Technische Universität München
TUM School of Social Sciences and Technology

The Effectiveness of Counterspeech in Mitigating Online Hate: Insights From a Multi-Method Investigation

Niklas Felix Cypris

Vollständiger Abdruck der von der TUM School of Social Sciences and Technology der Technischen Universität München zur Erlangung eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Jürgen Pfeffer

Prüfende der Dissertation:

1. Prof. Dr. Anna Baumert
2. Prof. Dr. Jens Großklags

Die Dissertation wurde am 03.04.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Social Sciences and Technology am 30.09.2024 angenommen.

Abstract

This dissertation investigated whether and how counterspeech can combat online hate speech. I examined counterspeech effects through a multi-method investigation encompassing a longitudinal study in a mock social media forum, cross-sectional vignette-based studies, and a field study on social media. This approach allowed me to isolate counterspeech effects on subsequent behavior and possible mediators in controlled experiments and further confirm their relevance in a real-world social media setting. I found that counterspeech not only emboldens bystanders to speak out against hate speech but that it can also reduce subsequent hate speech among both bystanders and transgressors. Across studies, counterspeech shaped collective and group-specific social norms and impacted how the severity of hate speech was perceived.

In Chapter A, I review contextual factors of online settings that may impede or facilitate morally courageous behavior in response to online hate speech.

Chapter B reports a longitudinal experiment in an interactive mock social media forum. The study assesses the direct impact of counterspeech on subsequent bystander counterspeech and potential mediators of its effect. Over the course of two weeks, participants interacted multiple times with a mock social media forum that contained hate speech comments by other ostensible users. I manipulated whether participants saw counterspeech or exclusively neutral replies to the hate speech. I report results regarding the participants' intentions to speak up themselves and actual behavior. In addition, I present mechanisms that mediate counterspeech impact on bystanders, focusing on perceptions of pro-counterspeech norms and hate speech severity. Finally, I discuss the longitudinal effects of counterspeech on the mediators and behavioral outcomes.

In Chapter C, I cover three cross-sectional experiments investigating the effect of counterspeech on further bystander counterspeech via ingroup norms as a potential mediating mechanism. In the experiments, I showed participants different social media post vignettes and comments by other ostensible participants, some of which were hate speech. I varied whether participants also saw counterspeech by an ingroup or an outgroup member. I present my findings regarding counterspeech impact on the participants' perception of pro-counterspeech ingroup norms and how these norms affect counterspeech endorsement and actual counterspeech. Moreover, I present the overall effects of counterspeech.

Chapter D, reports a field study on the social media platform X/Twitter. The chapter discusses how identity and status affect a counterspeaker's impact on average and radical transgressors and bystanders. I responded to users who had used a racial slur on Twitter and highlighted the harmful impact of their posts. As the experimental manipulation, I varied whether an account responded who had the transgressor's or the victimized group's ethnicity and either many or few followers. I discuss the effect of different counterspeakers on transgressor and bystander behavior. Moreover, I discuss how variations in the accounts' effectiveness indicate underlying psychological mechanisms driving counterspeech effectiveness with a focus on perceptions of ingroup norms and hate speech severity.

In the general discussion, I summarize the findings of my empirical chapters and assess the cumulative evidence for my two research questions, whether and how counterspeech affects bystanders and transgressors. I connect the evidence from my experiments in controlled settings to the real-world findings from my social media study. Based on this synopsis, I discuss congruencies across the different chapters as well as the crucial dimensions on which my findings vary and invite further research into the boundary conditions of counterspeech effectiveness. Finally, I end my dissertation with practical recommendations, informed by my findings, for those who wish to individually combat online hate speech.

Zusammenfassung

Die vorliegende Dissertation untersucht, ob und wie Gegenrede zur Bekämpfung von Online-Hassrede eingesetzt werden kann. Zur Untersuchung der Forschungsfragen wurde ein multimethodaler Ansatz gewählt – eine Längsschnittstudie in einem simulierten Social-Media-Forum, vignettenbasierte Querschnittsstudien und eine Feldstudie in einem realen Social-Media-Forum. Dieser Ansatz ermöglichte es, in kontrollierten Experimenten den Effekt von Gegenrede auf nachfolgendes Verhalten sowie mögliche Mediatoren zu isolieren und in einer realen Social-Media-Umgebung deren Relevanz zu bestätigen. Meine Forschung zeigte, dass Gegenrede weitere Menschen inspirieren sowie Hassrede bei Tätern und Umstehenden reduzieren kann. Über verschiedene Studien hinweg beeinflusste Gegenrede kollektive und gruppenspezifische soziale Normen und wirkte sich darauf aus, wie die Schwere von Hassrede wahrgenommen wurde.

In Kapitel A gehe ich auf Kontextfaktoren von Online-Umgebungen ein, die Zivilcourage als Reaktion auf Online-Hassrede behindern oder erleichtern können.

Kapitel B behandelt ein Längsschnitts-Experiment in einem interaktiven simulierten Social-Media-Forum. Die Studie untersuchte die direkte Wirkung von Gegenrede auf weitere Gegenrede durch Umstehende und mögliche Mediatoren dieses Effekts. Über einen Zeitraum von zwei Wochen interagierten Teilnehmende wiederholt mit einem Forum, das Hassrede-Kommentare von anderen vermeintlichen Nutzern und Nutzerinnen enthielt. Es wurde experimentell manipuliert, ob auch Gegenrede oder ausschließlich neutrale Antworten auf die Hassrede zu sehen waren. Ich berichte die Wirkung von Gegenrede auf die Absicht, selber Gegenrede zu veröffentlichen, sowie auf tatsächliches Verhalten. Darüber hinaus stelle ich Mechanismen vor, die die Wirkung von Gegenrede auf Umstehende mediieren, mit einem Fokus auf die Wahrnehmung von Pro-Gegenrede-Normen und der Schwere von Hassrede. Schließlich erörtere ich die zeitliche Wirkung von Gegenrede auf die Mediatoren und das Verhalten der Teilnehmenden.

In Kapitel C beschreibe ich drei Querschnitts-Experimente, in denen die Wirkung von Gegenrede auf weitere Gegenrede durch Umstehende und gruppeninterne Normen als potenzielle Mediatoren untersucht wurde. In den Experimenten sahen die Teilnehmenden verschiedene Nachrichtenartikel auf sozialen Medien und Kommentare von anderen vermeintlichen Teilnehmenden. Es wurde hierbei experimentell manipuliert, ob die Teilnehmenden auch Gegenrede von einem Mitglied der Eigengruppe oder einer anderen Gruppe sahen. Ich berichte meine Ergebnisse zur Wirkung von Gegenrede auf die Wahrnehmung von Pro-Gegenrede-Normen der Eigengruppe und wie diese Normen die Befürwortung von Gegenrede und tatsächliches Verhalten beeinflussen. Außerdem gehe ich auf die Gesamtwirkung von Gegenrede ein.

In Kapitel D präsentiere ich eine Feldstudie auf der Social-Media-Plattform X/Twitter. Es wurde untersucht, wie Identität und Status eines Gegenredners dessen Wirkung auf durchschnittliche und radikale Täter und Umstehende beeinflussen. Hierfür kontaktierte ich Nutzer, die auf Twitter eine rassistische Beleidigung verwendet hatten, und wies sie auf die schädliche Wirkung ihrer Beiträge hin. Ich variierte, ob ich einen Account zur Gegenrede verwendete, der dieselbe Ethnie wie der Täter oder die der betroffene Gruppe hatte und

der entweder viele oder wenige Follower besaß. Ich berichte die Wirkung der verschiedenen Gegenredner auf Täter und Umstehende. Darüber hinaus diskutiere ich, wie beobachtete Variationen auf zugrundeliegende psychologische Mediatoren hinweisen können, wobei ich mich auf die Wahrnehmung von Eigengruppen-Normen und der Schwere von Hassrede fokussiere.

In der allgemeinen Diskussion fasse ich die Ergebnisse meiner empirischen Kapitel zusammen und bewerte die kumulative Evidenz für meine beiden Forschungsfragen, ob und wie Gegenrede auf Umstehende und Täter wirkt. Ich verbinde die Erkenntnisse aus meinen Experimenten in kontrollierten Umgebungen mit den Ergebnissen aus meiner Studie in einem realen Social-Media-Forum. Auf der Grundlage dieser Synthese erörtere ich die Übereinstimmungen zwischen den verschiedenen Kapiteln sowie die Dimensionen, in denen sich meine Ergebnisse unterscheiden, und lade zur weiteren Erforschung der Randbedingungen für die Wirksamkeit von Gegenrede ein. Schließlich beende ich meine Dissertation mit praktischen Empfehlungen für diejenigen, die individuell gegen Hassrede im Internet vorgehen wollen.

Table of Contents

General Introduction

The Problem: Online Hate Speech

An ever-growing tide of online hate speech threatens societies across the world. Defined by the United Nations (2020) as communication that attacks individuals based on central identity characteristics such as race, religion, gender, or other identity factors, its prevalence in online discourse can hardly be understated. Half of the People of Color responding to a recent representative survey in the United States (ADL, 2023) indicated that they had received ethnicity-based online harassment in the preceding year - a steep rise from 28% just two years earlier (ADL, 2021). Additionally, one-third of Jewish respondents reported being targeted by hate speech (ADL, 2023). In a similar German survey, a staggering 76% of respondents had seen hate speech online (Forsa, 2023), matching the high prevalence in other European countries such as Finland (79%), France (65%), or Spain (75%) (Reichelmann et al., 2021). Roughly one in four German respondents had seen hate speech against Muslims and 43% reported having seen hate speech against refugees (Forsa, 2023). Online hate speech is a worldwide issue, extending from Slovakia (Miškolci et al., 2018) and Ethiopia (Chekol et al., 2023) to India (Sharma & Kaushal, 2023) and communities around the globe (United Nations, 2020).

Hate speech can inflict severe harm on its victims, reaching from immediate emotional distress to harmful long-term effects and even physical victimization (Leets, 2002; Mullen & Smyth, 2004; Müller & Schwarz, 2021). People regularly experience strong negative emotions when they are targeted by hate speech (Leets, 2002), which, over time, may evolve into increased levels of anxiety and depression (Keighley, 2022; Tynes et al., 2008; Wypych & Bilewicz, 2024) and decreased levels of life satisfaction (Keipi et al., 2018). The negativity of slurs against different immigrant groups has even been correlated with higher suicide rates in those groups (Mullen & Smyth, 2004). Beyond causing psychological suffering, hate speech can lead to withdrawal, silencing its targets and effectively pushing them out of the discourse (Keighley, 2022; Nadim & Fladmoe, 2021; Urbaniak et al., 2022). Victims also report increased feelings of insecurity in offline contexts (Dreißigacker et al., 2024), which is reasonable, as research shows that hate speech can inspire offline violence. For instance, a German study linked anti-refugee posts by the right-wing party Alternative für Deutschland (AfD) to a surge in attacks against refugees (Müller & Schwarz, 2021).

Moreover, the harmful effects of hate speech extend beyond its direct victims to a wider audience, desensitizing bystanders and deteriorating discourse norms (Bilewicz & Soral, 2020).

When individuals see others repeatedly engage in hateful language against a marginalized group, they consider such language less damaging (Schmid et al., 2022; Soral et al., 2018). This effect can even be observed for physical outcomes such as decreased heart rates in response to hateful language (Soral et al., 2023) and less neural activation of brain areas associated with responses to other people's suffering (Pluta et al., 2023). This

desensitization to hate speech, in turn, leads to more negative attitudes towards its victims (Soral et al., 2018).

Hate speech also negatively biases conversations and dissolves discursive norms (Alvarez-Benjumea, 2022; Bilewicz & Soral, 2020; Garland et al., 2020; Soral et al., 2020). Hate speech can substantially alter which derogatory comments are perceived as acceptable (Bilewicz & Soral, 2020; Forscher et al., 2015; Schmid et al., 2022; Soral et al., 2020), corroding established anti-hate speech sentiments (Alvarez-Benjumea, 2022; Crandall et al., 2002; Zitek & Hebl, 2007). Right-wing actors have been shown to consciously leverage this effect. For example, German right-wing demagogues use the term "to break open a discourse" to describe the insertion and subsequent normalization of hateful language (Ministerium für Inneres und Sport des Landes Sachsen-Anhalt, 2021). Seeing others derogate an outgroup can inspire bystanders to also actively voice prejudice (Alvarez-Benjumea, 2022; Bilewicz & Soral, 2020; Forscher et al., 2015; Hsueh et al., 2015). For example, people were more likely to compose negative comments against marginalized groups such as sexual minorities or refugees after seeing others post hate speech against them (Alvarez-Benjumea & Winter, 2018).

Online Contexts Enhance Hate Speech Effects

Several key characteristics of online environments amplify the prevalence and particularly harmful effects of online hate speech.

Firstly, it is easier for the average person to reach broad audiences online than offline. Users can widely disseminate their comments without encountering gatekeepers associated with traditional mass media (Bor & Petersen, 2021; Brady et al., 2019; Delgado & Stefancic, 2014; Obermaier et al., 2015; Ziegele et al., 2020). Offline, if a person wanted to disseminate hateful views to a wide audience, they would have to, for example, convince a newspaper editor or a TV station host to give them a platform. These gatekeepers would likely prevent the spread of hate speech due to ethical objections, legal concerns, or editorial policies. Online, users encounter many opportunities to publish hate speech without such editorial control, which can enable them to rapidly reach wide audiences (Mathew, Dutt, et al., 2019). Radical users exploit this feature extensively (Garland et al., 2022; Lopez-Sanchez & Müller, 2021).

Moreover, the internet enables hateful individuals to connect with others and spread hate speech in an organized fashion (Amichai-Hamburger, 2017; Evkoski et al., 2022; Garland et al., 2020, 2022; Goel et al., 2023; Johnson et al., 2019; Sunstein, 2007). Hate speakers do not need to act alone, but they can support each other and amplify each other's content (Goel et al., 2023). Such organized efforts were able to effectively hijack German discourse on Twitter in the wake of the 2017 federal elections (Garland et al., 2022).

