.004****	0.456	1.578	0.405
	M2: Supporting others	-0.406	0.667	0.209	0.235	1.265	0.695
	M3: Supporting self	0.884	2.419	0.272	**-0.952**	**0.386**	**0.040***
	M4: Confronting hate	**3.426**	**30.764**	**0.004****	-0.560	0.571	0.293
	M5: Educating ignorance	-1.291	0.275	0.191	-0.118	0.889	0.809
	M6: Signaling inclusion	-0.706	0.494	0.486	**0.834**	**2.302**	**0.050***
	M7: Issue focus	1.907	6.734	0.095	0.032	1.033	0.845
	M8: Venting emotions	0.384	1.467	0.674	-0.599	0.549	0.141

Barriers	**B1: Fear of public exposure**	**-3.528**	**0.029**	**0.004****	0.304	1.356	0.479
	B2: Fear of perpetrator retaliation	**2.600**	**13.464**	**0.032***	0.238	1.268	0.143
	B3: Fear of third-party harassment	0.732	2.078	0.477	-0.141	0.869	0.347
	B4: Time Concern	-1.074	0.342	0.228	-0.019	0.982	0.865
	B5: Emotional burden	0.562	1.754	0.569	-0.073	0.929	0.551
	B6: Skill gap	**3.572**	**35.584**	**0.001*****	**1.723**	**5.603**	**0.000*****
	B7: Engagement unqualified	-1.633	0.195	0.081	0.164	1.179	0.216
	B8: Engagement reluctance	0.769	2.158	0.309	-0.056	0.946	0.610
	B9: Engagement ineffective	-0.448	0.639	0.611	0.012	1.012	0.919
SNS & ChatGPT	**Past experience of online hate speech target**	**-1.476**	**0.229**	**0.035***	**-0.523**	**0.593**	**0.049***
	Frequency of encountering online hate speech	-2.085	0.124	0.069	-0.117	0.889	0.829
	Social media commenting frequency	-0.397	0.672	0.664	0.801	2.228	0.141
	Use of real name on social media	-0.651	0.522	0.313	-0.236	0.790	0.380
	Perceived usefulness of ChatGPT	/	/	/	**3.576**	**35.737**	**0.000*****
Demographic	**Age**	**0.051**	**1.053**	**0.024***	0.012	1.012	0.251
	Gender	-0.579	0.561	0.363	-0.025	0.976	0.929
	Ethnicity	0.755	2.127	0.170	0.186	1.205	0.502
	Education level	-0.261	0.770	0.232	0.169	1.184	0.097
	Sexual orientation	-0.473	0.623	0.578	0.093	1.097	0.774
	Political views	1.222	3.395	0.196	-0.839	0.432	0.056

β = regression coefficient; OR = odds ratio (exp(β)); P = p-value. *p < .05, **p < .01, ***p < .001.

People who have never used AI tools like ChatGPT: By contrast, among those who have never used ChatGPT, a stronger fear of perpetrator retaliation makes them significantly more willing to use AI for help when writing counterspeech on social media (*β = 2.600, OR = 13.464, p = .032*). Interestingly, non-users with a stronger motivation to confront hate are more willing to use AI assistance (*β = 3.426, OR = 30.764, p = .004*). The skill gap is also a significant factor for non-users (*β = 3.572, OR = 35.584, p = .001*). However, non-users are less willing to rely on AI assistance when the purpose of engaging in counterspeech is to defend close kin – family and close friends (*β = -3.161, OR = 0.042, p = .004*). Also, fear of public exposure significantly decreases willingness to use AI (*β = -3.528, OR = 0.029, p = .004*).

Both prior users and non-users are less willing to use AI for counterspeech writing if they have been targets of online hate speech in the past (prior users: *β = -0.523, OR = 0.593, p = .049;* non-users: *β = -1.476, OR = 0.229, p = .035*). Furthermore, age shows a positive effect only among non-users (*β = 0.051, OR = 1.053, p = .024*).

4.4.1 Qualitative Analysis: Motivations and Reservations for Using AI for Counterspeech Writing

Our qualitative analysis of participants' open responses resulted in a total of six themes associated with why people would or would not use AI assistance for writing online counterspeech. Tables 7 and 8 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. This indicates a very good level of agreement among the raters. Because some user statements

contained more than one theme, we coded them into multiple categories. Therefore, 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 faster for them. Some participants also highlighted that AI could help them more easily come up with ideas that they can elaborate on themselves and provide them with useful strategies for writing effective counterspeech that they can use later on.

Less Emotional Burden: Many participants highlight that AI tools like ChatGPT could help alleviate many of the negative emotions, such as anger and frustration, that arise when writing counterspeech. For example, one participant notes: "*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 quickly provide supporting evidence for arguments. For example, participants state that AI tools like ChatGPT can help them "*find the right words and vocabulary to express [their] thoughts more clearly and eloquently,*" as well as "*provide data and facts to make [their] argument stronger*".

Authenticity and Ethical Concerns: The most common reservations for using AI assistance in writing counterspeech are authenticity and ethical concerns. Some voiced feelings of cheating and unease, 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 AI usage detaches their words from personal ownership, 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 personal 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 7: 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 & 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.*	2.2%

	• *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.*	

Table 8: 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.* • *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%

In summary, RQ4 results demonstrate that prior experience of using AI tools like ChatGPT significantly influences people's willingness to use such tools to write online counterspeech. People who do versus do not have experience using AI tools also differ in terms of specific motivations and barriers that underly why they would or would not use AI for counterspeech writing. Prior users are more willing to use AI to signal inclusion to others through counterspeech, but not to defend themselves. In contrast, non-users (those who have never used ChatGPT or similar tools) are more willing to use AI to help them write counterspeech when they fear retaliation from the perpetrator or when motivated to confront hate, but they are significantly less willing when they fear public exposure and less inclined to rely on AI to defend friends and family. Notably, a perceived skill gap in writing effective counterspeech emerged as a significant factor for both groups, though with a stronger

effect among non-users. Interestingly, both groups show decreased willingness to use AI if they have been targets of online hate speech in the past.

5 DISCUSSION

5.1 Motivations and Barriers in Online Counterspeech

5.1.1 Main Deterrent of Online Counterspeech Engagement: General Reluctance to Engage on Social Media

RQ1 results show that over 60% of respondents rarely or never write counterspeech. The full sample analysis further reveals that peoples' general reluctance to engage on social media tends to outweigh other motivations for engaging in online counterspeech. This aligns with prior research in HCI that identifies reluctance towards social media interaction as a contributing factor to online bystander effect [50, 155, 165]. For this reason, while a majority of users observe online harassment, less than a third choose to intervene [8]. Nevertheless, our findings go a step further by examining key emotional and psychological factors that influence this behavior.

5.1.2 Main Driver of Online Counterspeech Engagement: Prior Victimization

Studies in online bystander intervention show that prior victimization is a key predictor of bystander action [145]. RQ1 results confirm these insights, demonstrating that having been a target of online hate speech in the past is a strong predictor that drives people to frequently engage in online counterspeech. We provide further nuance to prior scholarship by showing how counterspeech motivations and barriers in fact, significantly vary between former targets and non-targets. For targets, the perceived ineffectiveness of their counterspeech is the most significant barrier to engaging in counterspeech, while non-targets are primarily deterred by their reluctance to engage on social media in general. These findings are supported by existing work that suggests victims know well when an intervention may or may not be effective due to the memory of their own experiences as well as bystander effects on social media [83, 145]. With respect to peoples' motivations, Costello et al. (2016) found that past victims of online hate are more than three times as likely to defend fellow victims [43]. Our results provide context to this research, showing that targets engage in online counterspeech more often when they are primarily motivated to confront a hateful person or their behavior, while this motivation is insignificant for non-targets.

5.1.3 Multi-Item Survey Scale for Measuring Online Counterspeech Motivations and Barriers

While prior research in cyberbullying has created numerous scales to measure bystander motivations [157, 165], most are not generalizable to context of online counterspeech due to differences highlighted in prior research [145]. Furthermore, such studies are often based on children and adolescents [12, 13, 128], 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 online counterspeech motivation and barriers. Furthermore, as shown in RQ2, each motivation and barrier scale can be constructed into five key latent variables (LV1 – LV5). Researchers can use these variable items both individually and as latent factors in future studies. The choice between using individual items or latent factors depends on the research objectives. Individual items are suitable when examining specific motivations or barriers in relation to general perceptions, behaviors, or attitudes. In contrast, latent factors are more appropriate when studies aim to capture the overall effects of motivations and barriers.

5.1.4 Demographic Variances in Counterspeech Motivations and Barriers

Using our latent variables (LV1 -LV5), we demonstrate key demographic variances in counterspeech motivations and barriers (RQ2). We found that age was negatively associated with several barriers, such as fear, emotional and skill barriers, and engagement hesitation, meaning that younger adults are more likely to be deterred by these factors. Our findings are consistent with prior HCI research that highlights nuanced differences in how older versus younger adults perceive online risk [67] and safety [2, 3]. Another notable finding in our work is that women not only face a higher fear of retaliation from the perpetrator and third parties, but also fear being publicly exposed when countering hate through online counterspeech. This aligns with prior research documenting how women cope with online harassment by adopting gendered defensive strategies [29, 105]. Particularly, women tend to experience more depression and anxiety after being harassed online, partially explaining their heightened fear of retaliation [29, 122]. Moreover, we found that women

reported lower self-efficacy in their counterspeech writing skills and greater emotional burden concerns than men. Recent studies 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 [29]. Our work provides evidence that these factors may deter women from engaging in online counterspeech. In addition, research has shown that more educated individuals may better understand the potential risks and difficulties of public online engagement [14]. However, our findings show that participants with higher education levels report higher barriers stemming from fear-driven inhibition as well as emotional burden and skill barriers. Concerns of potential risks can prevent more educated individuals from writing counterspeech, even though they may have the necessary skills and knowledge to do so.

5.1.5 Demographic Factors That Shape Counterspeech Writing Experiences

RQ3 results show that compared to other demographic groups, younger, woman, more educated users tend to feel more satisfied towards their self-written counterspeech and perceive their counterspeech to be more effective. Interestingly, this group not only encounters hate speech more frequently, but is also more likely to have been a target of online hate speech in the past. Our results also show a positive correlation between the frequency of online hate speech exposure and fear-driven inhibition, meaning that the more often one encounters online hate speech, the stronger their fear-driven inhibitions to engage in counterspeech. As previously discussed, more exposure to online hate may increase peoples' awareness of the potential threats and challenges of countering it [14, 29, 107]. This may allow individuals to leverage their personal experiences and knowledge [75, 160], leading them to write more satisfying and effective counterspeech. Another explanation is that greater awareness of the potential risks and difficulties in countering online hate may motivate users to put more effort and care into crafting their responses [158]. This extra diligence could lead to more satisfaction and a stronger belief in the impact of their counterspeech [9]. Further research could explore these relationships in more depth.