The permanence of online content further exacerbates the impact of online hate speech (Barberá et al., 2015; Citron, 2014; Dillon & Bushman, 2015; Obermaier et al., 2015). For example, if a person decided to voice disgust for sexual minorities offline on a bus, only the other passengers around them would hear their hate speech. If that person wrote a similar statement on social media, their post would remain on the website, potentially affecting victims and bystanders indefinitely.

Lastly, the common anonymity of online contexts can further encourage hate speech (A. Brown, 2018; Citron, 2014; Obermaier et al., 2015; Postmes & Turner, 2015; Suler, 2004; Ziegele et al., 2020). Protected by it, users can compose vitriolic comments without the immediate threat of social and legal consequences (A. Brown, 2018; Citron, 2014). Not seeing the victims' reactions to hate speech further decreases the likelihood of feeling empathy with them (A. Brown, 2018; Suler, 2004), which could otherwise dissuade further transgressions (Chaney & Sanchez, 2018; Citron, 2014).

Taken together, these mechanisms make online hate speech a prevalent and pernicious threat, creating an urgent need to develop and apply effective countermeasures.

A Possible Countermeasure: Counterspeech

Problems With Deletion

At present, efforts by social media providers and state actors to combat online hate speech focus primarily on deletion (e.g., Council of the European Union, 2008; Meta, 2024; TikTok, 2022). However, deletion-only approaches suffer from multiple shortcomings.

First, deletion is not able to keep up with the ever-growing amount of online hate speech (FRA, 2023). Given its immense volumes, platforms cannot muster the necessary numbers of content moderators to effectively detect and remove hate speech manually and automated detection algorithms still yield an unsatisfying number of misses and false detections (Saleh et al., 2023). Automated deletion-based approaches face a dilemma: they can either reduce the number of misses at the expense of accidentally deleting posts that do not constitute hate speech, or they can delete fewer posts, prioritizing free speech at the expense of allowing more hate speech to persist. This dilemma is exacerbated by an ever-changing use of chiffres and so-called algo-speak to avoid detection of undesirable rhetoric (Chancellor et al., 2016; Saleh et al., 2023). For example, instead of explicitly calling for the murder of Jews, an anti-Semite might post something like "*(((They))) should be unalived".* The triple brackets are often used to denote Jews and the neologism *to unalive* was invented to avoid using the often-censored term *to kill* (Steen et al., 2023).

Second, deletion-based approaches raise concerns over free speech infringement (e.g., Lepoutre, 2017; Strossen, 2020). Free-speech purists often reference the Brandeis doctrine, which posits that "the remedy to be applied is more speech, not enforced silence" (*Whitney v. California*, 1927). Even those who acknowledge the potential necessity for hate speech laws often criticize the prevalent censorship by social media entities as illegitimate (Schaake, 2020).

Third, deletion may merely shift the place where users share hate speech (A. Brown, 2018; Jiménez Durán, 2022). Deletion does not address the root causes of hate speech proliferation, enticing users to move to less restrictive platforms instead of posting less hate speech (Mathew, Illendula, et al., 2019; Saleh et al., 2023).

Counterspeech as a Complement

Counterspeech could serve as a complement to deletion-only approaches that can mitigate some of their shortcomings (Bilewicz et al., 2021; Cepollaro et al., 2023; Cypris et al., 2022; Lepoutre, 2017). This dissertation defines counterspeech as communication that openly

confronts and rejects hate speech. The first advantage of counterspeech is that it can be applied more liberally than deletion. Unlike deletion, which risks illegitimate censorship when applied to misclassified comments, counterspeech maintains the original comment and enables others to judge whether a counterspeech response was appropriate or not. Second, counterspeech constitutes "more speech" instead of "enforced silence," as mentioned in the Brandeis Doctrine (*Whitney v. California*, 1927). Even in societies that legally require the deletion of certain types of hate speech (e.g., §130 German Criminal Code), counterspeech can address comments that are harmful but remain below the legal threshold. Third, as I will discuss in length below, counterspeech can potentially address the underlying causes and negative effects of online hate speech, influencing the future behavior of both transgressors and bystanders rather than simply removing offensive content.

To comprehensively assess counterspeech utility, it is necessary determine its effectiveness for three groups: hate speech victims, transgressors, and bystanders.

Encouragingly, research shows that counterspeech positively affects hate speech victims (Leets, 2002). Seeing others speak up on their behalf can offer support and reassurance to hate speech targets (Leets, 2002) and publicly reinforce their dignity (Cepollaro et al., 2023; Lepoutre, 2017). Counterspeech from members of the transgressor's group can also counteract the polarizing effects of hate speech by reducing negative reactions and perceptions of hate speech targets towards that group (Obermaier et al., 2021; Van Houtven et al., 2024).

Unfortunately, the effects of counterspeech are not sufficiently identified for bystanders and transgressors (Cepollaro et al., 2023; Rudnicki et al., 2023; Windisch et al., 2022). The limited research that exists, moreover, has yielded conflicting findings (e.g., Alvarez-Benjumea & Winter, 2018; Hangartner et al., 2021; Leonhard et al., 2018; Miškolci et al., 2018). However, reaching transgressors and bystanders is crucial for a sustainable reduction of online hate speech (Bilewicz & Soral, 2020; Garland et al., 2022). Therefore, my dissertation addresses the question:

RQ1: Does counterspeech have a positive impact on transgressors and further bystanders?

Impact on Transgressors

Concerning transgressors, "*Don't feed the troll"* is a prevalent adage on the internet (Hulk, 2018). Its underlying assumption is that engaging with hate speech is futile, based on the belief that transgressors are motivated by antisocial desires and thrive on conflict (Coles & Lane, 2023; March, 2019; March & Steele, 2020; Miškolci et al., 2018; Moor & Anderson, 2019). Indeed, research finds that initial responses to counterspeech often result in continued hate speech rather than an acknowledgment of wrongdoing (Garland et al., 2022; Miškolci et al., 2018). After receiving counterspeech, Facebook users who had posted anti-Romani hate speech were twenty times more likely to continue posting hate speech than to acknowledge their transgression (Miškolci et al., 2018). However, not all individuals who engage in hate speech should be considered *trolls.* Some may commit a mistake in the heat of the moment and not hold deeply ingrained hateful convictions (Cheng et al., 2017). Correspondingly, looking beyond immediate reactions, counterspeech can decrease

subsequent offenses in the long run (Bilewicz et al., 2021; Hangartner et al., 2021; Munger, 2017; Siegel & Badaan, 2020). For instance, Reddit users who received counterspeech engaged in less verbal aggression in the following 60 days compared to users who were not confronted (Bilewicz et al., 2021). On Twitter, transgressors were discouraged from further xenophobic hate speech for four weeks after receiving counterspeech that stressed the negative consequences for their victims (Hangartner et al., 2021). Also, on Twitter, Sunni Muslims were discouraged from further sectarian hate speech for a month after receiving counterspeech quoting religious authorities (Siegel & Badaan, 2020). These findings demonstrate the potential for counterspeech to sustainably alter transgressor behavior.

However, prior studies applied very specific settings, potentially overlooking dynamics present in the majority of online interactions. Most research observing counterspeech effects investigated anonymous counterspeakers (Bilewicz et al., 2021; Hangartner et al., 2021; Siegel & Badaan, 2020). In contrast, user profiles generally contain at least some identifying characteristics (Norberg et al., 2007; Nosko et al., 2010; Quinn et al., 2019). Therefore, findings for anonymous users are not representative of the average interaction. Once individuating characteristics are considered, the impact of one and the same counterspeech comment can vary significantly (Munger, 2017). For instance, empathy-based counterspeech had an overall positive effect for anonymous counterspeakers, but when counterspeaker status and ethnicity were visible, only high-status users of the transgressor's ethnicity effectively decreased subsequent hate speech (Hangartner et al., 2021; Munger, 2017). However, the research that found these identity effects only considered a pre-selected sample of the most offensive Twitter users (Munger, 2017). Consequently, it remains unclear how counterspeech affects transgressors in the most common scenario: An average and somewhat identifiable user speaks out against hate speech by an average transgressor.

In Chapter D, I will investigate how the intervener's identity can modulate counterspeech influence on average and extreme transgressors in real-life Twitter interactions. I will test whether members of the victimized group or those of the transgressor's group can have a positive impact and how their influence varies based on the counterspeaker's status.

Impact on Bystanders

Bystanders constitute another important target group for counterspeech, as they can play a crucial role in amplifying hate as well as counterspeech. As I discussed above, the impact of hate speech can substantially increase when further bystanders join in (Bilewicz & Soral, 2020). In contrast, even a handful of bystanders supporting an initial counterspeech comment can significantly shift audience attitudes in favor of counterspeech (Schieb & Preuss, 2016, 2018). Further underscoring this effect, organized counterspeech on Twitter has been linked to tangible reductions in the overall hostility of online conversations (Garland et al., 2020).

However, few studies specifically looked at the effect of counterspeech on further bystanders and the ones that did yielded mixed results (Cary et al., 2020; Leonhard et al., 2018; Miškolci et al., 2018). Studies that investigated real-world settings, either experimentally (Miškolci et al., 2018) or via surveys (Cary et al., 2020), found that

counterspeech can increase subsequent bystander counterspeech. For example, bystanders posted substantially more pro-Romani comments on Facebook when researchers had posted counterspeech compared to unresponded anti-Romani hate speech (Miškolci et al., 2018). Furthermore, online gamers reported a higher likelihood of posting counterspeech if their friends frequently did so as well (Cary et al., 2020). However, seeing counterspeech, rather than neutral responses, in a controlled experimental setting did not increase bystanders' intentions to speak out themselves against anti-refugee hate speech (Leonhard et al., 2018).

Similarly, the impact of counterspeech on bystander hate speech is mixed. Miškolci and colleagues (2018) found that counterspeech effectively decreased anti-Romani comments among bystanders. However, research in a controlled experimental setting once more failed to replicate counterspeech effects for bystanders (Alvarez-Benjumea & Winter, 2018). However, in this case, null findings could also have been caused by a small sample size, averaging only 45 participants per condition.

While Miškolci and colleagues (2018) found evidence that counterspeech can inspire other bystanders to speak up and suppress bystander hate on Facebook, their investigation in a dynamic social media environment leaves open the possibility of a multiplicity of confounding factors. For instance, it is plausible that counterspeech could merely enhance the hate speech comment's algorithmically determined visibility instead of inspiring other bystanders. A larger audience is more likely to contain individuals who are intrinsically motivated to speak up (Buerger, 2021; Sasse et al., 2023). Experiments that employed controlled settings and found no effect of prior counterspeech can rule out such alternative explanations (Alvarez-Benjumea & Winter, 2018; Leonhard et al., 2018). However, their cross-sectional designs might overlook important motivators for counterspeech, such as reputational concerns (Van Bommel et al., 2012; Ziegele et al., 2020), which are greatly decreased in anonymous one-shot interactions. In addition, deviations from real-life interactions such as focusing on hypothetical decisions (Leonhard et al., 2018) or forcing participants to write an answer (Alvarez-Benjumea & Winter, 2018) can further limit the applicability of these findings to real-life online behavior (Banyard & Moynihan, 2011; Baumert et al., 2013; A. L. Brown et al., 2014; Crosby & Wilson, 2015; Goodwin et al., 2020; Kawakami et al., 2009; Latané & Darley, 1970).

In addition to these mixed findings for overall counterspeech effects on bystanders, it remains unclear how a counterspeaker's identity modulates their influence. As discussed above, the impact on transgressors varies greatly depending on the counterspeaker's group affiliation and status (Munger, 2017). Bystanders could be similarly affected. The related field of cyberbullying also underscores the importance of counterspeaker identity, finding that peer approval strongly determines whether students engage in harassment or defend the victim (Bastiaensens et al., 2016; Espelage et al., 2012). Taken together, these results suggest that understanding the influence of the commenters' social identity is crucial to accurately assess counterspeech effects on further bystanders.

To address these issues, I will conduct a comprehensive investigation of counterspeech effects on bystanders, applying both controlled cross-sectional and longitudinal experiments

as well as a field experiment on social media. In Chapter B, I will test in a mock social media forum across two weeks whether counterspeech, in general, inspires bystanders to speak up. In Chapter C, I will assess counterspeech effects overall and how they are modulated by the counterspeaker's social identity in three experiments employing controlled cross-sectional designs. Finally, in Chapter D, I will investigate how counterspeech, modulated by the counterspeaker identity, affects subsequent bystander hate and counterspeech in real-life interactions on Twitter.

Mechanisms

This dissertation not only investigates *whether* counterspeech influences transgressors and bystanders but also *how* it impacts them. Counterspeech effectiveness can vary greatly depending on what is said (Hangartner et al., 2021; Siegel & Badaan, 2020) by whom (Munger, 2017). As mentioned above, similar counterspeech messages had wildly different results depending on the counterspeaker's identity (Hangartner et al., 2021; Munger, 2017). To effectively leverage counterspeech against online hate speech, it is, therefore, necessary to acquire a deeper understanding of the underlying mechanisms through which it influences transgressors and bystanders. Consequently, this dissertation addresses the second overarching question:

RQ2: Which mechanisms mediate counterspeech effects on transgressors and bystanders?

In this dissertation, I focus on two pivotal mechanisms suggested by social psychological theory - social norms (Cialdini et al., 1990) and perceptions of hate speech severity rooted in the bystander intervention model (Latané & Darley, 1970). I initially concentrated on social norms, but as my research progressed, severity perceptions emerged as an additional crucial mechanism. Notably, norm deterioration and desensitization to its harms have been identified as key processes through which hate speech exerts its adverse effects (Alvarez-Benjumea, 2022; Bilewicz & Soral, 2020; Soral et al., 2018). Counterspeech could be an effective antidote, directly targeting the processes that facilitate the destructive effects of hate speech.

Social Norms

Counterspeech could influence bystanders and transgressors through its effect on social norm perceptions. Social norms can be defined as the attitudes and behaviors that are prevalent and generally endorsed in relevant settings or by relevant groups (Cialdini et al., 1990; Tankard & Paluck, 2016). Social norms drive online behaviors in a wide range of issues, such as trolling in online chat rooms (Seering et al., 2017), cyberbullying and harassment (Henson et al., 2020), discussion toxicity (Matias, 2019), purchase decisions (R. Li et al., 2017), and attitudes about gay marriage (Tankard & Paluck, 2017). In the context of online hate and counterspeech, the potential of social norms as a driving mechanism has been suggested but remains underexplored (Bilewicz & Soral, 2020; Rudnicki et al., 2023).

They can exert influence as *collective* norms, representing what is normative in the broader context of an individual's society, school, or online conversation space (Paluck & Shepherd, 2012). Moreover, social norms can drive behaviors as *ingroup* norms by defining what is normative for more specific groups with which people identify (Klein et al., 2007; Turner et al., 1987). For instance, social norms could influence a male chat forum user as collective

norms by defining what is endorsed and normative for all users in the chat forum, as well as ingroup norms, by suggesting how men should act in particular.

Online environments often amplify the influence of social norms through a reduced visibility of individuating characteristics. Contrary to popular belief, being less individually identifiable leads to an increase, rather than a decrease, in the influence of social norms (Postmes et al., 2001; Reicher et al., 1995; Spears & Postmes, 2015). When people become less individually identifiable, they perceive themselves more strongly in terms of salient collective identities and act more in accordance with their norms (Reicher et al., 1995).

Individuals can infer such social norms by observing other people's conduct (Klein et al., 2007; Paluck & Shepherd, 2012; Tankard & Paluck, 2016) or by paying attention to other's overt endorsement or rejection of behaviors (Hogg & Rinella, 2018; Matias, 2019). Seeing others engage in hate speech positively correlates with perceiving such behaviors as normative (Wachs et al., 2021). Conversely, individuals who hear others reject prejudice consider it less admissible themselves (Bennett & Sekaquaptewa, 2014; Blanchard et al., 1994). The permanence of online communication facilitates norm learning as it enables users to infer social norms based on previous communication and modulate their own comments accordingly. This dynamic could be observed on Reddit, where newcomers adjusted their contributions to match the existing tone of discussions (Rajadesingan et al., 2020).

Collective Norms

On a broad level, counterspeech can signal the prevailing collective norms within a given conversation context, from social media platforms to specific forums and chats. Collective norm impact in online contexts is indicated by users regularly imitating the behaviors of others (Goldenberg & Gross, 2020; Mikal et al., 2014; Seering et al., 2017). Collective norms have been shown to override individual dispositions. On Reddit, the toxicity of a user's comment was more strongly determined by the prevalent norms of the conversation forum in which they posted than by their individual dispositions (Rajadesingan et al., 2020). Such norm influence can be a double-edged sword: Users are more inclined to post civil comments after observing similar behavior from others (Han et al., 2018; Han & Brazeal, 2015; Molina & Jennings, 2018; Ng et al., 2022; Seering et al., 2017), yet the reverse is also true for disruptive and uncivil behavior (Cheng et al., 2017; Seering et al., 2017). Prior hate speech inspired similar comments in adolescents (Wachs & Wright, 2018) and adults (Alvarez-Benjumea, 2022; Soral et al., 2020). The impact of collective social norms was notably exemplified following Elon Musk's acquisition of Twitter (now X) in 2022. Musk's signaling towards a more permissive stance on hate speech went along with a quadrupling of such content on the platform in the wake of the take-over (Hickey et al., 2023). Since Twitter's content moderation mechanisms were still in place then, this surge must rather be attributed to emboldened transgressors than to decreased vigilance on the platform's side.

Seeing others stand up to online hate speech could, therefore, shape bystander and transgressor perceptions of how common and endorsed counterspeech is in a given context. This could encourage bystanders to engage in counterspeech themselves. I will investigate this mechanism in Chapter B by measuring how seeing other users speak up against online

hate speech affects the perceived normativity of counterspeech in a mock social media forum over two weeks.

Ingroup Norms

In addition to collective norms, counterspeech can also affect ingroup norms (Klein et al., 2007; Tankard & Paluck, 2016; Turner et al., 1987). For example, if a man speaks out against sexism in a discussion forum, his counterspeech may not only convey that sexism is seen as illegitimate by other forum members but also by other men independent of the context. This may particularly influence men who strongly care about their gender identity.

According to the Social Identity Approach, people's self-image and self-worth are influenced by their social group affiliations (Tajfel, 1974; Tajfel & Turner, 1986). Moreover, according to Self-Categorization Theory, if a valued ingroup is salient, people are more likely to see themselves in terms of the group prototype and to adhere to its perceived norms (Turner et al., 1987). This conformity is further enhanced when individuating markers are scarce for reasons listed earlier (Lee, 2007b; Postmes et al., 2001; Reicher et al., 1995). In addition to this so-called cognitive route, in-group norms can influence an individual's behavior because they pursue strategic goals by conforming to ingroup norms. Klein and colleagues (2007) proposed that people can publicly conform to ingroup norms for two strategic reasons – mobilization and consolidation. People can engage in normative behavior to mobilize in- and outgroup members, attempting to inspire them to follow suit. In our case, if a person considers counterspeech normative for their ingroup, they might want to engage in it to inspire others to speak up as well (Coles & Lane, 2023). Ingroup norms can moreover be performed to consolidate one's identity. People can act in accordance with ingroup norms either to signal their belonging to the group to other ingroup members (Noel et al., 1995) or to signal to outgroup members that their group endorses these norms (Reicher & Levine, 1994). For instance, a study in the United Kingdom found that both the desire to achieve political change and the desire to show that one's ingroup opposes injustice can independently motivate people to join demonstrations (Saab et al., 2015).

Ingroup norms are potentially more powerful than collective forum or platform norms. Individuals are more likely to be influenced by fellow ingroup members than outgroup members (Hogg et al., 1990; Hogg & Turner, 1987; Lee, 2007a; Wilder, 1990). For instance, men are more likely swayed by another man speaking up against sexism than by a woman (Drury, 2013). Moreover, unlike the somewhat isolated influence of collective forum norms, which do not readily translate into other settings (Rajadesingan et al., 2020), ingroup norms can influence people across a wide array of different contexts (Morris et al., 2015).

Consequently, researchers have suggested ingroup norms to explain variations in counterspeech effectiveness (Munger, 2017; Siegel & Badaan, 2020). A study on Twitter showed that counterspeech from a user with a large following who shared the transgressor's white ethnicity effectively reduced racial slurs, while similar efforts from a member of the victimized group or users with fewer followers remained unsuccessful (Munger, 2017). The authors suggested that the ingroup counterspeaker shaped transgressor behavior by signaling how white people should act. Similarly, sectarian hate among Lebanese Twitter users was only reduced by counterspeech stressing disapproval by

religious authorities and endorsing a shared Muslim identity (Siegel & Badaan, 2020). Conversely, counterspeech had no impact when it failed to mention religious authorities or focused on national instead of religious unity. A second experiment, moreover, found similar results for bystanders. Survey respondents more strongly rejected sectarian hate speech and endorsed counterspeech when they saw an elite statement stressing a shared religious identity than non-elite statements or focusing on national unity (Siegel & Badaan, 2020).

However, the research mentioned above merely speculated about the impact of ingroup norms instead of directly measuring it. Moreover, both studies considered extreme transgressor samples who had either posted an abnormally high volume of offensive tweets (Munger, 2017) or religious hate in the past (Siegel & Badaan, 2020). The hyper-polarized Lebanese context - a country that experienced sectarian-based ethnic cleansing in living memory and that is highly segregated by religion (Majed, 2021) - also potentially impedes generalization to less polarized contexts. It remains unclear whether the reported findings replicate for the average transgressor or bystander, who might identify less strongly with their respective ingroups and, therefore, be less influenced by their norms.

In my dissertation, I will employ a controlled experimental setting in Chapter C to directly measure the effect of ingroup counterspeech on pro-counterspeech ingroup norm perceptions and their effect on subsequent bystander counterspeech. I will, moreover, test their effect on real-world interactions for a broad transgressor and bystander sample in Chapter D.

Severity

Counterspeech can also potentially increase perceptions of hate speech severity. This, in turn, can increase the likelihood that bystanders engage in counterspeech and decrease subsequent hate speech by transgressors (Chaney & Sanchez, 2018; Latané & Darley, 1970).

According to the bystander intervention model, for bystanders to take action, they need to categorize a situation as an emergency requiring intervention (Latané & Darley, 1970). Thus, whether individuals evaluate hate speech as a severe transgression or as a rather trivial misstep plays a pivotal role in determining their willingness to intervene. After repeated exposure, people start seeing online hate speech as less harmful, making it seem less of an emergency (Bilewicz & Soral, 2020; Schmid et al., 2022; Wachs et al., 2022). Counterspeech may remind its audience that hate speech is, in fact, a severe transgression. For example, when bystanders saw counterspeech calling out anti-Asian posts as racist, they viewed these posts as more offensive than people who saw more ambiguous counterspeech (Meyers et al., 2020). If bystanders consider hate and harassment a severe transgression, they are more inclined to speak out against it (Bastiaensens et al., 2014; Koehler & Weber, 2018; Leonhard et al., 2018; Rudnicki et al., 2023). Although the effect of counterspeech on severity perceptions and the effect of severity perceptions on bystander behavior are both independently plausible, they have not been investigated together to probe severity perceptions as a potential mechanism through which counterspeech influences bystanders.

In addition, counterspeech could impact transgressors through its effect on severity evaluations. University students who were confronted about their prejudiced behaviors

reported negative self-directed affect and feelings of guilt, leading to a reduction in subsequently reported prejudice (Chaney & Sanchez, 2018; Czopp et al., 2006). In a similar vein, Twitter users posted xenophobic tweets less frequently after seeing empathy-based counterspeech reminding them of the harmful consequences for their targets (Hangartner et al., 2021). However, this effect can fluctuate depending on the counterspeaker's identity (Munger, 2017). That said, the average transgressor may be more receptive to empathy-based counterspeech than offensive ones assessed by Munger (2017), possibly enhancing its effect on other counterspeakers as well (Crosby & Monin, 2013; Stone, 2011). Thus, it remains unclear how empathy-based counterspeech from different interveners affects the average transgressor.

In an experimental setting described in Chapter B, I will directly assess the impact of counterspeech on bystanders' hate speech severity perceptions and the effect of these perceptions on subsequent counterspeech. In addition, I will test the effect of empathy-based counterspeech by different counterspeakers in Chapter D.

Summary

In summary, although counterspeech has the potential to be a powerful antidote against online hate speech, its effectiveness remains empirically inconclusive. Prior research yielded mixed findings for its impact on further bystanders. It also remains unclear how the average transgressor is affected by an average intervener. Moreover, the mechanisms through which counterspeech exerts its effect remain underexplored. It seems plausible that counterspeech leverages social norms and hate speech severity perceptions to influence bystanders and transgressors. However, direct empirical evidence is needed to confirm such dynamics.

In my dissertation, I will, therefore, address two overarching questions. On the one hand, I will investigate whether counterspeech against online hate speech has a positive impact on bystanders and transgressors. On the other hand, I will examine mechanisms through which counterspeech influences its audience. Specifically, I will investigate whether counterspeech positively affects hate speech severity perceptions and pro-counterspeech forum and ingroup norms.

To determine causality and assure external validity, I will conduct a multi-method investigation across controlled settings and a field study. In the controlled settings of Chapters B and C, I test direct causal relationships between counterspeech and subsequent bystander behaviors as well as the mechanisms that may drive counterspeech effects. In Chapter D, I aim to confirm my findings in real-world social media interactions on Twitter while also investigating outcomes for transgressors. To assure the generalizability of my findings and uncover context-independent mechanisms, I will conduct my investigation across multiple target, intervener, and transgressor groups.

Chapter Overview

Chapter A

As discussed above, virtual settings can exacerbate the adverse effects of online hate speech. However, not only hate but also counterspeech impact can be modulated by online environments. In Chapter A, I review contextual factors of online settings that may impede or facilitate morally courageous behavior in response to online hate speech.

Chapter B

Chapter B reports a longitudinal experiment in an interactive mock social media forum. The study assesses the direct impact of counterspeech on subsequent bystander counterspeech and potential mediators of its effect. Over the course of two weeks, participants interacted multiple times with a mock social media forum that contained hate speech comments by other ostensible users. I manipulated whether participants saw counterspeech or exclusively neutral replies to the hate speech. I report results regarding the participants' intentions to speak up themselves and actual behavior. In addition, I present mechanisms that mediate counterspeech impact on bystanders, focusing on perceptions of pro-counterspeech norms and hate speech severity. Finally, I discuss the longitudinal effects of counterspeech on the mediators and behavioral outcomes.

Chapter C

In Chapter C, I cover three cross-sectional experiments investigating the effect of counterspeech on further bystander counterspeech via ingroup norms as a potential mediating mechanism. In the experiments, I showed participants different social media post vignettes and comments by other ostensible participants, some of which were hate speech. I varied whether participants also saw counterspeech by an ingroup or an outgroup member. I present my findings regarding counterspeech impact on the participants' perception of pro-counterspeech ingroup norms and how these norms affect counterspeech endorsement and actual counterspeech. Moreover, I present the overall effects of counterspeech.

Chapter D

Chapter D, reports a field study on the social media platform Twitter. The chapter discusses how identity and status affect a counterspeaker's impact on average and radical transgressors and bystanders. I responded to users who had used a racial slur on Twitter and highlighted the harmful impact of their posts. As the experimental manipulation, I varied whether an account responded who had the transgressor's or the victimized group's ethnicity and either many or few followers. I discuss the effect of different counterspeakers on transgressor and bystander behavior. Moreover, I discuss how variations in the accounts' effectiveness indicate underlying psychological mechanisms driving counterspeech effectiveness with a focus on perceptions of ingroup norms and hate speech severity.