One noteworthy aspect is that our measures of barriers and motivations capture general perceptions of counterspeech, whereas the measures of perceived satisfaction, difficulty, and effectiveness specifically relate to the counterspeech task in this study. Therefore, time concern showing no significant impact in our study could be because participants were explicitly asked to engage in counterspeech as part of the study, unlike real-world scenarios where time constraints might influence one's decision [10] to respond to hate speech. Similarly, fear-driven inhibition has a positive effect; this counterintuitive result can be explained by the absence of actual risks of retribution in the study setting. Participants who typically fear consequences in real-world interactions may have felt more at ease expressing themselves in this controlled environment, potentially leading to more positive self-evaluations of their counterspeech [38]. Future research could benefit from comparing controlled study environments with real-world counterspeech scenarios [5, 75].

5.1.6 The Role of AI Assistance in Counterspeech Writing: AI-mediated Counterspeech

AI-mediated communication can be defined as interpersonal communication that is not simply transmitted by technology but augmented or even generated by algorithms to achieve specific communicative or relational outcomes [88]. Since trust is fundamental to human relationships and manifests in collaborative behaviors such as a willingness to depend on and share information, previous research has placed a significant emphasis on examining people's perceptions of trustworthiness of messages generated with AI assistance [88, 106]. Our qualitative findings provide further nuance to prior research by revealing a tension between individuals' hesitations and motivations for using AI assistance in crafting online counterspeech. While the primary motivation for adopting such technology is rooted in its utilitarian advantages, reservations predominantly revolve around issues of trust. Most participants expressed distrust in AI's ability to accurately convey their emotions as well as the credibility of the information it presents. Given that prior research has found that peoples' trust in AI generated text decreases as the level of AI agency (the degree of autonomous content generation and decision-making by AI) increases, we propose AI-*mediated* counterspeech as an alternative to AI-*generated* counterspeech [106, 114]. We discuss specific design implications in 5.2.3.

5.2 Design Implications

5.2.1 *Incentivizing Participation Through Recognition*

Understanding the barriers that deter users from engaging in online counterspeech is a crucial step for developing effective design strategies to encourage proactive participation and intervention in online spaces. People's reluctance to engage on social media is one of the strongest barriers against participating in online counterspeech (RQ1). Prior research has shown that incentivizing user contribution through community acknowledgment through badge systems can increase engagement among those who prefer to be lurkers [27]. For instance, digital badges can function as an award mechanism through visible symbols of achievement and recognition [57, 104]. StackOverflow [57] and Reddit [104] use a system where users earn badges for varying levels of contribution, such as providing helpful answers or engaging in community moderation. Researchers have shown that these badges not only increase engagement, but also a sense of responsibility and motivation to ensure adherence to community norms in an empathetic manner [27]. However, in the context of counterspeech against hate speech, public recognition through badges may raise concerns about user safety and potentially inhibit participation. Research indicates that in hostile online environments, publicly visible identities can make users more vulnerable to negative experiences [17]. Given potential risks associated with public badges, platforms should explore alternative forms of recognition that balance encouragement with user safety. Relatedly, our work shows that confronting hate and supporting others are important motivators for engaging in counterspeech (RQ1). Platforms could leverage these motivations while addressing safety concerns by offering flexible recognition options. For instance, users could choose whether they want to receive public recognition or prefer more private forms of acknowledgment. Private feedback or anonymous point systems could encourage participation while protecting user identities [6].

5.2.2 *Mitigating Fear Through Personalized Interactions with Others' Responses to User's Counterspeech*

A notable finding in our study is that younger individuals, women, and those who are more educated feel greater fear related to engaging in online counterspeech, and that this fear is correlated with the amount of hateful content they encounter online (RQ2). To address this, HCI researchers can design personalized features to empower users to manage how they interact with responses, particularly harmful and retaliatory ones, to their counterspeech. Supporting this approach, prior HCI studies have identified a clear preference for personalized content moderation for harmful content, especially among users who have been victims of online hate [39, 138]. For instance, transgender Twitter users, who regularly encounter transphobic content online, appreciate the ability to automatically filter out posts containing offensive words specific to their personal preferences, eliminating the need to repetitively mute offenders or posts [84, 90]. In the context of reducing fear towards hateful responses to one's counterspeech, similar design features could be implemented to provide users with a more personalized approach in their ability to moderate responses to their counterspeech, while minimizing exposure to harmful responses from retaliators or third-parties. For example, users could have the choice to either subject these responses to pre-publication moderation or to engage an automatic filtering mechanism based on personally curated keywords for explicitly offensive reactions. Integrating personalized keyword filters or those based on community norms to blur out aggressive or offensive responses to one's counterspeech could potentially lessen the impact and, consequently, the fear of retaliation, thereby creating a safer and more controlled environment for users to engage in counterspeech.

5.2.3 *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 [70, 172]. Similarly, our qualitative analysis in RQ4 show that participants' willingness to use AI for counterspeech writing was often based on expectations of the AI's role and the degree of AI involvement in the writing process. For example, many expressed a preference to use AI to help them brainstorm ideas rather than having AI completely write the counterspeech for them. For such users, LLM-powered AI systems may be designed to facilitate brainstorming sessions, by allowing users to input words or phrases as fragments of their thoughts, or details they wish to provide based on personalized experiences. In response, the system may provide feedback and suggestions based on its training and understanding of what is considered constructive counterspeech. Furthermore, participants often indicated overcoming emotional burden or making sure than they do not sound overly angry in their responses as primary reasons for using AI

assistance in counterspeech writing. Such a desire for emotional regulation aligns with findings from Mun et al. (2024), where participants expressed wanting AI tools that could help them "remain composed" [119]. However, despite wanting to use AI for better articulation of their emotions and thoughts, users also expressed concerns around conveying authenticity. Interestingly, some users want AI to generate responses that were "exactly like they would" write [119]. The emphasis on personal authenticity in AI-assisted writing, alongside the need for emotional support, echoes broader discussions about user expectations in AI-mediated communication [74, 118, 174].

To address the complex needs of both emotional detachment and personal authenticity, research has shown that adaptive support systems can be highly effective when assistance is tailored to individual abilities and gradually reduced as competence increases [20, 177]. As a result, future research could explore adaptive AI assistance systems that could provide more extensive assistance when the user feels overwhelmed, but step back to a more advisory role when the user feels confident in responding authentically [49, 52]. Therefore, AI systems for counterspeech could be designed to provide more guidance when users struggle to formulate responses, but gradually step back to a more advisory role as users become more capable of expressing themselves confidently. This approach aligns with the concept of human-AI teaming, where the AI adapts its level of autonomy based on the human's cognitive state and task demands [70, 172]. Researchers in the HCI community, particularly those focusing on AI-assisted co-writing [66, 102], could examine these issues in future work by exploring ways to design human-AI interactions that can help users convey tone and emotion in their counterspeech, without diminishing the user's sense of personal agency and authenticity in their counterspeech writing process.

6 LIMITATIONS

While we have randomly assigned hate posts across diverse topics (gender, religion, disability, sexual orientation, and race) among our survey participants, we understand that such topics may impact each participant and their responses differently. Additionally, there is a difference in how survey questions on motivations and barriers to counterspeech were phrased, with barrier items framed generally (e.g., 'I feel that...') and motivation items framed conditionally (e.g., 'When I feel...'). This difference may have influenced how participants interpreted and responded to these items, as the conditional framing might have led respondents to consider specific scenarios, resulting in more contextualized responses for motivations compared to the more general responses for barriers. We recognize these limitations of our study, and plan to conduct a more detailed investigation in our future work to understand how these different topics in hate posts affect peoples' motivations and barriers when it comes to engaging in counterspeech to these posts, as well as their counterspeech to various topics.

Another limitation concerns the measurement scale used for RQ1, which employed relative quantifier labels ("Never", "Rarely", "Sometimes", "Often", "Frequently") to assess the frequency of counterspeech engagement on social media. As pointed out by Pohl (1981) [136], such relative quantifiers are prone to different interpretations among respondents. For instance, one participant's understanding of 'sometimes' may align with another's interpretation of 'frequently', potentially leading to inconsistent measurements. Nevertheless, in our context, relative quantifiers can provide certain contextual benefits. Schwarz et al. (1985) [151] note that providing specific frequency ranges can influence respondents' estimates of their own behavior. In the case of counterspeech, where frequency may vary greatly among individuals [23], predetermined ranges might not adequately capture the full spectrum of behaviors, thereby concentrating their responses in certain categories. Moreover, specific ranges could inadvertently suggest expected or "normal" frequencies of counterspeech engagement, potentially biasing respondents towards choosing what they perceive as socially desirable or expected responses. Considering these factors, we interpret our findings as representing participants' reported frequency of engaging in counterspeech, rather than absolute behavioral frequencies.

We understand that differences in participants' primary social media platform of choice may influence their survey responses. Therefore, we included this variable as a control in the linear regressions of RQ1 and RQ4 to verify the impact. However, our analysis revealed that this variable neither altered the significance nor the direction of the regression coefficient; hence, we moved these results to Table A5 (Appendix) and removed this non-significant variable to simplify the regression models in the main text. Another limitation is the high

correlation (r = 0.805) between the Fear-Driven Inhibition factor and the Emotional & Skill Barrier factor. This correlation indicates an intrinsic relationship between fear-based responses and emotional burdens in online counterspeech engagement. Future research can investigate the mechanisms underlying this relationship to better understand how these factors jointly influence counterspeech behavior. Finally, participants were asked to write a counterspeech in response to real-life posts containing hate speech in a survey setting, which is different from responding to a hateful post in real life. Hence, this may impact how they write counterspeech in addition to how they evaluate it.

7 CONCLUSION

Our work investigates the various motivations and barriers that underly online counterspeech engagement. To this end, we developed and validated a multi-item scale to assess these factors, demonstrating its significant influence on both the experience of writing counterspeech and people's perceptions towards their self-authored counterspeech. These measures can be used both as individual and latent factors, providing a scale that can be operationalized in future studies in relevant areas of scholarship. Using these latent variables, we demonstrate key demographic variances in counterspeech motivations and barriers, which have not been studied in prior research. Furthermore, we contribute to the emerging understanding of factors influencing people's openness to using AI assistance for crafting online counterspeech: while AI can assist in crafting responses, it cannot replace the human element essential for genuinely empathetic and contextually appropriate counterspeech. This finding underscores the necessity of human-AI collaboration, where AI's limitations are complemented by human insight and judgment. Our findings can guide tech firms and researchers in better understanding the role of AI in helping users combat hate speech on online platforms. This is especially timely as the implementation of AI technologies in facilitating online discourse is becoming more prominent.

References

[1] A better ChatGPT app: Poe wants to build the universal AI messaging client: 2023. https://www.theverge.com/23674656/poe-ai-chatbot-messaging-app. Accessed: 2023-09-06.

[2] Abraham, J. et al. 2022. Applying Behavioral Contagion Theory to Examining Young Adults' Participation in Viral Social Media Challenges. ACM Transactions on Social Computing. 5, 1–4 (Nov. 2022), 3:1-3:34. DOI:https://doi.org/10.1145/3538383.