General Discussion

In the general discussion, I summarize the findings of my empirical chapters and assess the cumulative evidence for my two research questions, *whether* and *how* counterspeech affects bystanders and transgressors. I connect the evidence from my experiments in

controlled settings to the real-world findings from my social media study. Based on this synopsis, I discuss congruencies across the different chapters as well as the crucial dimensions on which my findings vary and invite further research into the boundary conditions of counterspeech effectiveness. Finally, I end my dissertation with practical recommendations, informed by my findings, for those who wish to individually combat online hate speech.

Chapter A – Contextual Determinants of Online Moral Courage

This chapter is based on:

Sasse*, J., Cypris*, N. F., & Baumert, A. (2023). Online moral courage. In C. Cohrs, N. Knab, & G. Sommer (Eds.), *Handbook of Peace Psychology*. https://doi.org/10.17192/es2022.0074

**These authors contributed equally to this work*

Moral courage manifests in interventions intended to stop or redress others' transgressions of moral principles or social norms, despite the risk of incurring physical, financial, or social costs (Frey et al., 2006; Greitemeyer et al., 2006; Halmburger et al., 2016; Niesta Kayser et al., 2016). Following this definition, a broad range of actions qualify as moral courage, for example interventions against bullying, discrimination, or oppression (Baumert et al., 2020; M. Li et al., 2021). While such interventions may involve confrontation of and conflict with transgressors (see also Sasse et al., 2022), their ultimate goal is to uphold and defend moral principles or social norms that ensure the sound functioning of societies (Ellemers et al., 2019; Fehr & Gächter, 2002). In line with this, moral courage has also been considered as a behavior that is characterized by one's caring for others and that protects human and democratic values (Meyer, 2014; Staub, 2015). In the present chapter, we argue that moral courage manifests itself and plays an important role also in online contexts. We analyze the specific affordances and barriers posed by the online context, review empirical evidence on the positive effects of online moral courage, and propose practical recommendations for its enhancement.

Since our social interactions – spanning from friendships, dating, learning, to political debate – take place in considerable and increasing extent on social networking sites (SNS), also transgressions of moral principles and social norms occur in online contexts. According to a recent survey (Vogels, 2021), 41% of the participating US adults had personally experienced some form of online harassment. Within just six years (2014 to 2020), the share of people reporting severe forms of online harassment, such as physical threats or stalking, increased steeply, from 15% to 25%. Often, individuals and groups experience harassment or hate because of their political views, gender, ethnicity, religion, or sexual orientation (Vogels, 2021).

Citizens and policy makers alike have identified online norm transgressions as a major problem with a multitude of negative socio-psychological consequences (van der Wilk, 2018; Vogels, 2021). An Amnesty International survey (Dhorida, 2017), investigating the effects of online abuse and harassment on women, revealed that many targeted women subsequently experienced stress, anxiety, panic attacks, or lowered self-esteem as a consequence, and changed their own online behavior, up until the point of turning silent and withdrawing from online spaces altogether. Online norm transgressions can also aggravate social relations. Frequent exposure to hate speech against outgroups has been associated with increased prejudice towards those groups (Soral et al., 2018), and research from Germany has shown that increased anti-refugee sentiment on Facebook translated into higher crime rates against refugees (Müller & Schwarz, 2021), suggesting that online hate speech may spill over to physical violence offline.

The prevalence and ramifications of online norm transgressions call for effective countermeasures. Other online users can play an important role in this regard, just like bystanders in response to offline norm transgressions. If they perceive the actions of others as a violation of their moral convictions, or of social norms that they endorse, they may take steps to stop or redress these actions, for example by engaging in counterspeech or by reporting to authorities. While taking such steps is often socially desirable, it is not without risk to the person doing so. For instance, those who confront the norm transgressions of

others might themselves quickly become the next target of harassment and hate. As such, taking action against online norm transgressions can be considered *morally courageous*.

The Need for Online Moral Courage

Similar to offline norm transgressions, the types of situations and contexts in which they occur are highly diverse and encompass, for example, cyberbullying, sexual harassment, and hate speech. Despite their differences, all these violations have in common that perpetrators violate fundamental social norms and moral values, such as fairness, and that they cause harm. In many cases, they also constitute transgressions of international legal agreements (such as the EU Framework Decision of 2008) and national law (such as the "incitement to hatred" paragraph in Germany, §130 StGB).

While policy-makers and citizens see online platform providers as responsible for detecting and dealing with violations (Dhorida, 2017), doing so ex ante or proactively can prove difficult for them, often for technical (Ross et al., 2016), ethical (Lepoutre, 2017), or legal reasons (Zufall et al., 2019). This highlights the need for community engagement, by which users who encounter content that violates social norms or their moral beliefs intervene in order to uphold and ensure civil discourse. Depending on the online environment and user rights, they can do so in various ways, for example by directly confronting the transgressor (e.g., through counterspeech), by banning transgressors from groups, or indirectly by reporting them. All these forms of interventions require at least some time and effort (e.g., the interruption of conversations or work, writing a reply or a report), and it is plausible that they bear risks for the person taking the action, ranging from receiving unwanted attention, harsh criticism, to backlash as a direct response to actions, or to negative consequences that transpire into offline contexts and affect relationships, professional life, or physical well-being[1]. As such, intervening against online norm transgressions qualifies as online moral courage.

To date, research on moral courage has thus far mainly been conducted in offline environments. Here, theoretical and empirical work has pointed out that moral courage requires complex psychological processes, and whether or not individuals intervene against others' norm transgressions may depend on a range of individual and situational factors (Baumert et al., 2013; Halmburger et al., 2016; M. Li et al., 2021; Niesta Kayser et al., 2010; Toribio-Flórez et al., 2023). As offline and online environments differ in various ways, for example with regard to anonymity, situational factors in online contexts may shape the psychological processes of moral courage in unique ways.

In this chapter, drawing from a theoretical model of moral courage – the *integrative model of moral courage* by Halmburger and colleagues (2016) – we first identify several crucial situational characteristics of online environments and discuss how they may obstruct or facilitate the psychological processes underlying online moral courage. Second, we discuss both potential beneficial and adverse consequences of online moral courage. Third, we synthesize insights on the psychological processes and the consequences of online moral

[1] While these risks may seem less apparent for reporting perpetrators to authorities, bystanders may still be concerned about them. Depending on the platform, the reporting process may be somewhat intransparent so that the own anonymity may not be seen as ensured or there may be concerns that perpetrators can infer who reported them.

courage to derive practical recommendations that may inform platform policies and the work of practitioners.

Most evidence reviewed in this chapter stems from research on interventions against hate speech on SNS and we highlight whenever we draw from further research on further forms of online norm transgressions.

Obstacles and Facilitators of Online Moral Courage

What determines whether individuals show moral courage? According to the integrative model of moral courage (Halmburger et al., 2016, adapted from Latané & Darley, 1970), prior to acting, observers must *detect* the norm violation and *interpret* it as such, and they must then *assume responsibility* and the necessary *skills to intervene*, and finally *decide to intervene*. According to the model, only if each of these stages is passed successfully moral courage will be shown. For example, even if an observer interprets an instance of hate expressed against members of a minority as wrong, but do not feel responsible to address it, they will not do so.

Whether or not the stages of psychological processes are passed successfully should depend on characteristics of the individual person, as well as of the situation (Halmburger et al., 2016). Situation characteristics, in particular, may differ between online and offline contexts. In this chapter, we focus on five prominent characteristics of online contexts, which, we argue, can work as both facilitators and obstacles of moral courage, namely

a. reach (Bor & Petersen, 2021; Brady et al., 2019; Obermaier et al., 2015; Ziegele et al., 2020)
b. connectedness (Amichai-Hamburger, 2017)
c. permanence (Barberá, 2015; Dillon & Bushman, 2015; Obermaier et al., 2015)
d. asynchrony (K. R. Allison & Bussey, 2016; Obermaier et al., 2015; Suler, 2004)
e. anonymity (Obermaier et al., 2015; Postmes & Turner, 2015; Suler, 2004; Ziegele et al., 2020)

With reach, we refer to the fact that online environments provide the opportunity to communicate with large or distant audiences with little effort. Moreover, people cannot only unidirectionally reach out to other people across the world via the internet, but they can just as easily communicate multidirectionally and network with others, for example to mobilize and organize like-minded individuals. Especially SNS facilitate this *connectedness*. The reach of online communication is further enhanced through a temporal component. While statements made in face-to-face conversations are often of an ephemeral nature, those made online are rather *permanent,* as they remain accessible for a long time, providing the chance that more people will become aware of them or reproduce them at a later point. The permanence of online communication also allows for it to happen *asynchronously*. That is, interactions do not need to be temporally contingent. Instead, people can reply to messages months after they were originally posted. Another critical characteristic of online contexts is *anonymity*. In many online environments, users have – or can choose to have – no or few personal markers that make them identifiable. As such, communication partners can remain anonymous, rendering it uncertain who is making or reading a statement. We argue that these aspects of *reach*, *connectedness*, *permanence,*

asynchrony, and *anonymity* can be both obstacles and facilitators for the psychological processes of online moral courage.

Detection and Interpretation of Online Norm Violations

For moral courage to occur, observers first need to detect the norm transgression and interpret it as such. While this may seem trivial, these processes are not always straightforward. For example, imagine coming across a comment on social media in which one user calls another '*bitch*'. From reading just this term, it is difficult to infer whether this is a sexist insult or whether a group of friends uses the term in a playful way to address each other. In other words, the intention for using the term is ambiguous and thus difficult to infer for observers. Consequently, ambiguity is a barrier to moral courage (Bowes-Sperry & O'Leary-Kelly, 2005; Halmburger et al., 2016; Toribio-Flórez et al., 2023). In online communication, some factors can increase – and others reduce – ambiguity.

Often, individuals and groups who intentionally and frequently transgress norms online disguise their communication to make it particularly difficult for witnesses to detect and interpret transgressions. For instance, transgressors use ciphers to refer to specific marginalized groups without detection from outside witnesses and prosecution. For example, Black people are sometimes referred to with a capitalized "N" or Jews with three parentheses (e.g., commenting "(((they))) are behind everything"). Similarly, transgressors use codes to communicate hateful sentiments, such as '88' instead of 'Heil Hitler'. Plausibly, the connectedness in the online context facilitates the rapid development of hateful jargon, making it particularly difficult for users to detect and interpret transgressions.

Just as connectedness can contribute to norm transgressions, it may also facilitate their detection. Bystanders do not need to act alone, but instead may form groups to coordinate the detection of transgressions and initiate concerted interventions. For instance, groups such as Reconquista Internet (Garland et al., 2020) and #ichbinhier (#iamhere) (Ley, 2018; Ziegele et al., 2020) inform their members about occurrences of hate and vitriolic language, so that members can seek them out and counter them collectively. That way, the detection of transgressions and their interpretation as such do not fall upon individuals, but are organized, thereby facilitating the passing of the first stages of the psychological processes in moral courage.

The interpretation of norm transgressions may also be affected by temporal asynchrony and permanence. On the one hand, if norm-transgressing posts remain visible for a long period of time without being visibly challenged, users might question whether any gut feelings of inappropriateness are in fact warranted. On the other hand, permanence and temporally asynchronous interaction provides users who suspect a norm transgression, for example behind jargon, with time to reflect and inform themselves. This way, ambiguity can be reduced which should facilitate subsequent psychological processes of online moral courage.

Assuming Responsibility

Once observers have interpreted a norm transgression as such, they need to determine whether intervening falls within their responsibility.

Here, the prevalent asynchrony of online contexts may pose a hurdle. In case of older hate speech, people may assume that the issue has been resolved outside of the visible communication channel (K. R. Allison & Bussey, 2016), or that the communication had moved on with no further need to circle back (Leonhard et al., 2018).

In addition, the assumption of responsibility seems to depend on the number of bystanders present, and in online contexts with typically high reach, they are often many. For example, in the context of cyberbullying, Obermaier and colleagues (2016; Study 2) found that students had lower intentions to intervene against cyberbullying when many bystanders were present, compared to very few. This effect was mediated by (lower) feelings of responsibility (see also Machackova et al., 2015; Song & Oh, 2018). Similarly, Leonhard et al. (2018) found that people were more likely to speak up against anti-immigrant hate speech when only four other SNS users saw the transgressive post as opposed to 4,000. These findings suggest that, with an increasing number of bystanders, *diffusion of responsibility* may occur (Darley & Latané, 1968; Fischer et al., 2011).

However, the negative association between number of bystanders and intervention behavior in computer-mediated communication does not always seem to be linear, as the actual and the perceived number of bystanders do not increase proportionately (Machackova et al., 2015; Obermaier et al., 2016). Instead, increases up to two dozen bystanders are perceived disproportionately larger than increases above that, and 24 bystanders are already considered rather many (Obermaier et al., 2016). This might lead to the finding that the bystander effect is more pronounced for increases in smaller groups of bystanders than for increases in bigger groups of bystanders (Machackova et al., 2015) and that there is no linear trend at all once hundreds of bystanders are involved (Obermaier et al., 2016). Potentially, this is because, at a certain point, the sheer number and heterogeneity of individuals in a large audience increase the chances that other factors facilitating interventions are present and outweigh the diffusion of responsibility. For example, Voelpel, Eckhoff and Förster (2008) proposed that the number of so-called "perpetual helpers", individuals with a generally elevated disposition to help, increases with audience size. Hence, while findings suggest that diffusion of responsibility may be prevalent in the online context, the vast reach might at a certain point also serve to counter-act this effect by increasing the odds for the presence of more individuals who are generally disposed to act prosocially.

Subjective Intervention Skills

Beyond assuming responsibility, observers need to determine whether they dispose of the necessary – and effective – skills to intervene and have the opportunity to do so.

As mentioned earlier, norm transgressions are often committed by organized groups, facilitated by the connectedness in online contexts. For example, in the context of the 2017 German elections, the right-wing hate group *Reconquista Germanica*, which had only 1,500 to 3,000 members, published millions of vitriolic posts on Twitter in order to shift the online discourse in the direction of right-wing populism (Garland et al., 2020). In the face of concerted incivility and hate, it seems plausible that bystanders might feel unable to counter such attacks substantially and effectively.

At the same time, the majority of SNS provides guidelines for intervention, and many forms of intervention require little skill or effort, which might lower the threshold for intervening. Due to temporal asynchrony, even users who may not be familiar with given standards have the opportunity to inform themselves about different intervention options. In general, interventions can be conducted either directly or indirectly (Latané & Darley, 1970). In the online context, direct interventions refer to actions such as writing counter-comments against group-based hate comments on a SNS. Also, other easy-to-implement measures can be taken in many online settings to express disagreement with norm transgressions, such as dislike functions to reject hate speech by others. An indirect way of intervening, instead, would be to notify the relevant authorities, for example by reporting the post to the SNS provider or a moderator. Indirect interventions are facilitated across most social media through functions like flagging and reporting of transgressive comments, which can normally be done with a few clicks (Naab et al., 2018). Plausibly, connectedness between users and providers or moderators enhances the knowledge of effective intervention options.

Decision to Intervene

According to the integrative model of moral courage, the final step of the psychological process is the decision to intervene. The model proposes that, at this point, individuals weigh the expected benefits against personal costs which they might suffer as a result of intervening. Those costs can range from the mere investment of time and effort to intervene to the loss of money, physical harm, or backlash from transgressors, as well as drawing unwanted attention to themselves or being evaluated by others (Latané & Darley, 1970).

Plausibly, permanence that characterizes communication in many online environments may foster concerns about the costs of interventions. The fact that in online environments evidence of one's actions often prevails until long after the exchange has taken place (Slonje & Smith, 2008) might trigger fears that one's intervention might be perceived negatively by a wider audience (Dillon & Bushman, 2015; Fischer et al., 2011). Moreover, when engaging with users who use uncivil language, the fear of being associated with them for an unforeseeable amount of time could further raise perceived personal costs (Ziegele et al., 2020). The long-term documentation of one's direct intervention might also invite direct retaliation, such as online harassment or physical violence in the offline world, not only by the original transgressor, but also their sympathizers. Due to the broad reach of online environments, their number can be assumed to be high, but is often unknown to interveners, which may be perceived as particularly threatening.

However, with a large audience, potential interveners might not only fear backlash, but also anticipate support from like-minded individuals. A strong predictor for people speaking up against uncivil language online is expected positive social appraisal (Ziegele et al., 2020) and SNS offer various ways for bystanders to reward morally courageous comments (e.g., likes, following accounts, writing a supportive comment of one's own, retweeting, etc.). Thus, if individuals anticipate support from others, large reach might also promote the decision to intervene.

Moreover, barriers to indirect means of intervention are often explicitly reduced in online environments. As mentioned above, most SNS offer low-effort ways to flag or report norm transgressions. Given that indirect interventions can often preserve the anonymity of interveners, such clear sets of indirect intervention measures should reduce the perceived riskiness of (indirect) intervention.

The anonymity of many online contexts, which emerges due to a scarcity of individualizing markers (e.g., a lack of visual representation of individuals), can also shape decisions to intervene by affecting the salience of group norms. A person's self-image is made up of individual characteristics as well as social group memberships (Tajfel, 1974; Tajfel & Turner, 1986). When a particular group membership becomes salient, people tend to see themselves more in terms of that group membership and consequently to act more in line with the respective group norms (Turner et al., 1987). According to the *Social Identity Model of Deindividuation Effects* (SIDE; Reicher et al., 1995), this salience is increased in environments with few individualizing markers. If individual characteristics of a person recede in a situation due to anonymity, their salient group membership becomes more influential and shapes attitudes and behavior. Thus, in contexts where norms of a salient group favor moral courage, members of that group can actually be more likely to engage in interventions (Levine & Crowther, 2008) and this effect can be especially strong in contexts of computer-mediated communication where reduced individuating cues trigger increased conformity with one's group (Lee, 2004; Postmes et al., 2001). In summary, group norms that support bystander interventions can shape the decision to intervene - in particular in an environment such as computer mediated communication as it does not contain many individuating components.

Conclusion and Outlook

Online norm transgressions such as hate speech and bullying have become a major issue, as they harm individuals, groups, and the societal discourse (Bilewicz & Soral, 2020; United Nations, 2020). Moral courage could play an important role in the attempt to reduce their occurrence and attenuate their detrimental effects especially on social media; yet, to date, this role is not well understood. In this chapter, following the integrative model of moral courage (Halmburger et al., 2016), we highlighted how some defining features of online environments and online communication, namely *reach*, *connectedness*, *permanence, asynchrony,* and *anonymity*, might be both facilitators and obstacles at different stages of the psychological process of moral courage. These considerations are of theoretical relevance for our understanding of online moral courage and may provide a road map for its future comprehensive investigation.

Chapter B – Counterspeech Impact, Collective Norms, and Severity Perceptions

A rising flood of online hate speech threatens discourse and societies around the globe (United Nations, 2020). Online hate speech not only inflicts psychological harm on its victims (Tynes et al., 2008), but may even escalate into offline violence (Müller & Schwarz, 2021). However, bystanders who speak out against online hate speech can substantially attenuate its harmful effects, supporting victims (Leets, 2002) and discouraging perpetrators from further hate speech (Hangartner et al., 2021). Counterspeech effectiveness further increases with the number of bystanders who speak up (Garland et al., 2020; Schieb & Preuss, 2016).

In the present research, we therefore shift the focus from victims and perpetrators to bystanders and ask: Does counterspeech also inspire further bystanders to speak up? If so, through which psychological mechanism and at which time scale does it exert its influence?

Prior research in controlled cross-sectional settings or on social media has yielded inconsistent findings regarding the effects of counterspeech (Alvarez-Benjumea & Winter, 2018; Leonhard et al., 2018; Miškolci et al., 2018). While the hypothetical one-shot nature of the former might underestimate effects of potentially crucial factors like social norms, the uncontrolled nature of field studies cannot rule out contextual confounders which might bias results. To address these issues and to comprehensively assess counterspeech impact, we conducted a longitudinal experiment in which participants (N = 856) repeatedly interacted with a mock social media forum for two weeks ($N_{Observations}$ = 3,605).

Hate Speech and Counter Speech

Hate speech can be defined as communication attacking individuals based on identity factors like race, religion, or gender (United Nations, 2020). It has a multiplicity of adverse effects for its victims, such as causing anxiety and depression (Keighley, 2022; Tynes et al., 2008). Moreover, it negatively affects bystanders by desensitizing them to harmful language and changing which kind of language is deemed acceptable (Bilewicz & Soral, 2020; Hsueh et al., 2015; Soral et al., 2018). It can even lead to offline violence. For instance, a study in Germany found that surges in anti-refugee rhetoric on the social media platform Facebook significantly elevated offline crimes against refugees (Müller & Schwarz, 2021). It is thus paramount to investigate effective countermeasures.

Counterspeech, defined as the direct and overt rejection of hate speech, has been proposed as an effective antidote (Bilewicz et al., 2021; Garland et al., 2020; Hangartner et al., 2021; Miškolci et al., 2018). Victims benefit from seeing others taking their side (Leets, 2002), and transgressors can be effectively discouraged from posting further hate (Hangartner et al., 2021; Munger, 2017; Siegel & Badaan, 2020). For example, users of the microblogging service Twitter (now X) who had posted xenophobic hate speech substantially reduced further transgressions after being asked to consider the effects of hate speech on their victims (Hangartner et al., 2021). In addition to its beneficial effects for victims and perpetrators, counterspeech could inspire further bystanders who might otherwise be reluctant to publicly stand up to online haters. This carries great potential since the increase from one counterspeaker to a handful can already markedly increase the persuasiveness of counterspeech (Schieb & Preuss, 2016) and concerted counterspeech was shown to decrease overall discourse toxicity on Twitter (Garland et al., 2020).

Does Counterspeech Facilitate Further Counterspeech?

Generally, people tend to adopt the predominant conduct in online contexts (Sasse et al., 2023 for a review). This tendency encompasses a wide range behaviors such as emotion expression (Goldenberg & Gross, 2020), language toxicity (Rajadesingan et al., 2020), or constructive and destructive contributions in online chats (Seering et al., 2017).

However, for counterspeech, this disposition could be attenuated by several factors. Individuals may hesitate to copy behavior that would prompt them to directly and overtly oppose another person. Their reluctance could be exacerbated by fears of retaliation from a transgressor who indicated a clear propensity for verbal aggression, through their previous hate comment. Moreover, since hate speakers regularly form like-minded groups, counterspeech opens bystanders up to potential backlash from a great number of other users (Goel et al., 2023). Moreover, internet users could further be discouraged by the perceived futility of counterspeech, assuming a minuscule impact on transgressors and not considering a potential positive effect on other users (Coles & Lane, 2023).

The small body of research investigating whether counterspeech motivates further bystanders to speak up yields mixed results (Alvarez-Benjumea & Winter, 2018; Cary et al., 2020; Leonhard et al., 2018; Miškolci et al., 2018). Cross-sectional experiments in controlled environments did not find that counterspeech increased subsequent bystander counterspeech or their intentions to speak up (Alvarez-Benjumea & Winter, 2018; Leonhard et al., 2018). However, their hypothetical nature might diminish counterspeech effects. When hate speech has real-life consequences, participants might be more motivated to oppose comments they consider harmful. Furthermore, the studies' cross-sectional designs could have reduced participants' motivation. Participants interacted with strangers with whom they expect no future contact. This diverges greatly from real-world interactions that can regularly last for extended periods and where people can reencounter conversation partners or other bystanders later. Such potential for long-term interactions can increase the importance of social norms (Van Bommel et al., 2012), which could serve as an important mediator for counterspeech effects on further bystanders, as we will discuss below. Moreover, the above-mentioned threat of negative feedback in response to counterspeech can only manifest in prolonged interactions.

Overcoming the potential limitations of controlled cross-sectional experiments, other research directly measured counterspeech effects on Facebook by comparing bystander reactions after either posting counterspeech or not responding to anti-Romani hate speech (Miškolci et al., 2018). The study found that counterspeech was followed by more pro-Roma comments, suggesting that counterspeech can inspire further bystanders. However, since the study was performed on an actual social media platform, confounding processes cannot be ruled out. For example, responses to posts, regardless of their content, can boost the initial post's algorithmically determined visibility (Buerger, 2021). Bigger audience sizes of hate speech posts can simply increase the number of bystanders intrinsically motivated to speak up (Sasse et al., 2023) rather than encouraging passive bystanders to become active.

In summary, while prior research offers preliminary information about possible counterspeech effects on bystanders, it does not allow for definitive conclusions. Limited

immersion, cross-sectional designs, and confounding factors could have biased results. We, therefore, conducted a controlled study in which participants interacted with a realistic-looking mock social media forum over an extended period to test whether

> H1: Prior counterspeech against hate speech increases the (actual and self-reported) likelihood of engaging in counterspeech against hate speech oneself compared to exclusively neutral prior responses.

Temporal Effects

Moreover, the likelihood of bystander counterspeech could increase if they see counterspeech not just once but various times across multiple weeks. From a social learning perspective, repetition increases the influence of other people's behaviors (Bandura, 2002; Perry & Bussey, 1979). While prior research suggests that repeated organized counterspeech correlates with an improvement of the overall discourse across time (Garland et al., 2020), it remains uncertain whether such effects are caused by inspiring users to speak up or through confounding factors such as self-selection of toxic users out of such discourses. We, therefore, explored the effect of repeated exposure on further bystander counterspeech as well as on possible mediators, which we discuss in detail in the following sections.

Through Which Mechanisms Does Counterspeech Facilitate Further Counterspeech?

In addition to examining *whether* counterspeech inspires further bystanders to speak up, we were also interested in *how* it exerts its influence. Considering the mixed findings of prior research, pinpointing the conditions and processes through which counterspeech influences bystander behavior is essential. We, therefore, aimed to comprehensively assess the pathways through which counterspeech prompts further bystander engagement.

Bystander Intervention Model

The Bystander Intervention Model (Latané & Darley, 1970) can serve as a valuable framework to understand the conditions under which bystanders decide to engage in online counterspeech (Sasse et al., 2023). The model comprehensively describes the sequence of cognitive appraisals that occur for bystanders between encountering a norm violation and deciding to intervene. It posits that bystanders need to detect the event requiring intervention, interpret it as such, feel responsible and competent, and finally decide to intervene after assessing possible costs and benefits.

We hypothesized that counterspeech inspires bystanders to intervene via two pathways: by heightening their perception of hate speech being an emergency and by decreasing the perceived costs of speaking up.

For bystanders to view a hate speech comment as an emergency that demands countermeasures, they must deem the transgression severe enough to warrant countermeasures and feel certain about their judgment. However, frequent exposure to online hate speech can decrease peoples' severity evaluations (Bilewicz & Soral, 2020; Soral et al., 2018). Alarmingly, seeing online hate speech and harassment as less severe correlates with a decreased willingness to intervene (Bastiaensens et al., 2014; Koehler & Weber,

2018; Leonhard et al., 2018; Lu & Luqiu, 2023; Rudnicki et al., 2023). Observing counterspeech could remind bystanders that online hate speech is not a trivial offense but a highly harmful and damaging issue requiring intervention. Moreover, counterspeech can provide social proof, reaffirming bystanders in their own evaluation that online hate speech comments constitute a transgression (Rendsvig, 2014).

Counterspeech may also reduce the perceived cost of intervening. Concerns about potential embarrassment and negative judgments from others often deter bystander intervention (Rosenberg, 2009; Sabini et al., 2001). Many hesitate to confront online hate speech to avoid looking overly sensitive and pedantic (Buerger, 2021). However, prior counterspeech can mitigate these fears, demonstrating to further bystanders that they are not alone in their disapproval, thereby encouraging them to speak up.