[3] Ali, S. et al. 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, Apr. 2022), 1–14.

[4] Allison, K.R. and Bussey, K. 2016. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. Children and Youth Services Review. 65, (Jun. 2016), 183–194. DOI:https://doi.org/10.1016/j.childyouth.2016.03.026.

[5] Álvarez-Benjumea, A. and Winter, F. 2020. The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. Proceedings of the National Academy of Sciences. 117, 37 (Sep. 2020), 22800–22804. DOI:https://doi.org/10.1073/pnas.2007977117.

[6] Amichai-Hamburger, Y. et al. 2016. Psychological factors behind the lack of participation in online discussions. Computers in Human Behavior. 55, (Feb. 2016), 268–277. DOI:https://doi.org/10.1016/j.chb.2015.09.009.

[7] Ashktorab, Z. and Vitak, J. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, May 2016), 3895–3905.

[8] Author, N. 2017. 3. Witnessing online harassment. Pew Research Center: Internet, Science & Tech.

[9] Bac, M. 2014. Opinion expressions under social sanctions. International Review of Law and Economics. 38, (Jun. 2014), 58–71. DOI:https://doi.org/10.1016/j.irle.2014.03.002.

[10] Baek, Y.M. et al. 2012. Online versus face-to-face deliberation: Who? Why? What? With what effects? New Media & Society. 14, 3 (May 2012), 363–383. DOI:https://doi.org/10.1177/1461444811413191.

[11] Baider, F. 2023. Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech. Politics and Governance. 11, 2 (May 2023), 249–260. DOI:https://doi.org/10.17645/pag.v11i2.6465.

[12] Balakrishnan, V. 2018. Actions, emotional reactions and cyberbullying – From the lens of bullies, victims, bully-victims and bystanders among Malaysian young adults. Telematics and Informatics. 35, 5 (Aug. 2018), 1190–1200. DOI:https://doi.org/10.1016/j.tele.2018.02.002.

[13] Balakrishnan, V. and Fernandez, T. 2018. Self-esteem, empathy and their impacts on cyberbullying among young adults. Telematics and Informatics. 35, 7 (Oct. 2018), 2028–2037. DOI:https://doi.org/10.1016/j.tele.2018.07.006.

[14] Bauman, S. 2023. Cyberbullying and online harassment: The impact on emotional health and well-being in higher education. Cyberbullying and Online Harms. Routledge.

[15] Benesch, S. et al. 2016. Considerations for successful counterspeech. Dangerous speech project. (2016).

[16] Billings, M. and Watts, L.A. 2010. Understanding dispute resolution online: using text to reflect personal and substantive issues in conflict. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA, Apr. 2010), 1447–1456.

[17] Blackwell, L. et al. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. Proceedings of the ACM on Human-Computer Interaction. 1, CSCW (Dec. 2017), 24:1-24:19. DOI:https://doi.org/10.1145/3134659.

[18] Boateng, G.O. et al. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. Frontiers in Public Health. 6, (Jun. 2018). DOI:https://doi.org/10.3389/fpubh.2018.00149.

[19] Bollen, K.A. 1989. Structural Equations with Latent Variables. John Wiley & Sons.

[20] Brawner, K.W. and Gonzalez, A.J. 2016. Modelling a learner's affective state in real time to improve intelligent tutoring effectiveness. Theoretical Issues in Ergonomics Science. 17, 2 (Mar. 2016), 183–210. DOI:https://doi.org/10.1080/1463922X.2015.1111463.

[21] Brody, N. and Vangelisti, A.L. 2016. Bystander Intervention in Cyberbullying. Communication Monographs. 83, 1 (Jan. 2016), 94–119. DOI:https://doi.org/10.1080/03637751.2015.1044256.

[22] Buerger, C. 2021. Counterspeech: A Literature Review.

[23] Buerger, C. 2021. #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. Social Media + Society. 7, 4 (Oct. 2021), 20563051211063843. DOI:https://doi.org/10.1177/20563051211063843.

[24] Buerger, C. 2022. Why They Do It: Counterspeech Theories of Change. SSRN Electronic Journal. (2022). DOI:https://doi.org/10.2139/ssrn.4245211.

[25] Carlo, G. et al. 2012. The interplay of emotional instability, empathy, and coping on prosocial and aggressive behaviors. Personality and Individual Differences. 53, 5 (Oct. 2012), 675–680. DOI:https://doi.org/10.1016/j.paid.2012.05.022.

[26] Cavusoglu, H. et al. 2015. Can Gamification Motivate Voluntary Contributions? The Case of StackOverflow Q&A Community. Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (New York, NY, USA, Feb. 2015), 171–174.

[27] Cavusoglu, H. et al. 2015. Can Gamification Motivate Voluntary Contributions?: The Case of StackOverflow Q&A Community. Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing. (2015). DOI:https://doi.org/10.1145/2685553.2698999.

[28] Cepollaro, B. et al. 2023. Counterspeech. Philosophy Compass. 18, 1 (2023), e12890. DOI:https://doi.org/10.1111/phc3.12890.

[29] Chadha, K. et al. 2020. Women's Responses to Online Harassment. International Journal of Communication. 14, (2020).

[30] Chancellor, S. et al. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (New York, NY, USA, Feb. 2016), 1201–1213.

[31] Chandrasekharan, E. et al. 2017. 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 (Dec. 2017), 1–22. DOI:https://doi.org/10.1145/3134666.

[32] Chen, A. 2015. Conversion via Twitter. The New Yorker.

[33] Chetty, N. and Alathur, S. 2018. Hate speech review in the context of online social networks. Aggression and Violent Behavior. 40, (May 2018), 108–118. DOI:https://doi.org/10.1016/j.avb.2018.05.003.

[34] Chung, Y.-L. et al. 2021. Multilingual Counter Narrative Type Classification. Proceedings of the 8th Workshop on Argument Mining (2021), 125–132.

[35] Citron, D.K. and Norton, H. 2011. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. Boston University Law Review. 91, (2011), 1435.

[36] Clark, M. and Bussey, K. 2020. The role of self-efficacy in defending cyberbullying victims. Computers in Human Behavior. 109, (Aug. 2020), 106340. DOI:https://doi.org/10.1016/j.chb.2020.106340.

[37] Cohen-Almagor, R. 2011. Fighting Hate and Bigotry on the Internet. Policy & Internet. 3, 3 (2011), 1–26. DOI:https://doi.org/10.2202/1944-2866.1059.

[38] Coles, B.A. and West, M. 2016. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. Computers in Human Behavior. 60, (Jul. 2016), 233–244. DOI:https://doi.org/10.1016/j.chb.2016.02.070.

[39] Cook, C. et al. 2021. Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. 3, (2021). DOI:https://doi.org/10.3389/fhumd.2021.626409.

[40] Cortiz, D. and Zubiaga, A. 2021. Ethical and technical challenges of AI in tackling hate speech. (2021). DOI:https://doi.org/10.29173/irie416.

[41] Costello, A.B. and Osborne, J. 2019. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical assessment, research, and evaluation. 10, 1 (2019), 7.

[42] Costello, M. et al. 2017. 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 (Oct. 2017), 587–605. DOI:https://doi.org/10.1177/0894439316666272.

[43] Costello, M. et al. 2016. Virtually Standing Up or Standing By? Correlates of Enacting Social Control Online. International Journal of Criminology and Sociology. 6, (Feb. 2016), 16–28. DOI:https://doi.org/10.6000/1929-4409.2017.06.03.

[44] Croasmun, J.T. and Ostrom, L. 2011. Using Likert-Type Scales in the Social Sciences. Journal of Adult Education. 40, 1 (2011), 19–22.

[45] Cuadrado, E. et al. 2016. Determinants of Prosocial Behavior in Included Versus Excluded Contexts. Frontiers in Psychology. 6, (2016).

[46] Dahlberg, L. 2001. The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. Information, Communication & Society. 4, 4 (Jan. 2001), 615–633. DOI:https://doi.org/10.1080/13691180110097030.

[47] DeVellis, R.F. and Thorpe, C.T. 2021. Scale Development: Theory and Applications. SAGE Publications.

[48] DeVito, M.A. et al. 2018. "Too Gay for Facebook": Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. Proceedings of the ACM on Human-Computer Interaction. 2, CSCW (Nov. 2018), 44:1-44:23. DOI:https://doi.org/10.1145/3274313.

[49] Dhillon, P.S. et al. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, May 2024), 1–18.

[50] DiFranzo, D. et al. 2018. Upstanding by Design: Bystander Intervention in Cyberbullying. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC Canada, Apr. 2018), 1–12.

[51] Dinan, E. et al. 2022. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2022), 4113–4133.

[52] Ding, X. et al. 2024. CounterQuill: Investigating the Potential of Human-AI Collaboration in Online Counterspeech Writing. arXiv.

[53] Domínguez-Hernández, F. et al. 2018. A systematic literature review of factors that moderate bystanders' actions in cyberbullying. Cyberpsychology: Journal of Psychosocial Research on Cyberspace. 12, 4 (Dec. 2018). DOI:https://doi.org/10.5817/CP2018-4-1.

[54] Douek, E. 2021. Governing online speech: From" posts-as-trumps" to proportionality and probability. Colum. L. Rev. 121, (2021), 759.

[55] Dow, M. and Frenett, R. 2014. One to One Online Interventions A pilot CVE methodology. (Jan. 2014).

[56] Duncan, O.D. 2014. Introduction to Structural Equation Models. Elsevier.

[57] Easley, D. and Ghosh, A. 2016. Incentives, Gamification, and Game Theory: An Economic Approach to Badge Design. ACM Transactions on Economics and Computation. 4, 3 (Jun. 2016), 16:1-16:26. DOI:https://doi.org/10.1145/2910575.

[58] Edwards, A. 2002. Bowling Together. Online Public Engagement in Policy Deliberation, by Stephen Coleman and John Gøtze. Information Polity. 7, (Dec. 2002), 247–252. DOI:https://doi.org/10.3233/IP-2002-0021.

[59] Elsaesser, C.M. et al. 2021. Avoiding fights on social media: Strategies youth leverage to navigate conflict in a digital era. Journal of Community Psychology. 49, 3 (2021), 806–821. DOI:https://doi.org/10.1002/jcop.22363.

[60] Ernst, J. et al. 2017. Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. Journal for Deradicalization. 10 (2017), 1–49.

[61] Fanton, M. et al. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. 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) (Online, Aug. 2021), 3226–3240.

[62] Fortuna, P. et al. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Information Processing & Management. 58, 3 (May 2021), 102524. DOI:https://doi.org/10.1016/j.ipm.2021.102524.

[63] Freis, S.D. and Gurung, R.A.R. 2013. A Facebook analysis of helping behavior in online bullying. Psychology of Popular Media Culture. 2, 1 (2013), 11–19. DOI:https://doi.org/10.1037/a0030239.