We, therefore, hypothesized that

H2: Seeing others speak out against hate speech

a. increases the perception of hate speech severity
b. increases the certainty of one's assessment
c. decreases perceived costs of counterspeech

H3: These effects increase the actual likelihood of engaging in counterspeech oneself.

We moreover explored the effects of counterspeech on the other steps of the bystander model.

Emotional Reactions

The focus of the Bystander Intervention Model on cognitive appraisals fails to consider effects of emotions which can be important motivators for bystander decisions to intervene against norm violations (Greitemeyer et al., 2006; Halmburger et al., 2015; Kayser et al., 2010). For hate speech evaluations in particular, prior research has highlighted the importance of emotional reactions (Soral et al., 2018). Potentially, counterspeech could also inspire subsequent counterspeech by affecting emotional reactions – specifically, through anger or fear (Sasse, Halmburger, et al., 2022).

Anger has been postulated as a central motivator for bystander interventions against norm transgressions (Greitemeyer et al., 2006; Halmburger et al., 2016; Osswald et al., 2011; Sasse, Halmburger, et al., 2022). For instance, anger about a fascist political party predicted whether people risked personal harm to promote its prohibition (Kayser et al., 2010). People regularly experience anger when others violate salient norms (Ellsworth & Scherer, 2003). Since hate speech is generally considered anti-normative (Bilewicz et al., 2017), counterspeech by others could increase the salience of this norm violation, resulting in increased feelings of anger. This anger could, in turn, lead participants to perceive the hate speech as more harmful.

H4: Seeing others speak out against hate speech enhances anger, which in turn increases severity evaluations.

In addition, witnessing prior counterspeech could decrease the fear associated with posting counterspeech oneself. In addition to fearing adverse reactions from other forum users (Buerger, 2021; Rosenberg, 2009), bystanders might also be afraid of backlash from the transgressor who, after all, had already displayed a willingness to engage in vitriolic language against others (Buerger, 2021; Sasse et al., 2023). Prior counterspeech could reassure bystanders by signaling potential support from other forum users. Reduced fear, in turn, would lead to reduced expected costs of one's own counterspeech.

> H5: Seeing others speak out against hate speech reduces fear, which in turn decreases cost evaluations.

Counterspeech Norms

Finally, social norms make up another potentially crucial driver of bystander counterspeech. They can be defined as the behaviors that are common and endorsed in a certain group or context (Tankard & Paluck, 2016). Individuals are generally motivated to adhere to social norms driven by a desire to fit in and to be accepted by their peers (Asch, 1955; Tankard & Paluck, 2016). They affect a variety of cognitive appraisals in the Bystander Intervention Model (Latané & Darley, 1970), such as feeling responsible (Levine & Manning, 2013) and the assessment of the costs and benefits of an intervention (Levine & Crowther, 2008). Therefore, we investigated social norms as a complementary factor rather than subsuming their effect in one of the model's steps.

In digital spaces, norm influence is regularly more pronounced because individuating features are less prominent than in psychical settings (Postmes et al., 2001). For instance, visual cues are greatly reduced in most text-based interactions, such as chat rooms or discussion forums, compared to face-to-face conversations. The Social Identity model of Deindividuation Effects posits that a scarcity of individuating cues causes people to see themselves in terms of collective identities and their perceived norms (Reicher et al., 1995; Spears & Postmes, 2015).

In online contexts, individuals use other users' behaviors to assess what is considered acceptable, tailoring their actions to these perceived norms (Postmes et al., 2001; Rajadesingan et al., 2020; Seering et al., 2017). Therefore, counterspeech could inspire further bystanders to speak up via its effect on social norms. Offline, observing others speak out against prejudice can have a lasting impact on the perceived admissibility of prejudice (Blanchard et al., 1994) and on self-reported interventions against it (Bennett & Sekaquaptewa, 2014). Online, social norms have been suggested as a motivator for bystander counterspeech but have not yet been directly tested (Siegel & Badaan, 2020). We thus predicted that,

> H6: Seeing others speak out against hate speech enhances the perception of pro-counterspeech norms, which in turn increases the (actual and self-reported) likelihood to engage in counterspeech oneself.

Present Research

Over two weeks, participants interacted with a mock social media forum whose content we had curated and that – depending on experimental condition – contained counterspeech or not. This enabled us to measure actual behavior in a controlled setting, closely resembling

real-world social media. Following their forum interaction, participants indicated their perceptions of the forum and answered questions about a hypothetical transgression scenario in a separate survey at each time point. This allowed us to assess underlying psychological mechanisms that may have driven their forum behavior. In addition, our longitudinal multi-week design allowed us to explore temporal dynamics.

Method

The design, procedure, and analysis plan were pre-registered. Data, analysis code, materials, the supplementary materials, and the pre-registration be found at https://osf.io/rd3my/?view_only=d93e634783f045468cfc90e36a187bb0.

Participants

The study was conducted in German with a German sample recruited through a panel provider. We screened 1229 participants, of whom 856 decided to take part in our study, resulting in a total of 3605 observations across the five time points. The participants' mean age was 43.20 (*SD* = 13.94). Our sample consisted of 413 men, 442 women, and one person who indicated another gender identity. The sample was slightly left-leaning with a mean political orientation of 3.73 (*SD* = 1.13; 1 = "left," 7 = "right"). Participants were excluded from further data collection if they incorrectly answered all three attention checks at a given time point (see below for details). We had to exclude 16 participants at T1, nine at T2, 10 at T3, five at T4, and four at T5. Moreover, we were unable to match 24 forum responses to surveys due to missing identifiers.

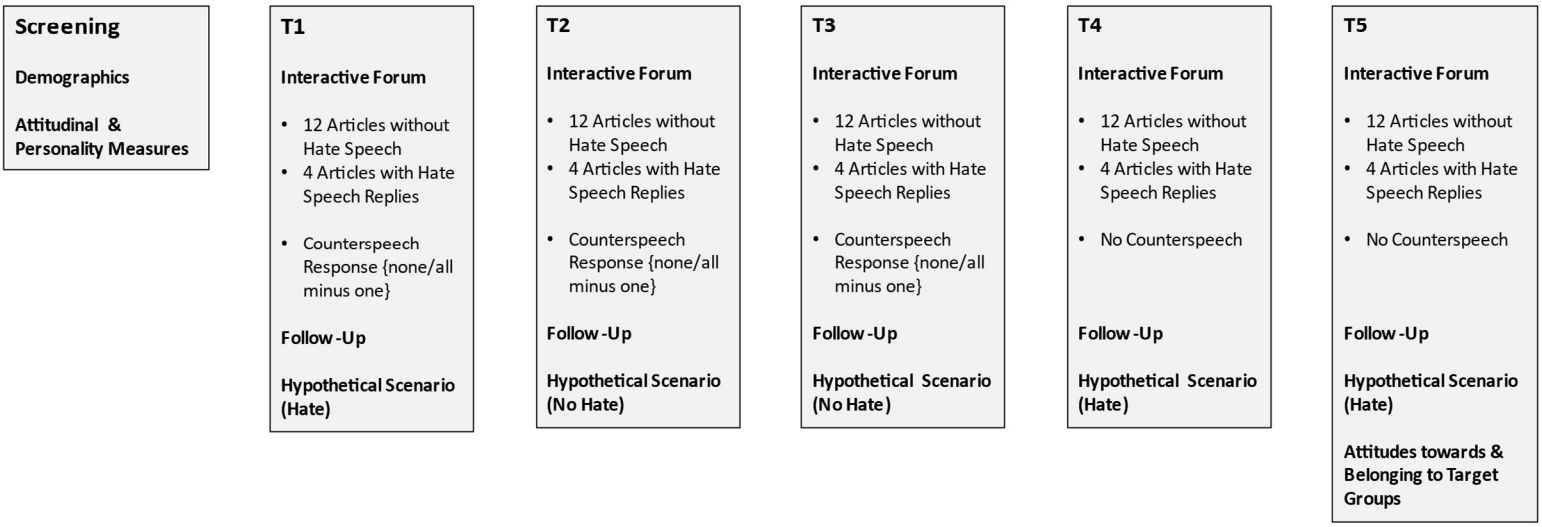
Design & Procedure

Before the actual experiment, participants provided demographic information and answered attitudinal and personality measures in a screening session in an online survey. Participants were excluded if they indicated they did not speak German fluently or were younger than 18 years old. After the screening session, eligible participants were invited to take part in the actual experiment (see *Figure 1* for an overview).

The general procedure was similar for each time point: After indicating their consent, participants interacted with the mock social media forum. After that, they answered some follow-up questions regarding their forum experience. Finally, we presented participants with a hypothetical scenario that included a norm violation and asked them about their appraisals.

Figure 1

Study Design



Interactive Forum

At each time point, participants interacted with a mock social media forum with an interactive interface that gave the users a realistic social media experience (Jagayat et al., 2021). In the forum, participants saw posts containing headlines of sixteen news articles covering a wide array of topics such as celebrity gossip, recent scientific discoveries, current political events, or recipes. Other ostensible forum users had responded to some of these articles. While responses to most articles were unproblematic, participants encountered hate speech replies to four articles at each time point. In a between-subjects design, we manipulated whether the hate speech replies were followed by *neutral speech* or *counterspeech* at T1-T3. In the neutral condition, participants saw exclusively neutral or no replies to the hate speech. In the counterspeech condition, participants saw counterspeech to three of the four hate speech comments they encountered; the unresponded fourth hate speech comment is henceforth referred to as *focal* and relevant for our analyses. At T4 and T5, all participants saw exclusively neutral or no responses[2] to hate speech, allowing us to investigate whether prior counterspeech could exert lasting effects.

Using the forum's interactive functionality, participants could compose comments and click on any of the news headlines to read the corresponding article.

[2]We had intended to only show exclusively neutral comments at T5 to the participants in the counterspeech condition, however, due to a programming error, all participants already saw exclusively neutral comments at T4. We conducted robustness tests to rule out any biasing effects of the coding error as reported below.

Follow-Up and Hypothetical Scenario

After interacting in the forum, participants were re-directed to the survey platform and responded to three attention checks and to follow-up questions about their experience in the mock social media forum. Finally, we showed subjects a hypothetical scenario of an online news article and user responses. The user responses included a hate speech comment only at T1, T4, and T5. The other two scenarios served as distractors to mask our study's research question. Participants answered multiple questions about their anticipated reactions and their perceptions of the scenario, after which data collection was completed for the respective time point.

Target Groups

Prior research found that hate speech is perceived differently depending on the victimized groups (Obermaier et al., 2023). We, therefore, included four different hate speech target groups for our main analyses to increase generalizability. In a pre-study (*N* = 105), participants indicated their liking of different groups in German society and their assumed prevalence of hate speech against the respective groups. We applied the following criteria for our final selection: The average endorsement of the group was above the scale midpoint, and participants considered the group to be a plausible target of online hate speech. Moreover, we aimed to include groups whose evaluations did not differ substantially across political orientation. The four groups that fulfilled these criteria to the greatest extent were overweight people, Black people, Jews, and homeless people (*see Supplement for pre-study analyses*). We, therefore, primarily assessed responses to hate speech against these groups. Groups that fulfilled the criteria to a lesser extent were chosen for the hypothetical scenarios and the non-focal hate speech (*see Supplement*).

In a balanced design, participants saw one hate speech comment against one of these groups per time point T1-T4 in random order. At T5, subjects saw one hate speech comment targeting each group. The other three non-focal hate speech comments during T1-T4, respectively, were randomly distributed across other groups (*see Supplement*).

During the hypothetical hate speech scenarios at T1, T4, and T5, participants were presented with hate speech comments against jobless people, gay men, and Turks in random order. We assessed participants' attitudes towards the respective groups and their own belonging to any of the groups at T5.

Measures

We assessed the central measures at different points of our study, as outlined below. For a comprehensive list of all exploratory measures, see *Supplement*.

Screening

Right-Wing Authoritarianism

We assessed agreement with the three items on the Ultra Short Authoritarianism Scale (Heller et al., 2020) (e.g., "Established conducts should not be questioned."; 1 - *strongly disagree* to 7 - *strongly agree*).

Political Orientation

Political orientation was measured on a seven-point Likert scale (1 = *left*, 7 = *right*).

Interactive Forum

Counterspeech Comments

Responses to the hate speech comments were coded as counterspeech or no counterspeech, each by two independent raters blind to the conditions and hypotheses. The principal investigator resolved inter-rater disagreements. As the main dependent counterspeech variable, we coded whether participants had responded with a counterspeech comment (1 = yes; 0 = no) to the unresponded hate comment presented across experimental conditions at T1-T3 and to the first hate speech comment at T4 and T5.

Attention Checks

To ensure that participants had properly engaged with the forum, they answered three questions regarding the articles' content and the comments. Questions depended on the presented content and hence varied between time points (e.g., "Which famous person was mentioned in one of the articles?" at T1).

Follow-Up Questionnaire

In addition to some distractor questions, participants indicated *Counterspeech Norms* at every time point and *Audience Size, Pro-Target Group Attitudes*, and *Belonging to Target Groups* at T5.

Counterspeech Norms

Participants indicated agreement with the statement: "Users of this forum speak out against hate speech." (1 - *strongly disagree* to 7 - *strongly agree*)

Credibility Check: Audience Size

Participants indicated how many people they expected to potentially see their comments on a scale from 0 to "more than 100".

Pro-Target Group Attitudes

Participants indicated how positively they felt about the hate speech target groups from 1 – *very negative* to 7 – *very positive*.

Belonging to Target Groups

Participants indicated whether they belonged to one or more target groups (yes/no).

Hypothetical Scenario

At T1, T4, and T5, participants answered the following items for a hypothetical hate speech comment.

Counterspeech Intentions

Participants indicated their likelihood to show different reactions to the hate speech comment if they encountered it in a forum (e.g., "criticize the comment") on a scale from 1 (*very unlikely*) to 7 (*very likely*).

Severity

Participants indicated the perceived severity of the hate speech comment on a four-item, seven-point Likert scale (e.g., "The comment is hurtful.")

Further Bystander Variables

Subjective *Certainty* (e.g., "I am unsure whether I am assessing the situation correctly."), *Responsibility (*e.g., "I personally feel obliged to oppose the comment."), and *Competence (*e.g. "I feel competent to speak up against the comment.") were assessed with three items each, and subjective *Cost* (e.g., "I feel uncomfortable at the thought of positioning myself against the comment.") with two items. Response options ranged from 1 (*strongly disagree*) to 7 (*strongly agree*). The items measuring responsibility were adapted from Obermaier et al. (2021).

Emotions

Participants indicated their *Anger* and *Fear* by answering the questions: "When thinking about the above-mentioned comment [hate in the hypothetical scenario] and whether to react, did you feel [emotion]? (1 – *strongly disagree* to 7 – *strongly agree*).

Results

Descriptive statistics for the variables assessed in the screening session are depicted in Table 1, for the outcome variables per time point and condition in *Table 2*, and for the bystander variables per time point and condition in *Table 3*.

Table 1

Attitudinal and Personality Variable Distributions.

	Neutral Condition	Counterspeech Condition	Cronbach's α
Right-Wing Authoritarianism	3.98 (1.21)	3.94 (1.17)	0.54
Attitudes: Black People	4.67 (1.56)	4.6 (1.58)	
Attitudes: Overweight People	4.33 (1.44)	4.7 (1.42)	
Attitudes: Homeless People	4.58 (1.08)	5 (1.15)	
Attitudes: Jews	5.33 (1.15)	4.9 (1.29)	

Mean values per condition and standard deviations are in parentheses. Where applicable, Cronbach's α is listed. Answers ranged from 1 (*strongly disagree*) to 7 (*strongly agree*) for the first three items and from 1 (*very negative*) to 7 (*very positive*) for the last four.

Table 2

Counterspeech Variable Distributions.

	T1		T2		T3		T4		T5	
Condition	neutral	counter	neutral	counter	neutral	counter	neutral	counter	neutral	counter
N	397	387	336	343	356	354	379	344	365	344
Any Comment	164 (41.3%)	176 (45.5%)	138 (41.1%)	153 (44.6%)	145 (40.7%)	160 (45.2%)	163 (43%)	157 (45.6%)	142 (38.9%)	159 (46.2%)

Counter Comment	38 (9.6%)	66 (17.1%)	46 (13.7%)	50 (14.6%)	58 (16.3%)	64 (18.1%)	78 (20.6%)	91 (26.5%)	64 (17.5%)	87 (25.3%)
Counterspeech Outside	3.53 (1.97)	3.52 (1.99)					3.49 (1.98)	3.64 (2.01)	3.53 (1.99)	3.52 (2.09)
Counterspeech Norms	3.09 (1.81)	4.37 (1.59)	2.92 (1.65)	4.72 (1.5)	2.95 (1.6)	4.8 (1.54)	3.24 (1.69)	3.74 (1.55)	3.31 (1.65)	3.76 (1.67)

Average Cronbach's α for *Counterspeech Outside* = 0.96. Number of observations and percent values in parentheses or means and standard deviations in parentheses. Answers ranged from 1 (*strongly disagree*) to 7 (*strongly agree*) for the last two items.

Table 3

Distributions of Variables Related to Cognitive and Emotional Appraisals.

	T1		T4		T5	
Condition	neutral	counter	neutral	counter	neutral	counter
Certainty	5.05 (1.28)	5.06 (1.29)	5.27 (1.35)	5.43 (1.21)	5.32 (1.37)	5.4 (1.25)
Severity	5.02 (1.52)	5.18 (1.32)	4.8 (1.44)	5.05 (1.36)	5.04 (1.48)	5.23 (1.43)
Responsibility	3.15 (1.89)	3.24 (1.89)	3.06 (1.87)	3.22 (1.93)	3.22 (1.93)	3.27 (1.96)
Competence	4.39 (1.67)	4.41 (1.63)	4.75 (1.68)	4.78 (1.62)	4.58 (1.77)	4.63 (1.73)
Costs	2.72 (1.52)	2.63 (1.54)	2.59 (1.5)	2.39 (1.58)	2.44 (1.44)	2.41 (1.58)
Anger	4.54 (2)	4.58 (1.96)	4.15 (1.98)	4.24 (2.01)	4.29 (2.11)	4.32 (2.16)
Fear	2.41 (1.62)	2.41 (1.57)	1.99 (1.4)	1.87 (1.38)	2.15 (1.46)	1.94 (1.44)

Average Cronbach's α: *Certainty* = 0.68, *Severity* = 0.82, *Responsibility* = 0.98, *Competence* = 0.88; average correlation *Cost* = 0.62. Number of observations and percent values in parentheses or means and standard deviations in parentheses. Answers ranged from 1 (*strongly disagree*) to 7 (*strongly agree*).

Credibility Check

Participants indicated, on average, that 730 other participants would be able to see their posts (*SD* = 540). Only 7 participants indicated that they expected no other forum users to read their potential comments.

Analyses

We had pre-registered multi-level regression analyses to account for having multiple observations per participant. However, these models underestimated the effect of the

predictors on the participants' counterspeech, likely due to zero-inflation. We, therefore, conducted regression analyses with clustered robust standard errors (clustered per participant) instead. This method has been highlighted as a valid alternative to multi-level models (Cameron & Miller, 2015, for an overview). This resulted in no notable changes in significance test results but in a more accurate prediction of our outcome measures (*see Supplement* for multi-level analyses). For example, the frequency of counterspeech predicted by regression analysis with clustered robust standard errors more closely matched the observed counterspeech frequency than that estimated by the correspondent multi-level analysis.

Effect of Prior Counterspeech

Overall Effects

First, we tested whether seeing prior counterspeech motivated individuals to (a) engage in more counterspeech in the forum and (b) indicate that they would speak up against hypothetical hate speech outside the study.

A logistic regression analysis that tested whether *Condition* predicted *Counterspeech Comment* yielded a significantly higher log-likelihood in the intervention condition than the neutral condition, β = 0.32, *SE* = 0.13, *p* = .014. Conversely, a linear regression with *Counterspeech Outside* as the criterion yielded no effect of *Condition* on *Counterspeech Comment,* β = 0.02, *SE* = 0.06, *p* = .723.

Temporal Effects

An exploratory linear regression analysis with *Counterspeech Comment* as the outcome variable and *Time Point* (T1 = 0, T2 = 1, …, T5 = 4), *Condition*, and their interaction as predictors yielded that across time points, counterspeech increased, independent of conditions (*Time Point*: β = 0.35, *SE* = 0.17, *p* = .045; *Condition*: β = 0.18, *SE* = 0.04, *p* < .001; *Time Point X Condition*: β = -0.01, *SE* = 0.05, *p* = .908).

Cognitive and Emotional Appraisals

Overall Effects

Next, we tested the effect of *Condition* on *Severity, Certainty,* and *Cost with linear regressions.* Only *Severity* was positively affected by prior counterspeech.

Logistic regressions with *Counterspeech Comment* as the criterion and, respectively, *Severity, Certainty,* and *Cost* as the predictor yielded that each variable individually predicted counterspeech in the mock social media forum.

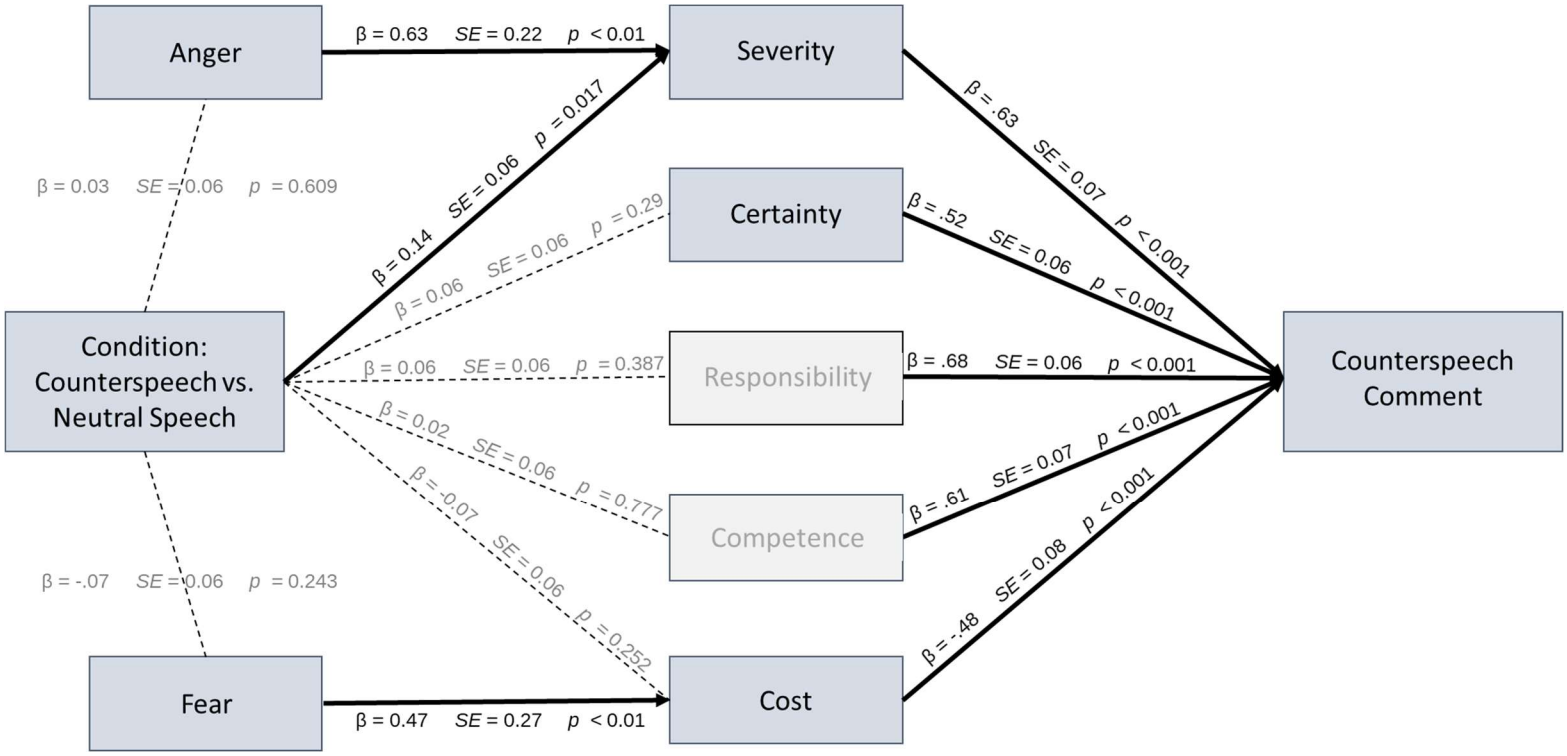
Linear regressions with *Anger* and *Fear* as the outcomes and *Condition* as the predictor further yielded no effect*.* When we regressed *Severity* on *Anger* and *Costs* on *Fear*, we did find a significant association in both cases.

Exploratory analyses testing the effect of *Condition* on *Responsibility* and *Competence* also found no effect, but both variables predicted *Counterspeech Comments* in the mock social media forum.

The results of the analyses are shown in *Figure 2.*

Figure 2

Prior Counterspeech, Cognitive and Emotional Appraisals, and Subsequent Bystander Counterspeech



Separate univariate regressions (e.g., *Severity* regressed on *Anger*) depicted together. Standardized regression weights, standard errors, and p-values of logistic regressions for the outcome variable *Counterspeech Comment* and linear regressions for any other outcome variables. Analyses involving *Responsibility* and *Competence* were exploratory, the others confirmatory and pre-registered.

To confirm the mediating effect of *Severity*, we conducted a non-preregistered mediation analysis as described above, exchanging *Counterspeech Norms* for *Severity* this time. The causal mediation effect was 0.01 (95% CI [0.00, 0.03], p = .015) at the intervention condition and 0.01 (95% CI [0.00, 0.02], p = .015) at the neutral condition, amounting to an average 17.4% of the total effect of the intervention condition on bystander counterspeech being mediated by *Severity*. For full results, see *Supplement*.

Temporal Effects

An exploratory linear regression analysis with *Severity* as the outcome and *Condition*, *Time Point* (T1 = 0, T4 = 1, T5 = 2), and their interaction as predictors yielded no effect for *Time Point* (β = 0.01, *SE* = 0.03, p = .846) or the interaction (β = 0.01, *SE* = 0.04, p = .813) and a positive non-significant effect for *Condition* (β = 0.13, *SE* = 0.07, p = .059). Simple slope analyses showed that *Condition* had a significant effect at T4 (β = 0.14, *SE* = 0.06, p = .016) and at T5 (β = 0.15, *SE* = 0.07, p = .041).

Counterspeech Norms

Overall Effects

Next, we tested whether perceptions of a pro-counterspeech forum norm as another possible mediator. We first regressed *Counterspeech Norms* on *Condition,* finding a significant positive effect, β = 0.67, *SE* = 0.05, p < .001.

Next, we conducted a logistic regression analysis with *Counterspeech Comment* as the criterion and *Counterspeech Norms*, *Condition*, and their interaction as the predictors. Analyses yielded a negative main effect for *Counterspeech Norms* (β = -0.33, *SE* = 0.08, p < .001) and a positive effect for the *Condition* (β = 0.35, *SE* = 0.14, p = .015) as well as a positive interaction between the two terms (β = 0.55, *SE* = 0.11, p < .001). These results indicate that *Counterspeech Norms* had the predicted positive effect on *Counterspeech Comment* only in the intervention condition and their effect was, unexpectedly, negative in the neutral condition.