[64] Garland, J. et al. 2020. Countering hate on social media: Large scale classification of hate and counter speech. Proceedings of the Fourth Workshop on Online Abuse and Harms (Online, Nov. 2020), 102–112.

[65] Garland, J. et al. 2020. Impact and dynamics of hate and counter speech online. EPJ Data Science. 11, (2020). DOI:https://doi.org/10.1140/epjds/s13688-021-00314-6.

[66] Gero, K.I. et al. 2023. Social Dynamics of AI Support in Creative Writing. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, Apr. 2023), 1–15.

[67] Ghaiumy Anaraky, R. et al. 2021. To Disclose or Not to Disclose: Examining the Privacy Decision-Making Processes of Older vs. Younger Adults. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, May 2021), 1–14.

[68] Gillespie, T. et al. 2023. Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates.

[69] Goddard, R.D. et al. 2004. Collective Efficacy Beliefs:Theoretical Developments, Empirical Evidence, and Future Directions. Educational Researcher. 33, 3 (Apr. 2004), 3–13. DOI:https://doi.org/10.3102/0013189X033003003.

[70] Grudin, J. 2022. From Tool to Partner: The Evolution of Human-Computer Interaction. Springer Nature.

[71] Guo, L. and Johnson, B.G. 2020. Third-Person Effect and Hate Speech Censorship on Facebook. Social Media + Society. 6, 2 (Apr. 2020), 2056305120923003. DOI:https://doi.org/10.1177/2056305120923003.

[72] Haimson, O.L. et al. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. Proceedings of the ACM on Human-Computer Interaction. 5, CSCW2 (Oct. 2021), 466:1-466:35. DOI:https://doi.org/10.1145/3479610.

[73] Han, X. and Tsvetkov, Y. 2020. Fortifying Toxic Speech Detectors Against Veiled Toxicity. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Online, Nov. 2020), 7732–7739.

[74] Hancock, J.T. et al. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. Journal of Computer-Mediated Communication. 25, 1 (Mar. 2020), 89–100. DOI:https://doi.org/10.1093/jcmc/zmz022.

[75] Hangartner, D. et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. Proceedings of the National Academy of Sciences. 118, 50 (Dec. 2021), e2116310118. DOI:https://doi.org/10.1073/pnas.2116310118.

[76] Hargittai, E. 2007. Whose Space? Differences among Users and Non-Users of Social Network Sites. Journal of Computer-Mediated Communication. 13, 1 (Oct. 2007), 276–297. DOI:https://doi.org/10.1111/j.1083-6101.2007.00396.x.

[77] Hassan, S. et al. 2018. Social media influencer and cyberbullying: A lesson learned from preliminary findings. (2018).

[78] Hawdon, J. et al. 2017. Exposure to Online Hate in Four Nations: A Cross-National Consideration. Deviant Behavior. 38, 3 (Mar. 2017), 254–266. DOI:https://doi.org/10.1080/01639625.2016.1196985.

[79] Hawdon, J. et al. 2015. Online extremism and online hate: Exposure among adolescents and young adults in four nations. NORDICOM INFORMATION. 37, 3–4 (2015), 29–37.

[80] Hayduk, L.A. and Littvay, L. 2012. Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? BMC Medical Research Methodology. 12, 1 (Oct. 2012), 159. DOI:https://doi.org/10.1186/1471-2288-12-159.

[81] He, B. et al. 2021. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (Virtual Event Netherlands, Nov. 2021), 90–94.

[82] Heise, D.R. 1975. Causal analysis. John Wiley & Sons.

[83] Henson, B. et al. 2020. There Is Virtually No Excuse: The Frequency and Predictors of College Students' Bystander Intervention Behaviors Directed at Online Victimization. Violence Against Women. 26, 5 (Apr. 2020), 505–527. DOI:https://doi.org/10.1177/1077801219835050.

[84] Ho, L. 2022. Countering Personalized Speech. SSRN Electronic Journal. (2022). DOI:https://doi.org/10.2139/ssrn.4117895.

[85] Horta Ribeiro, M. et al. 2023. Deplatforming did not decrease Parler users' activity on fringe social media. PNAS Nexus. 2, 3 (Mar. 2023), pgad035. DOI:https://doi.org/10.1093/pnasnexus/pgad035.

[86] Hu, L. and Bentler, P.M. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 6, 1 (Jan. 1999), 1–55. DOI:https://doi.org/10.1080/10705519909540118.

[87] Ireland, L. et al. 2020. Preconditions for guardianship interventions in cyberbullying: Incident interpretation, collective and automated efficacy, and relative popularity of bullies. Computers in Human Behavior. 113, (Dec. 2020), 106506. DOI:https://doi.org/10.1016/j.chb.2020.106506.

[88] Jakesch, M. et al. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow Scotland Uk, May 2019), 1–13.

[89] Jhaver, S. et al. 2018. Online Harassment and Content Moderation. ACM Transactions on Computer-Human Interaction (TOCHI). 25, (2018). DOI:https://doi.org/10.1145/3185593.

[90] Jhaver, S. et al. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. Proceedings of the ACM on Human-Computer Interaction. 7, CSCW2 (Oct. 2023), 289:1-289:33. DOI:https://doi.org/10.1145/3610080.

[91] Ji, Z. et al. 2023. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. 55, 12 (Mar. 2023), 248:1-248:38. DOI:https://doi.org/10.1145/3571730.

[92] Joinson, A.N. 2008. Looking at, looking up or keeping up with people? motives and use of facebook. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA, Apr. 2008), 1027–1036.

[93] Jr, H.M.B. 2018. Causal Inferences in Nonexperimental Research. UNC Press Books.

[94] Keipi, T. et al. 2016. Online Hate and Harmful Content: Cross-National Perspectives. Taylor & Francis.

[95] Kline, R. and St, C. 2022. Principles and Practice of Structural Equation Modeling.

[96] Lai, V. et al. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, Apr. 2022), 1–18.

[97] Lammerts, P. et al. 2023. How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (New York, NY, USA, Aug. 2023), 834–844.

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

[99] Laville, S. 2016. Top tech firms urged to step up online abuse fightback. The Guardian.

[100] Lee, A.L.C., Howard B. 2013. A First Course in Factor Analysis. Psychology Press.

[101] Lee, H. et al. 2022. ELF22: A Context-based Counter Trolling Dataset to Combat Internet Trolls. Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022), 3530–3541.

[102] Lee, M. et al. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. CHI Conference on Human Factors in Computing Systems (Apr. 2022), 1–19.

[103] Leonhard, L. et al. 2018. Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. SCM Studies in Communication and Media. 7, 4 (Dec. 2018), 555–579. DOI:https://doi.org/10.5771/2192-4007-2018-4-555.

[104] Lewis, C. 2014. Irresistible Apps: Motivational Design Patterns for Apps, Games, and Web-based Communities. Apress.

[105] Lindsay, M. et al. 2016. Experiences of Online Harassment Among Emerging Adults. Journal of Interpersonal Violence. 31, (2016). DOI:https://doi.org/10.1177/0886260515584344.

[106] Liu, Y. et al. 2022. Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing. CHI Conference on Human Factors in Computing Systems (New Orleans LA USA, Apr. 2022), 1–13.

[107] Luong, G. and Charles, S. 2014. Age differences in affective and cardiovascular responses to a negative social interaction: the role of goals, appraisals, and emotion regulation. Developmental psychology. 50 7, (2014). DOI:https://doi.org/10.1037/a0036621.

[108] Machackova, H. et al. 2015. Brief report: The bystander effect in cyberbullying incidents. Journal of Adolescence. 43, (Aug. 2015), 96–99. DOI:https://doi.org/10.1016/j.adolescence.2015.05.010.

[109] Madden, C. and Loh, J. (M. I.) 2020. Workplace cyberbullying and bystander helping behaviour. The International Journal of Human Resource Management. 31, 19 (Oct. 2020), 2434–2458. DOI:https://doi.org/10.1080/09585192.2018.1449130.

[110] Marques, D.R. 2021. What Type of Factor Analysis Are You Doing? Implications for Sleep Medicine Field. Sleep and Vigilance. 5, 2 (Dec. 2021), 337–338. DOI:https://doi.org/10.1007/s41782-021-00173-1.

[111] Masullo Chen, G. and Lu, S. 2017. Online Political Discourse: Exploring Differences in Effects of Civil and Uncivil Disagreement in News Website Comments. Journal of Broadcasting & Electronic Media. 61, 1 (Jan. 2017), 108–125. DOI:https://doi.org/10.1080/08838151.2016.1273922.

[112] Mathew, B. et al. 2020. Hate begets Hate: A Temporal Study of Hate Speech. Proceedings of the ACM on Human-Computer Interaction. 4, CSCW2 (Oct. 2020), 92:1-92:24. DOI:https://doi.org/10.1145/3415163.

[113] Mathew, B. et al. 2019. Thou Shalt Not Hate: Countering Online Hate Speech. Proceedings of the International AAAI Conference on Web and Social Media. 13, (Jul. 2019), 369–380. DOI:https://doi.org/10.1609/icwsm.v13i01.3237.

[114] McKnight, D.H. et al. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. Information Systems Research. 13, 3 (Sep. 2002), 334–359. DOI:https://doi.org/10.1287/isre.13.3.334.81.

[115] Meske, C. and Bunde, E. 2023. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. Information Systems Frontiers. 25, 2 (Apr. 2023), 743–773. DOI:https://doi.org/10.1007/s10796-021-10234-5.

[116] Miškolci, J. et al. 2020. Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. Social Science Computer Review. 38, 2 (Apr. 2020), 128–146. DOI:https://doi.org/10.1177/0894439318791786.

[117] Mollas, I. et al. 2022. ETHOS: an Online Hate Speech Detection Dataset. Complex & Intelligent Systems. 8, 6 (Dec. 2022), 4663–4678. DOI:https://doi.org/10.1007/s40747-021-00608-2.

[118] Monrad, M. 2024. Feeling rules in artificial intelligence: norms for anger management. (Jul. 2024). DOI:https://doi.org/10.1332/26316897Y2024D000000016.

[119] Mun, J. et al. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. Proceedings of the CHI Conference on Human Factors in Computing Systems (New York, NY, USA, May 2024), 1–22.

[120] Munger, K. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. Political Behavior. 39, 3 (Sep. 2017), 629–649. DOI:https://doi.org/10.1007/s11109-016-9373-5.

[121] N et al. 2017. Megan Phelps-Roper: If You're Raised To Hate, Can You Reverse It? NPR.

[122] Nadim, M. and Fladmoe, A. 2019. Silencing Women? Gender and Online Harassment. Social Science Computer Review. 39, (2019). DOI:https://doi.org/10.1177/0894439319865518.

[123] Nextdoor Is Integrating Generative AI to Drive Engaging and Kind Conversations in the Neighborhood: 2023. https://finance.yahoo.com/news/nextdoor-integrating-generative-ai-drive-103000201.html. Accessed: 2023-09-06.

[124] Nickerson, A.B. et al. 2014. Perceptions of School Climate as a Function of Bullying Involvement. Journal of Applied School Psychology. 30, 2 (Apr. 2014), 157–181. DOI:https://doi.org/10.1080/15377903.2014.888530.

[125] Obermaier, M. et al. 2016. Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. New Media & Society. 18, 8 (Sep. 2016), 1491–1507. DOI:https://doi.org/10.1177/1461444814563519.

[126] Obermaier, M. et al. 2021. I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. New Media & Society. (Aug. 2021), 146144482110175. DOI:https://doi.org/10.1177/14614448211017527.

[127] Obermaier, M. et al. 2023. Too civil to care? How online hate speech against different social groups affects bystander intervention. European Journal of Criminology. 20, 3 (May 2023), 817–833. DOI:https://doi.org/10.1177/14773708231156328.

[128] Obermaier, M. 2022. Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. New Media & Society. (Oct. 2022), 14614448221125417. DOI:https://doi.org/10.1177/14614448221125417.

[129] O'brien, R.M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity. 41, 5 (Oct. 2007), 673–690. DOI:https://doi.org/10.1007/s11135-006-9018-6.

[130] Ousidhoum, N. et al. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China, Nov. 2019), 4675–4684.

[131] Papacharissi, Z. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. New Media & Society. 6, (2004), 259–283. DOI:https://doi.org/10.1177/1461444804041444.

[132] Parker, S. and Ruths, D. 2023. Is hate speech detection the solution the world wants? Proceedings of the National Academy of Sciences. 120, 10 (Mar. 2023), e2209384120. DOI:https://doi.org/10.1073/pnas.2209384120.

[133] Pater, J.A. et al. 2016. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (New York, NY, USA, Feb. 2016), 1185–1200.

[134] Petterson, A. et al. 2023. Supporting Social Movements Through HCI and Design. Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, Apr. 2023), 1–5.

[135] Piliavin, J.A. et al. 1976. Time of Arrival at an Emergency and Likelihood of Helping. Personality and Social Psychology Bulletin. 2, 3 (Jul. 1976), 273–276. DOI:https://doi.org/10.1177/014616727600200314.

[136] Pohl, N.F. 1981. Scale Considerations in Using Vague Quantifiers. The Journal of Experimental Education. 49, 4 (Jun. 1981), 235–240. DOI:https://doi.org/10.1080/00220973.1981.11011790.

[137] Qian, J. et al. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019), 4755–4764.

[138] Reid Chassiakos, Y.L. et al. 2016. Children and Adolescents and Digital Media. Pediatrics. 138, 5 (Nov. 2016), e20162593. DOI:https://doi.org/10.1542/peds.2016-2593.

[139] Reuter, C. et al. 2017. Social Media in Emergencies: A Representative Study on Citizens' Perception in Germany. Proceedings of the ACM on Human-Computer Interaction. 1, CSCW (Dec. 2017), 90:1-90:19. DOI:https://doi.org/10.1145/3134725.

[140] Reynolds, L. and Tuck, H. THE COUNTER-NARRATIVE MONITORING & EVALUATION HANDBOOK.

[141] Richards, R.D. and Calvert, C. 2000. Counterspeech 2000: A New Look at the Old Remedy for Bad Speech. Brigham Young University Law Review. 2000, (2000), 553.

[142] Rieger, D. et al. 2018. Hate and counter-voices in the Internet: Introduction to the special issue. SCM Studies in Communication and Media. 7, 4 (Dec. 2018), 459–472. DOI:https://doi.org/10.5771/2192-4007-2018-4-459.

[143] Rösner, L. and Krämer, N.C. 2016. Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. Social Media + Society. 2, 3 (Jul. 2016), 2056305116664220. DOI:https://doi.org/10.1177/2056305116664220.

[144] Rosseel, Y. 2012. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software. 48, (May 2012), 1–36. DOI:https://doi.org/10.18637/jss.v048.i02.

[145] Rudnicki, K. et al. 2023. Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. Behaviour & Information Technology. 42, 5 (Apr. 2023), 527–544. DOI:https://doi.org/10.1080/0144929X.2022.2027013.

[146] Ruths, D.R. et al. 2016. Considerations for Successful Counterspeech. Dangerous Speech Project.

[147] Saha, P. et al. 2022. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. (Jul. 2022), 5157–5163.

[148] Saltman, E. et al. 2023. New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis. Studies in Conflict & Terrorism. 46, 9 (Sep. 2023), 1547–1574. DOI:https://doi.org/10.1080/1057610X.2021.1888404.

[149] Sasse, J. and Grossklags, J. 2023. Breaking the Silence: Investigating Which Types of Moderation Reduce Negative Effects of Sexist Social Media Content. Proceedings of the ACM on Human-Computer Interaction. 7, CSCW2 (Oct. 2023), 327:1-327:26. DOI:https://doi.org/10.1145/3610176.

[150] Schieb, C. and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook.

[151] SCHWARZ, N. et al. 1985. Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. Public Opinion Quarterly. 49, 3 (Jan. 1985), 388–395. DOI:https://doi.org/10.1086/268936.

[152] Schwertberger, U. and Rieger, D. 2021. Hass und seine vielen Gesichter: Eine sozial- und kommunikationswissenschaftliche Einordnung von Hate Speech. Hate Speech - Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen. S. Wachs et al., eds. Springer Fachmedien. 53–77.

[153] Seering, J. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. Proceedings of the ACM on Human-Computer Interaction. 4, CSCW2 (Oct. 2020), 107:1-107:28. DOI:https://doi.org/10.1145/3415178.

[154] Seering, J. et al. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (New York, NY, USA, Feb. 2017), 111–125.

[155] Seo, M. 2020. Bystanders' Experience in Cyber Bullying among Adolescents: Focused on Group Chat Room. THE KOREAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY. 33, 3 (Sep. 2020), 65–88. DOI:https://doi.org/10.35574/KJDP.2020.9.33.3.65.

[156] Shi, D. et al. 2019. Understanding the Model Size Effect on SEM Fit Indices. Educational and Psychological Measurement. 79, 2 (Apr. 2019), 310–334. DOI:https://doi.org/10.1177/0013164418783530.

[157] Shultz, E. et al. 2014. Cyber-bullying: An exploration of bystander behavior and motivation. Cyberpsychology: Journal of Psychosocial Research on Cyberspace. 8, 4 (Dec. 2014). DOI:https://doi.org/10.5817/CP2014-4-3.

[158] Siebert, J. and Siebert, J.U. 2023. Effective mitigation of the belief perseverance bias after the retraction of misinformation: Awareness training and counter-speech. PLOS ONE. 18, 3 (Mar. 2023), e0282202. DOI:https://doi.org/10.1371/journal.pone.0282202.

[159] Silverman, T. et al. 2016. The impact of counter-narratives. Institute for Strategic Dialogue. 54, (2016).

[160] Smeenk, W. et al. 2018. A systematic validation of the Empathic Handover approach guided by five factors that foster empathy in design. CoDesign. 15, (2018). DOI:https://doi.org/10.1080/15710882.2018.1484490.

[161] Smith, P.K. et al. 2008. Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry. 49, 4 (2008), 376–385. DOI:https://doi.org/10.1111/j.1469-7610.2007.01846.x.

[162] Soral, W. et al. 2018. Exposure to hate speech increases prejudice through desensitization. Aggressive Behavior. 44, (2018). DOI:https://doi.org/10.1002/ab.21737.

[163] Strauss, A.L. and Corbin, J.M. 1998. Basics of qualitative research: techniques and procedures for developing grounded theory. Sage Publications.

[164] Stroud, S.R. and Cox, W. 2018. The Varieties of Feminist Counterspeech in the Misogynistic Online World. Mediating Misogyny: Gender, Technology, and Harassment. J.R. Vickery and T. Everbach, eds. Springer International Publishing. 293–310.

[165] Taylor, S.H. et al. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. Proceedings of the ACM on Human-Computer Interaction. 3, CSCW (Nov. 2019), 118:1-118:26. DOI:https://doi.org/10.1145/3359220.

[166] Tekiroğlu, S.S. et al. 2022. Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study. Findings of the Association for Computational Linguistics: ACL 2022 (2022), 3099–3114.

[167] Tokunaga, R.S. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior. 26, 3 (May 2010), 277–287. DOI:https://doi.org/10.1016/j.chb.2009.11.014.

[168] Ubangha, C. 2016. Hate Speech in Cyberspace: Why Education is Better than Regulation.

[169] Vashistha, A. et al. 2019. Threats, Abuses, Flirting, and Blackmail: Gender Inequity in Social Media Voice Forums. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (New York, NY, USA, May 2019), 1–13.

[170] Wachs, S. et al. 2021. 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, (Oct. 2021), 106872. DOI:https://doi.org/10.1016/j.chb.2021.106872.

[171] Wan, T. et al. 2015. Kappa coefficient: a popular measure of rater agreement. Shanghai archives of psychiatry. 27, 1 (2015), 62.

[172] Wang, D. et al. 2021. Designing AI to Work WITH or FOR People? Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama Japan, May 2021), 1–5.

[173] Wang, Y. et al. 2015. Coming of Age (Digitally): An Ecological View of Social Media Use among College Students. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (New York, NY, USA, Feb. 2015), 571–582.

[174] Weber-Guskar, E. 2021. How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. Ethics and Information Technology. 23, 4 (Dec. 2021), 601–610. DOI:https://doi.org/10.1007/s10676-021-09598-8.

[175] Westboro Baptist Church | History & Facts | Britannica: 2024. https://www.britannica.com/topic/Westboro-Baptist-Church. Accessed: 2024-03-05.

[176] Wong, R.Y.M. et al. 2021. Standing Up or Standing By: Understanding Bystanders' Proactive Reporting Responses to Social Media Harassment. Information Systems Research. 32, 2 (Jun. 2021), 561–581. DOI:https://doi.org/10.1287/isre.2020.0983.

[177] Wood, D. et al. 1976. The Role of Tutoring in Problem Solving. Journal of Child Psychology and Psychiatry. 17, 2 (1976), 89–100. DOI:https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.

[178] Worthington, R.L. and Whittaker, T.A. 2006. Scale Development Research: A Content Analysis and Recommendations for Best Practices. The Counseling Psychologist. 34, 6 (Nov. 2006), 806–838. DOI:https://doi.org/10.1177/0011000006288127.

[179] Wright, S. 1921. Correlation and causation. (1921).

[180] Wright, S. 1934. The Method of Path Coefficients. The Annals of Mathematical Statistics. 5, 3 (1934), 161–215.

[181] Wright, S. and Street, J. 2007. Democracy, deliberation and design: the case of online discussion forums. New Media & Society. 9, 5 (Oct. 2007), 849–869. DOI:https://doi.org/10.1177/1461444807081230.

[182] Zhang, Q. et al. 2023. Internet altruistic behavior among Chinese early adolescents: Exploring differences in gender and collective efficacy using a latent growth modeling. Current Psychology. (May 2023). DOI:https://doi.org/10.1007/s12144-023-04660-8.

[183] Zhu, W. and Bhat, S. 2021. Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021), 134–149.

Received 11 March 2024; Revised 1 February 2025; accepted 10 May 2025

APPENDICES

Survey Questions

This section presents the key questions from the survey; for the complete list of questions, please refer to the link available on the Open Science Framework (OSF) [3].

Screen question: Imagine you are a user of an online group on social media. Another user (perpetrator) in the group posted the following: *[hateful post]*. Do you consider this post to be hateful? (Yes, No)

Satisfaction: How satisfied are you with the counterspeech that you've written? (Extremely dissatisfied, Somewhat dissatisfied, Neither satisfied nor dissatisfied, Somewhat satisfied, Extremely satisfied)

Difficulty: How difficult was it to write this counterspeech? (Extremely difficult, Somewhat difficult, Neither easy nor difficult, Somewhat easy, Extremely easy)

Self-perceived effectiveness: How effective do you think your counterspeech would be in preventing the perpetrator from engaging in further hateful behavior? (Not effective at all, Slightly effective, Moderately effective, Very effective, Extremely effective)

Prior experience of online hate speech target: Have you been a target of hateful speech on the internet? (Yes, No)

Barriers: How much do the following factors prevent you from writing a counterspeech on social media? (None at all, A little, A moderate amount, A lot, A great deal)

- B1: I fear being publicly exposed.
- B2: I'm afraid of retaliation from the perpetrator.
- B3: I'm afraid that I will be harassed by people (other than the perpetrator).
- B4: I don't want to spend time on this.

[3] https://osf.io/rzmg3/?view_only=6b2fd3a3d42b4b25a37f014612fac18a

- B5: Writing a counterspeech is emotionally burdensome.
- B6: I don't know how to write an effective counterspeech.
- B7: I feel that it's not my place to engage in counterspeech.
- B8: I don't like to engage in social media conversations.
- B9: I feel that my counterspeech would not make a difference.

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

- M1: When I feel the need to stand up for people I care about (e.g., family, close friends).
- M2: When I feel the need to stand up for people in general.
- M3: When I feel the need to stand up for myself.
- M4: When I want to confront a hateful person or behavior.
- M5: When I want to educate an ignorant person.
- M6: To signal that I stand for inclusion.
- M7: When it concerns issues or topics I care about.
- M8: When I want to blow off steam.

Frequency of writing online counterspeech: How often do you write counterspeech on social media? (Never, Rarely, Sometimes, Often, Frequently)

Prior use of ChatGPT: Have you used ChatGPT before? ChatGPT is an artificial intelligence tool that allows you to have human-like conversations by generating human-like responses to text-based inputs. ChatGPT can answer questions, and assist you with tasks such as composing emails, essays, and code. (Yes, No)

Perceived usefulness of ChatGPT: How useful do you find ChatGPT? (Not at all useful, Slightly useful, Moderately useful, Very useful, Extremely useful)

Willingness to use ChatGPT to help write counterspeech: If you were writing a counterspeech on social media, would you use artificial intelligence technology like ChatGPT to help you write it?

Social media platforms: What social media platform do you use most often? (Facebook, Instagram, Twitter/X, Linkedin, Reddit, YouTube, TikTok, Snapchat, Other, I don't currently use social media)

Social media usage: I have used social media to. (Stay informed on current events/news, Shop, Learn, Socialize, Entertain myself, Advocate for social issues I care about)

Post frequency: How often did you post on social media? (Never, Rarely, Sometimes, Often, Always)

Comment frequency: How often did you comment on content that you encountered on social media? (Never, Rarely, Sometimes, Often, Always)

Online Anonymous: Did you use your real name on social media? (Never, Rarely, Sometimes, Often, Always)

Social media experience: I encountered the following on social media. (Never, Rarely, Sometimes, Often, Always)

- Content that I disagree with.
- Content that I find hateful.
- Content that I find controversial.

Exploratory Factor Analysis (EFA) Procedures and Factor Selection

Initial Factor Extraction and Eigenvalues

The scree plot in Figure A1 displays the eigenvalues associated with each factor in our analysis. The graph shows a steep decline in eigenvalues for the first three factors, followed by a more gradual decrease. There is an inflection point at the fourth factor, where the curve begins to level off more distinctly. Following the Kaiser criterion, which recommends retaining factors with eigenvalues greater than 1.0 [41], we noted that the fourth factor's eigenvalue (0.960) was below this threshold. As a result, we decided to retain three factors for our EFA.

Determination of Number of Factors

The EFA resulted in a three-factor solution for the motivation and barrier items related to counterspeech on social media. Table A1 presents the rotated factor loading matrix, showing how each item loads onto the three extracted factors. Factor 1 primarily consists of motivation items (M1-M7). Factor 2 is predominantly composed of fear-related barrier items (B1-B3), time concern (B4), and emotional burden (B5). Factor 3 includes a mix of barrier items related to time, emotional burden, and barriers to engagement (B4, B5, B7-B9). Some items show cross-loadings across factors. For instance, B5 (Emotional burden) loads on both Factor 2 (.304) and Factor 3 (.453). Table A2 displays the correlations between the three extracted factors.

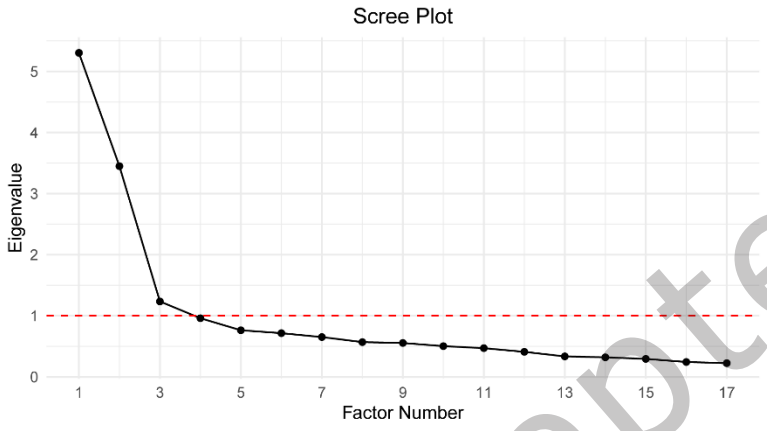

Figure A1: Scree Plot of Eigenvalues for Factor Analysis. It illustrates the relationship between eigenvalues and the number of factors. The red dashed line represents the traditional eigenvalue threshold of 1.0.

Table A1. Rotated Factor Loading Matrix for Motivation and Barrier Items

	Motivation and Barrier Items	Factor 1	Factor 2	Factor 3
Barriers	B1: Fear of public exposure		0.618	<0.200
	B2: Fear of perpetrator retaliation		0.937	
	B3: Fear of third-party harassment		0.789	<0.200
	B4: Time Concern			0.702
	B5: Emotional burden		0.304	0.453
	B6: Skill gap		0.361	<0.200
	B7: Engagement unqualified			0.576
	B8: Engagement reluctance		<0.200	0.470
	B9: Engagement ineffective			0.515
Motivations	M1: Supporting kin	0.767		
	M2: Supporting others	0.837		
	M3: Supporting self	0.714		
	M4: Confronting hate	0.818		
	M5: Educating ignorance	0.782		
	M6: Signaling inclusion	0.660		

	M7: Issue focus	0.828		
	M8: Venting emotions	0.363		

Note. Factor loadings < .200 are suppressed. All displayed loadings are significant at p < .05.

Although the Kaiser criterion, which recommends retaining factors with eigenvalues greater than 1.0, is a common rule of thumb, research has shown that this approach can sometimes lead to the extraction of too few factors [41]. While the EFA suggested a three-factor solution, we decided to adjust the model to five factors to conceptually distinct groupings that warranted separation. We applied a robust cut-off threshold of 0.55 for the factor loadings, following [100], and included variable items with strong correlations to the latent variables to ensure that our model was parsimonious and reliable [100]. The fear-related items (B1-B3) formed a clear "Fear-Driven Inhibition" factor (LV1), distinct from other barriers. The final results of this five-factor model can be found in Table 5 in Section 4.2 of the main text. Time concern (B4) emerged as a standalone factor (LV2), reflecting its unique nature as a practical constraint. Emotional burden and skill gap (B5-B6) clustered together as "Emotional & Skill Barrier" (LV3), representing internal challenges. The engagement-related barriers (B7-B8) formed an "Engagement Hesitation" factor (LV4), capturing reluctance to participate. Finally, all motivation items (M1-M7) loaded onto a single "Motivation" factor (LV5). This structure represents our proposed five-factor model for understanding motivations and barriers to counterspeech.

Table A2. Factor Correlation Matrix

	Factor 1	Factor 2	Factor 3
Factor 1	1.000		
Factor 2	0.009	1.000	
Factor 3	-0.327*	0.409*	1.000

* significant at 5% level

To validate our decision, we compared the performance of the EFA-derived three-factor model with our manually adjusted five-factor model. The results of this comparison can be found in Table A3. Comparing the three-factor and five-factor models reveals that the five-factor solution provides a better fit to the data across all indices. The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) are both higher for the five-factor model, indicating better overall fit. The Root Mean Square Error of Approximation (RMSEA) is lower for the five-factor model, with the confidence intervals overlapped.

Table A3. Model Fit Indices Comparison for Three-Factor and Five-Factor Models

	3 Factors Model	**5 Factors Model**
Comparative Fit Index (CFI)	0.934	0.952
Tucker-Lewis Index (TLI)	0.921	0.944
RMSEA (90% CI)	0.070 [0.062-0.077]	0.066 [0.057-0.076]

Note. RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval.

Based on these results, we ultimately selected the five-factor model for our analysis. This decision is supported by both statistical and theoretical considerations. Statistically, the five-factor model demonstrates superior fit across all indices. Theoretically, it provides a more granular and interpretable structure of motivations and barriers to counterspeech. This enhanced differentiation among barrier types (fear-driven, time-related, emotional/skill-based, and engagement-related) offers richer insights into the factors influencing counterspeech behavior. While the three-factor model from the EFA provided a useful starting point, the five-factor model ultimately offers a more comprehensive and nuanced framework for understanding the complex dynamics of counterspeech engagement. Table A4 shows the correlations between the five factors.

Reliability Analysis: We conducted reliability analysis for the Five-Factor Model of counterspeech motivations and barriers. Cronbach's alpha, omega, and average variance extracted (AVE) were calculated for four of the five factors. The Fear-Driven Inhibition factor ($\alpha = 0.84$, $\omega = 0.85$, $AVE = 0.65$) and Motivation factor ($\alpha = 0.91$, $\omega = 0.91$, $AVE = 0.60$) demonstrated good reliability. However, the Emotional & Skill Barrier factor ($\alpha = 0.51$, $\omega = 0.52$, $AVE = 0.36$) and the Engagement Hesitation factor ($\alpha = 0.56$, $\omega = 0.56$, $AVE = 0.39$) exhibited lower

reliability scores. Notably, reliability coefficients for the Time Concern factor were not computed, likely because this factor consists of only a single item, making internal consistency calculations not applicable.

Table A4. Factor Correlation Matrix for Five-Factor Models

	Fear-Driven Inhibition	Time Concern	Emotional & Skill Barrier	Engagement Hesitation	Motivation
Fear-Driven Inhibition	1.000				
Time Concern	0.216*	1.000			
Emotional & Skill Barrier	0.805*	0.471*	1.000		
Engagement Hesitation	0.547*	0.644*	0.838*	1.000	
Motivation	-0.006	-0.288*	-0.096	-0.446*	1.000

* significant at 5% level

Reliability and Validity

Discriminant Validity: We use factor correlation estimates and nested model comparisons to assess the discriminant validity for the Five-Factor Model. The results (Table A4) revealed varying degrees of distinctiveness among the factors. The Fear-Driven Inhibition factor showed moderate correlations with Time Concern (r = 0.22, 95% CI [0.13, 0.30]) and Engagement Hesitation (r = 0.55, 95% CI [0.45, 0.64]), but a strong correlation with Emotional & Skill Barrier (r = 0.80, 95% CI [0.71, 0.90]). The Time Concern factor demonstrated moderate correlations with Emotional & Skill Barrier (r = 0.47, 95% CI [0.37, 0.57]) and Engagement Hesitation (r = 0.64, 95% CI [0.56, 0.73]). Notably, the Emotional & Skill Barrier and Engagement Hesitation factors showed a very high correlation (r = 0.84, 95% CI [0.71, 0.96]), suggesting potential overlap between these constructs. The Motivation factor exhibited weak to moderate negative correlations with Time Concern (r = -0.29, 95% CI [-0.37, -0.21]) and Engagement Hesitation (r = -0.45, 95% CI [-0.55, -0.35]), and was not significantly correlated with Fear-Driven Inhibition or Emotional & Skill Barrier.

Content Validity: To assess the content validity of our measurement instrument, we conducted a thorough review of existing literature and theoretical frameworks related to motivations for and barriers to engaging in online counterspeech (Section 2.2.1 and 2.2.2). This approach allowed us to evaluate whether our items comprehensively captured the various aspects of the constructs under investigation. While our method did not involve expert ratings, the use of literature-based content validation is a widely accepted practice in scale development [47, 178]. This approach is particularly valuable when access to a panel of experts is limited [18]. Our analysis suggests that the items developed for both motivations and barriers demonstrate good content validity, as they align closely with the themes and factors identified in previous research on online behavior, bystander intervention, and counterspeech engagement. However, future studies could further strengthen the validity of these measures through expert evaluation or cognitive interviewing techniques.

Description of Latent Variables

Fear-Driven Inhibition (LV1): This latent factor captures a range of fears related to engaging in online counterspeech. It includes concerns about public exposure (B1), fears of retaliation from the perpetrator (B2), and fear of harassment from third parties (B3).

Time Concern (LV2): This one-item latent variable is the barrier associated with time concern in engaging in online counterspeech.

Emotional and Skill Barriers (LV3): This dual-variable factor addresses both the emotional toll and skill-related concerns when engaging in online counterspeech. The factor consists of the emotional burden of engaging in counterspeech (B5) along with uncertainty regarding how to write an effective counterspeech (B6).

Engagement Hesitation (LV4): This latent variable captures engagement-related barriers in engaging in online counterspeech, which includes feeling unqualified to engage in counterspeech (B7), and a general reluctance to engage in social media conversations as a reason for not engaging in online counterspeech (B8).

Motivation (LV5): Except for venting emotions (M8), this latent variable embodies all motivation variables for engaging in counterspeech – including standing up for kin (M1), others in general (M2), and oneself (M3), confronting hate (M4), educating the ignorant (M5), signaling inclusion (M6), focusing on issues of personal importance (M7).

Influence of Social Media Platforms

We also considered the influence of different social media platforms on counterspeech. We asked participants to indicate which platforms they currently mainly use from a list of 10 options: Facebook, Instagram, Twitter (X), Linkedin, Reddit, YouTube, TikTok, Snapchat, Other, and I don't currently use social media. Since the number of participants selecting "I don't currently use social media" was small (n=1), we combined this category with "Other" to meet the assumptions of ordinal logistic regression. We used Facebook as the reference level. The results are shown in Table A5. However, we found that these differences were not substantial enough to significantly affect the overall findings. Therefore, we did not include the platform variable in our main analysis.

Table A5. Effects of Social Media Platforms on All Dependent Variables

	Counterspeech Writing Frequency		**Willingness to Use AI Assistance**	
	β	**P**	**β**	**P**
Facebook (Ref level)	/	/	/	/
Instagram	0.730	0.057	-0.038	0.884
Twitter (X)	0.422	0.323	0.049	0.867
Linkedin	-0.665	0.664	-0.149	0.893
Reddit	0.134	0.790	-0.249	0.385
YouTube	0.123	0.760	-0.125	0.625
TikTok	-0.422	0.366	0.316	0.281
Snapchat	1.076	0.275	-0.680	0.413
Other/ I don't currently use social media	-2.230	0.051	-1.645	0.057

Tables

Table A6. Participant Demographics (N = 458)

Factor	Category	N
Age group	18-30	138
	31-60	293
	61-81	27
Gender	Man	226
	Woman	232
	Other or prefer not to answer	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

Table A7. Skewness and Kurtosis Values for Study Variables

Variables	Skewness	Kurtosis
B1: Fear of public exposure	1.0199	-0.2868
B2: Fear of perpetrator retaliation	0.8075	-0.5243
B3: Fear of third-party harassment	0.4763	-1.1007
B4: Time concern	0.1807	-1.2792
B5: Emotional burden	0.4477	-0.9449
B6: Skill gap	1.0397	0.0121
B7: Engagement unqualified	0.8927	-0.2607
B8: Engagement reluctance	-0.1199	-1.3187
B9: Engagement ineffective	-0.1231	-1.2474
M1: Supporting kin	-0.6866	-0.6277
M2: Supporting public	-0.2769	-1.0541
M3: Supporting self	-0.3554	-1.0229
M4: Confronting hate	-0.1312	-1.0783
M5: Educating ignorance	-0.1923	-1.2322
M6: Signaling inclusion	0.3174	-1.2194
M7: Issue focus	-0.3922	-0.8701
M8: Venting emotions	1.1159	0.1617
Age	0.5624	-0.2453
Gender	-0.0261	-2.0037
Ethnicity	0.0435	-2.0025
Education level	-0.3489	-0.9445
Sexual orientation	1.3746	-0.1106
Political View	-0.2163	-0.8723
Social media commenting frequency	-0.0202	-0.5396
Use of real name on social media	0.1299	-1.6865
Prior experience of online hate speech target	0.4180	-1.8292
Frequency of encountering online hate speech	0.2885	-0.2835
Prior use of ChatGPT	-0.6099	-1.6315
Perceived usefulness of ChatGPT	-0.1342	-0.8015
Satisfaction	-0.6209	0.1688
Difficulty	0.1199	-0.5220
Effectiveness	0.3032	-0.5542
Willingness to use ChatGPT to write counterspeech	0.3724	-0.8221

Table A8: Ordinal Logistic Regression Results for Counterspeech Writing Frequency on Social Media (N=458)

		β	SE	OR	P	VIF
Motivations	M1: Supporting kin	0.124	0.460	1.133	0.787	1.326
	M2: Supporting others	**1.471**	**0.515**	**4.353**	**0.004****	**1.416**
	M3: Supporting self	0.037	0.381	1.037	0.923	1.292
	M4: Confronting hate	**1.080**	**0.455**	**2.943**	**0.018***	**1.347**
	M5: Educating ignorance	0.644	0.419	1.905	0.125	1.327
	M6: Signaling inclusion	0.624	0.347	1.867	0.072	1.279
	M7: Issue focus	-0.299	0.516	0.742	0.562	1.345
	M8: Venting emotions	**0.743**	**0.362**	**2.101**	**0.040***	**1.167**
Barriers	B1: Fear of public exposure	0.325	0.410	1.384	0.428	1.299
	B2: Fear of perpetrator retaliation	-0.482	0.496	0.617	0.331	1.394
	B3: Fear of third-party harassment	-0.004	0.457	0.996	0.993	1.397
	B4: Time Concern	-0.630	0.337	0.533	0.061	1.270
	B5: Emotional burden	0.240	0.377	1.271	0.525	1.263
	B6: Skill gap	**-1.043**	**0.397**	**0.353**	**0.009****	**1.183**
	B7: Engagement unqualified	-0.479	0.456	0.619	0.293	1.228
	B8: Engagement reluctance	**-1.893**	**0.327**	**0.151**	**0.000*****	**1.237**
	B9: Engagement ineffective	0.064	0.347	1.067	0.853	1.253
SNS	**Past experience of online hate speech target**	**1.190**	**0.243**	**3.288**	**0.000*****	**1.207**
	Frequency of encountering hateful content	**1.439**	**0.463**	**4.216**	**0.002****	**1.154**
	Social media commenting frequency	**1.264**	**0.469**	**3.539**	**0.007****	**1.136**
	Use of real name on social media	**0.502**	**0.237**	**1.651**	**0.035***	**1.174**
Demographic	Age	0.000	0.009	1.000	0.960	1.260
	Gender	-0.004	0.245	0.996	0.988	1.271
	Ethnicity	-0.144	0.234	0.865	0.536	1.210
	Education level	-1.417	0.950	0.242	0.136	1.157
	Sexual orientation	-0.127	0.294	0.881	0.665	1.260
	Political views	0.200	0.344	1.221	0.561	1.162

β = regression coefficient; SE = standard error; OR = odds ratio (exp(β)); P = p-value; VIF = variance inflation factor.

*p < .05, **p < .01, ***p < .001.

McFadden R^2 = 0.341

Table A9. Interaction Effects Between Past Experience of Online Hate Speech Target and Motivation/Barrier Variables. This table presents interaction coefficients from ordinal logistic regression analyses. Main effects are omitted for brevity. Interaction terms represent the moderating effect of past experience as a target of online hate speech on the relationship between each motivation/barrier variable and the dependent variable.

Interaction Terms			β	SE	z-value	P
Past experience of online hate speech target	x	M1: Supporting kin	-1.512	0.672	-2.249	0.024*
		M2: Supporting others	1.671	0.575	2.907	0.004**
		M3: Supporting self	-1.357	0.656	-2.067	0.039*
		M4: Confronting hate	1.72	1.169	1.472	0.141
		M5: Educating ignorance	-1.334	1.004	-1.328	0.184
		M6: Signaling inclusion	-1.445	0.758	-1.906	0.057
		M7: Issue focus	1.86	1.225	1.519	0.129
		M8: Venting emotions	0.286	0.763	0.375	0.707
		B1: Fear of public exposure	0.787	0.937	0.84	0.401
		B2: Fear of perpetrator retaliation	-0.546	1.074	-0.508	0.611
		B3: Fear of third-party harassment	-0.387	0.978	-0.396	0.692
		B4: Time Concern	0.754	0.737	1.023	0.306
		B5: Emotional burden	1.711	0.88	1.945	0.052
		B6: Skill gap	0.622	0.909	0.684	0.494
		B7: Engagement unqualified	-1.362	0.669	-2.036	0.042*
		B8: Engagement reluctance	1.442	0.704	2.047	0.041*
		B9: Engagement ineffective	-1.758	0.764	-2.301	0.021*

β = regression coefficient; SE = standard error; z = z-score; P = p-value.

*p < .05, **p < .01, ***p < .001.

McFadden R^2 = 0.375

Table A10. Structural Equation Model Results for Factors Influencing Self-Perceived Satisfaction, Difficulty, and Effectiveness of Counterspeech

DVs	Factors	β	SE	z	P
Self-Perceived Satisfaction	**Fear-Driven Inhibition**	**0.536**	**0.181**	**2.294**	**0.022***
	Time Concern	0.106	0.061	0.897	0.370
	Emotional and Skill Barriers	**-1.255**	**0.348**	**-3.147**	**0.002****
	Engagement Hesitation	0.508	0.266	1.660	0.097
	Motivation	**0.450**	**0.090**	**3.998**	**0.000*****
Self-Perceived Difficulty	Fear-Driven Inhibition	-0.287	0.189	-1.456	0.145
	Time Concern	-0.138	0.066	-1.346	0.178

	Emotional and Skill Barriers	**0.921**	**0.366**	**2.714**	**0.007****
	Engagement Hesitation	-0.436	0.288	-1.625	0.104
	Motivation	**-0.313**	**0.096**	**-3.201**	**0.001****
	Fear-Driven Inhibition	**0.812**	**0.378**	**2.336**	**0.019***
	Time Concern	-0.108	0.132	-0.595	0.552
Self-Perceived Effectiveness	**Emotional and Skill Barriers**	**-1.993**	**0.729**	**-3.341**	**0.001****
	Engagement Hesitation	**-1.269**	**0.580**	**2.669**	**0.008****
	Motivation	**0.650**	**0.190**	**3.812**	**0.000*****

Note. β = standardized coefficient; SE = standard error; z = z-score; P = p-value. *p < .05, **p < .01, ***p < .001.

Table A11. Correlation Matrix of Self-Perceived Satisfaction, Difficulty, and Effectiveness in Counterspeech Writing

	Satisfaction	Difficulty	Effectiveness
Satisfaction	1.000		
Difficulty	-0.284*	1.000	
Effectiveness	0.069	0.132	1.000

* significant at 5% level

Table A12. Ordinal Logistic Regression of Willingness to Use ChatGPT to Write Counterspeech on Social Media

		β	**SE**	**OR**	**P**	**VIF**
Motivations	M1: Supporting kin	-0.383	0.412	0.682	0.353	1.380
	M2: Supporting others	-0.068	0.428	0.935	0.874	1.468
	M3: Supporting self	0.178	0.333	1.194	0.594	1.325
	M4: Confronting hate	0.056	0.401	1.057	0.889	1.409
	M5: Educating ignorance	-0.309	0.374	0.734	0.409	1.389
	M6: Signaling inclusion	0.485	0.318	1.624	0.127	1.299
	M7: Issue focus	0.642	0.430	1.901	0.136	1.400
	M8: Venting emotions	**-0.721**	**0.334**	**0.487**	**0.031***	**1.183**
Barriers	B1: Fear of public exposure	0.503	0.353	1.653	0.155	1.299
	B2: Fear of perpetrator retaliation	**0.939**	**0.411**	**2.557**	**0.022***	**1.393**
	B3: Fear of third-party harassment	-0.608	0.395	0.544	0.124	1.405
	B4: Time Concern	0.082	0.289	1.086	0.775	1.294
	B5: Emotional burden	-0.534	0.322	0.586	0.097	1.251
	B6: Skill gap	**1.200**	**0.347**	**3.319**	**0.001*****	**1.196**
	B7: Engagement unqualified	0.137	0.352	1.146	0.698	1.247
	B8: Engagement reluctance	-0.029	0.272	0.972	0.916	1.241
	B9: Engagement ineffective	0.023	0.305	1.023	0.940	1.271

SNS & ChatGPT	**Prior use of ChatGPT**	**1.031**	**0.211**	**2.805**	**0.000*** **	**1.159**
	Past experience of online hate speech target	**-0.515**	**0.217**	**0.598**	**0.018***	**1.213**
	Frequency of encountering online hate speech	-0.480	0.406	0.619	0.237	1.147
	Social media commenting frequency	0.383	0.411	1.467	0.350	1.161
	Use of real name on social media	-0.029	0.207	0.972	0.890	1.178
Demographic	Age	0.013	0.008	1.013	0.105	1.249
	Gender	-0.213	0.220	0.808	0.332	1.273
	Ethnicity	0.246	0.209	1.278	0.240	1.209
	Education level	-0.407	0.981	0.666	0.678	1.153
	Sexual orientation	0.149	0.269	1.160	0.581	1.281
	Political views	-0.506	0.329	0.603	0.124	1.186

β = regression coefficient; SE = standard error; OR = odds ratio (exp(β)); P = p-value; VIF = variance inflation factor.

*p < .05, **p < .01, ***p < .001.

McFadden R^2 = 0.121

Table A13. Interaction Effects Between Prior ChatGPT Usage and Motivation/Barrier Variables. This table presents interaction coefficients from ordinal logistic regression analyses. Main effects are omitted for brevity. Interaction terms represent the moderating effect of prior ChatGPT usage on the relationship between each motivation/barrier variable and the dependent variable.

		Interaction Terms	**β**	**SE**	**z-value**	**P**
Prior use of ChatGPT	x	M1: Supporting kin	4.229	1.093	3.87	0.000***
		M2: Supporting others	2.437	0.796	3.063	0.002**
		M3: Supporting self	-1.673	0.803	-2.083	0.037*
		M4: Confronting hate	-3.132	1.086	-2.885	0.004**
		M5: Educating ignorance	1.203	0.597	2.015	0.044*
		M6: Signaling inclusion	1.422	0.853	1.668	0.095
		M7: Issue focus	-1.154	0.566	-2.039	0.041*
		M8: Venting emotions	-0.702	0.827	-0.849	0.396
		B1: Fear of public exposure	3.172	0.999	3.176	0.001**
		B2: Fear of perpetrator retaliation	-1.807	1.107	-1.633	0.103
		B3: Fear of third-party harassment	-2.247	0.887	-2.533	0.011*
		B4: Time Concern	1.836	0.648	2.831	0.005**
		B5: Emotional burden	2.11	0.794	2.658	0.008**
		B6: Skill gap	-2.593	0.715	-3.629	0.000***
		B7: Engagement unqualified	1.898	0.961	1.975	0.048*
		B8: Engagement reluctance	-1.113	0.716	-1.555	0.12
		B9: Engagement ineffective	1.858	0.655	2.837	0.005**

β = regression coefficient; SE = standard error; z = z-score; P = p-value.

*p < .05, **p < .01, ***p < .001.

McFadden R^2 = 0.201

Table A14. Relationship Between Latent Variables and Covariates in the Structural Equation Model

IVs	Covariates	Estimate	Std. Est	Std. Error	z value	P-value
Fear-Driven Inhibition	Age	-1.948	-0.162	0.616	-3.160	0.001*
	Gender	0.114	0.228	0.024	4.723	0.000*
	Ethnicity	-0.019	-0.040	0.022	-0.853	0.394
	Education level	0.136	0.194	0.034	3.994	0.000*
	Sexual orientation	0.025	0.064	0.019	1.359	0.174
	Political views	0.009	0.008	0.052	0.177	0.860
	Past experience of online hate speech target	0.016	0.035	0.022	0.742	0.458
	Frequency of encountering hateful content	0.117	0.126	0.045	2.625	0.009*
	Social media commenting frequency	0.014	0.015	0.045	0.320	0.749
	Use of real name on social media	-0.055	-0.034	0.076	-0.718	0.473
Time Concern	Age	0.086	0.005	0.859	0.101	0.920
	Gender	-0.033	-0.047	0.032	-1.034	0.301
	Ethnicity	-0.022	-0.031	0.032	-0.689	0.491
	Education level	0.082	0.081	0.047	1.762	0.078
	Sexual orientation	0.013	0.023	0.027	0.506	0.613
	Political views	-0.050	-0.031	0.075	-0.670	0.503
	Past experience of online hate speech target	0.015	0.021	0.032	0.466	0.641
	Frequency of encountering hateful content	-0.032	-0.023	0.063	-0.501	0.616
	Social media commenting frequency	-0.269	-0.190	0.067	-4.023	0.000*
	Use of real name on social media	-0.144	-0.059	0.110	-1.302	0.193
Emotional & Skill Barrier	Age	-1.270	-0.115	0.558	-2.277	0.023*
	Gender	0.096	0.298	0.023	4.108	0.000*
	Ethnicity	-0.034	-0.083	0.021	-1.616	0.106
	Education level	0.092	0.181	0.031	2.997	0.003*
	Sexual orientation	-0.002	-0.007	0.017	-0.145	0.884
	Political views	0.016	0.016	0.051	0.310	0.757
	Past experience of online hate speech target	0.018	0.045	0.020	0.904	0.366
	Frequency of encountering hateful content	0.069	0.085	0.041	1.699	0.089
	Social media commenting frequency	-0.116	-0.139	0.044	-2.608	0.009*
	Use of real name on social media	-0.074	-0.053	0.072	-1.031	0.303
Engagement Hesitation	Age	-1.403	-0.124	0.624	-2.248	0.025*
	Gender	0.025	0.059	0.023	1.063	0.288
	Ethnicity	-0.009	-0.021	0.023	-0.393	0.694

	Education level	0.065	0.107	0.034	1.944	0.052
	Sexual orientation	-0.009	-0.025	0.019	-0.461	0.645
	Political views	-0.153	-0.134	0.055	-2.387	0.010*
	Past experience of online hate speech target	-0.007	-0.017	0.022	-0.322	0.747
	Frequency of encountering hateful content	0.022	0.027	0.045	0.500	0.617
	Social media commenting frequency	-0.229	-0.270	0.051	-4.513	0.000*
	Use of real name on social media	-0.003	-0.002	0.079	-0.037	0.970
Motivation	Age	0.992	0.080	0.555	1.786	0.074
	Gender	0.016	0.035	0.021	0.793	0.428
	Ethnicity	0.020	0.043	0.021	0.969	0.332
	Education level	-0.026	-0.039	0.030	-0.868	0.385
	Sexual orientation	0.011	0.028	0.017	0.622	0.534
	Political views	0.170	0.150	0.050	3.401	0.001*
	Past experience of online hate speech target	0.061	0.134	0.021	2.937	0.003*
	Frequency of encountering hateful content	0.163	0.181	0.042	3.859	0.000*
	Social media commenting frequency	0.265	0.286	0.046	5.716	0.000*
	Use of real name on social media	0.142	0.090	0.071	1.988	0.047*