Finally, we conducted a mediation analysis using the `mediation` package (Tingley et al., 2014). We estimated confidence intervals with 5000 simulation draws and clustered standard errors per participant. The causal mediation effect was 0.02 (95% CI [0.01, 0.04], p = .003) in the intervention condition and -0.03 (95% CI [-0.04, -0.02], p < .001) in the neutral condition. Direct effects were 0.07 (95% CI [0.03, 0.11], p < .001) in the intervention condition and 0.02 (95% CI [-0.02, 0.06], p = .288) in the neutral condition. Counterspeech Norms mediated 51% of the total effect in the counterspeech condition.

Temporal Effects

To test the temporal effects of counterspeech on norm perceptions, we conducted separate exploratory analyses for only the first three time points, during which participants in the counterspeech condition saw counterspeech by others and for all five time points.

A linear regression analysis across T1-T3 with *Counterspeech Norms* as the outcome and *Time Point* (T1 = 0, T2 = 1, T3 = 2), *Condition*, and their interaction as the predictors yielded no main effect of *Time Point* (β = -0.04, *SE* = 0.03, p = .159), a positive effect for the *Condition* (β = 0.74, *SE* = 0.06, p < .001), and a positive interaction effect (β = 0.16, *SE* = 0.04, p < .001). Simple slope analyses using the `marginaleffects` package (Arel-Bundock, 2023) yielded that the time point had an effect in the counterspeech condition (β = 0.12, *SE* = 0.03, p < .001), but not in the neutral condition (β = -0.04, *SE* = 0.03, p = .159).

Counterspeech Norms decreased at T4 when participants in the counterspeech condition did not see any more counterspeech (*see Table 2*), so the linear trend observed for T1-T3 disappeared in a linear regression analysis considering T1-T5 (*see Supplement*).

Reverse Forum Norms Effect in Neutral Condition

Finally, we explored why the relationship between *Counterspeech Norms* and *Counter Comment* could have been reversed in the neutral condition. We reasoned that political bias could have driven the effect. Linear regressions with *Counterspeech Norms* as the criterion showed the same pattern for *Right-Wing Authoritarianism* and *Political Orientation* as a predictor together with *Condition* and the interaction between *Condition* and the respective predictor. In the neutral condition, *Right-Wing Authoritarianism* (β = 0.19, *SE* = 0.04, p < .001) and *Political Orientation* (β = 0.11, *SE* = 0.04, p = .010) predicted *Counterspeech Norms*. However, in the counterspeech condition, this pattern was not observed as indicated by an opposing interaction term (*Right-Wing Authoritarianism*: β = -0.08, *SE* = 0.05, p = .095; *Political Orientation*: β = -0.12, *SE* = 0.05, p = .033).

Robustness Tests

The results of our confirmatory analyses were confirmed via multiple robustness tests (*see Supplement*).

Discussion

We set up a mock social media forum to investigate the effect of counterspeech on subsequent bystander counterspeech over two weeks. We measured actual behavior as well as behavioral intentions and found that seeing counterspeech by other forum users increased bystander counterspeech against hate speech but not intentions to speak up. The effect of counterspeech on subsequent bystander behavior was mediated by perceiving counterspeech as normative in the forum and by perceiving hate speech as a severe, rather than a trivial, transgression. Norm perceptions and severity assessments were positively affected from the first time bystanders saw counterspeech; repeated encounters were not necessary for this effect to emerge. Moreover, perceptions of pro-counterspeech norms increased with the number of expositions. Our results further showed that the positive effects of counterspeech could last for multiple days.

Counterspeech Increases Subsequent Bystander Counterspeech

Participants who saw counterspeech by others were more likely to speak up themselves than individuals who saw only neutral replies. Our findings suggest that counterspeech can not only positively affect victims (Leets, 2002) and transgressors (Hangartner et al., 2021) but also increase the likelihood that subsequent bystanders speak up. These results align with observations on social media that counterspeech positively influences other bystanders (Miškolci et al., 2018) and with self-reports that show a positive correlation between one's own propensity to counterspeech and that of one's surroundings (Cary et al., 2020). Our controlled experimental setting enabled us to rule out alternative explanations, such as post visibility or biased recall. Our results, therefore, provide strong evidence that counterspeech motivates subsequent bystanders to speak up.

We did not observe the same effect for mere intentions to speak up against hypothetical hate speech. These findings match prior studies, which found no impact on intentions to speak up (Leonhard et al., 2018; Obermaier et al., 2021). Prior research has shown that intentions to speak out against hate speech generally exceed actual counterspeech (Crosby & Wilson, 2015; Kawakami et al., 2009). Also, in our study, the average self-reported likelihood to speak out against hate speech was generally just below the scale mid-point, whereas actual counterspeech in the forum never exceeded 26%. Moreover, intentions to intervene against norm violations were found to be predicted by different personality traits than actual bystander interventions (Baumert et al., 2013; Goodwin et al., 2020). Similarly, our results suggest that bystander counterspeech might influence these two types of outcomes differently. Our results thus highlight the importance of selecting appropriate measures to assess the effects of online hate and counterspeech.

Counterspeech Norms

We found that the perception of pro-counterspeech forum norms mediated the relationship between prior counterspeech and subsequent bystander behavior. After seeing others speak up against hate, participants felt more strongly that it was normative in the forum to

speak up against hate speech. Thus, complementary to the observation that hate speech can deteriorate conversation norms (Alvarez-Benjumea, 2022; Bilewicz & Soral, 2020), our results suggest that counterspeech can address the same mechanism, thereby constituting a potential antidote to repair some of the damage caused by hate speech.

While we observed a positive effect of counterspeech norms on bystander counterspeech in the counterspeech condition, we found an unexpected negative effect in the neutral condition. This is puzzling because it means that in that condition, bystanders who assumed that people in the forum generally stand up to hate speech were less likely to speak out against it. Since participants in the neutral condition did not see any actual counterspeech by others, their assumption must have been driven by another reason, possibly a different definition of hate speech. In our study, the four focal groups were more positively evaluated by left-leaning than right-leaning individuals. The latter could, therefore, not have deemed attacks against these groups hate speech. Right-leaning participants perceived counterspeech against online hate speech as more normative than left-leaning ones in the neutral but not the counterspeech condition. This indicates that they may have driven the reverse effect of counterspeech norm perceptions in the neutral condition. This pattern might have been the opposite if we had considered target groups that are more strongly endorsed by right-leaning individuals.

Cognitive Appraisals and Emotional Evaluations

We identified the perceived severity of hate speech as a second factor mediating the relationship between prior counterspeech and subsequent bystander counterspeech. All factors related to the Bystander Intervention Model (Latané & Darley, 1970) predicted whether our participants spoke up against hate speech, but only severity assessments were positively influenced by prior counterspeech. That is, seeing counterspeech in the mock social media forum increased the participants' severity evaluations of hate speech in a separate hypothetical scenario, which in turn significantly predicted counterspeech in the mock social media forum. This indicates that prior counterspeech can serve as a reminder that online hate speech is, in fact, a grave and harmful transgression and not just a trivial nuisance. Hence, it could counteract the desensitizing effects of online hate speech, which normalize its occurrence in the eyes of bystanders (Pluta et al., 2023; Soral et al., 2018). Our findings align with prior research that found similar effects for transgressors, showing they can be discouraged from further offenses by counterspeech that stresses the harmful effects of hate speech (Hangartner et al., 2021).

We did not find effects of prior counterspeech on bystander certainty about their judgments and expected intervention costs nor on anger and fear regarding hate speech posts and possible interventions. The absence of these effects could have been caused by contextual factors immanent to our experimental design. For instance, participants did not receive responses to their comments. This could have caused participants' low fear and expected intervention costs across all time points, potentially attenuating any beneficial counterspeech effects. However, the absence of counterspeech effects on these dimensions could also indicate their general absence, for example due to concurrent processes. For example, bystanders who had seen counterspeech before possibly felt more anger at the transgression but could also more likely envision reducing that anger through own

counterspeech. Although our data cannot definitively tell why we did not observe some mechanisms, our results show a consistent positive effect of counterspeech on severity evaluations and pro-counterspeech norms, underscoring their importance as mediators of counterspeech effects.

Temporal Effects

We found that counterspeech had an immediate positive effect on subsequent bystander counterspeech. In line with prior research (Miškolci et al., 2018), these results indicate that very few counterspeech comments can already impact bystanders positively. Our results, therefore, reject the notion that previous cross-sectional studies failed to observe a counterspeech impact on bystanders due to a delayed effect or the need for multiple expositions. However, it remains plausible that the prospect of interacting repeatedly with the environment enhanced counterspeech impact on subsequent bystander interventions.

Counterspeech influenced norm perceptions from the first instance. In addition, the perception of pro-counterspeech forum norms increased with the number of times that participants saw others speak up. Our results match findings that people quickly pick up on social norms (Cialdini et al., 1990) and that these norms are reenforced with increasing exposition (Bandura, 2002; Postmes et al., 2001). This offers encouraging evidence that even a handful of counterspeech instances can be enough to improve bystander perceptions of pro-counterspeech norms and that social media in which counterspeech is frequent can foster an especially inclusive environment.

In contrast, we observed that the positive effect of counterspeech on severity assessments remained relatively constant across time. Since they were assessed only at time points one, four, and five, we could not directly measure a learning effect as we did for pro-counterspeech forum norms. Indirectly supporting such an effect, we only observed a statistically significant impact on severity assessments at time points four and five and not at the first one. Nevertheless, the actual effect size difference across time was minimal and not statistically significant. This suggests that counterspeech overall and especially repeated counterspeech positively affects severity assessments. It remains less certain how quickly this effect sets in. Future research could examine the impact of multiple counterspeech expositions on severity assessments with a higher temporal resolution to address this question.

Encouragingly, even six days after last seeing counterspeech by others, participants displayed heightened levels of own counterspeech engagement, pro-counterspeech norm perceptions, and hate speech severity assessments compared to participants who had not seen any counterspeech. Our results match findings that counterspeech can also positively affect transgressors for up to a month after exposure (Hangartner et al., 2021; Munger, 2017), recommending counterspeech as a sustainable tool against online hate speech.

Limitations

While our interactive mock social media forum closely imitated real-world interactions, the generalizability of our findings is limited by two factors. First, participants exclusively interacted with previously unknown users. Conversely, in natural social media settings, a user's potential audience regularly includes friends, family, and colleagues. The possibility of

offline consequences by valued others could increase the role of social norms (Klein et al., 2007). However, on the internet, people also tend to form groups with like-minded others, especially users who regularly engage in hate speech (Goel et al., 2023). Bystanders could consider hate speech less severe if acquainted with the transgressor. Second, participants in our study did not receive direct responses to their comments. The possibility of being confronted with negative feedback can increase peoples' adherence to social norms (Ellingsen & Johannesson, 2008), possibly increasing the influence of pro-counterspeech norms. In contrast, fear of adverse transgressor reactions could suppress some bystander interventions (Buerger, 2021), possibly more so if participants had seen no one else speak up. Both, no self-selection of one's audience and no responses to one's comments, were necessary to keep the setting constant for all participants. However, by identifying severity and social norms as mediators, our results can inform future research in real-life social media settings of possible boundary conditions for counterspeech effects. To comprehensively investigate the limits of counterspeech effects on bystanders, future research could moreover employ controlled experiments that iteratively increase the complexity of their setup to approach real-world levels.

Another limitation of our study was its reliance on single-item measures to assess constructs such as forum norms. Considering the extensive design of our study and the concern that participants might tire or drop out if confronted with too many questions, we were compelled to reduce the overall length of our follow-up surveys. Having identified the central mechanisms, future studies could reduce the number of assessed constructs in favor of a more comprehensive examination of the remaining ones.

An additional limitation was that some of our items used the term "hate speech," which could have been interpreted heterogeneously. This might have been involved in reversing the effect of perceived pro-counterspeech speech norms in the neutral condition. To decrease ambiguity, future studies could provide a more explicit definition of hate speech or let participants indicate what they consider hate speech.

Practical Implications & Conclusion

Our findings show that counterspeech positively influenced bystanders in two ways: It generated the impression that counterspeech is normative and reminded people that online hate speech is a harmful transgression. Counterspeech could thus potentially counteract the deleterious effects of hate speech, which work through desensitization and the deterioration of inclusive social norms (Bilewicz & Soral, 2020). However, internet users are seldom willing to speak out against online hate speech, citing the perceived futility (Coles & Lane, 2023). Encouragingly, our findings showed that bystander counterspeech can substantially impact its audience even when relatively few people speak up. Our longitudinal design further yielded that its positive effect can last for days.

Chapter C – Ingroup Norms and Bystander Counterspeech

This chapter is based on:

Cypris, N.F., Sasse, J., Grossklags, J., & Baumert, A. (*under review*). Countering Online Hate Speech: Social Norms Predict Bystander Intervention.

Hate speech is associated with a multitude of the most harmful outcomes for its targets as well as society as a whole (Bilewicz & Soral, 2020). It constitutes an especially prevalent problem on the internet. In a recent survey in the United States, 33% of the respondents reported having experienced online harassment because of their identity in the last year and 28% reported race-based harassment (ADL, 2021). Fortunately, bystander interventions in the shape of counterspeech can be a potent weapon against online hate (Garland et al., 2020; Hangartner et al., 2021). Counterspeech appears particularly powerful if many bystanders intervene (Schieb & Preuss, 2016). Thus, an eminent question in the fight against online hate speech is how uninvolved bystanders can be motivated to endorse and engage in counterspeech.

Prior counterspeech can increase subsequent bystander counterspeech in virtual settings (Garland et al., 2020; Miškolci et al., 2018). We propose that this effect could be partly due to the performance of social norms. The online environment is characterized by an often heightened salience and importance of social dynamics (e.g., Spears & Postmes, 2015). People adjust their tone and rhetoric to prior comments in online conversations (Rajadesingan et al., 2020), with prior comments from ingroup members exerting an especially strong influence (Seering et al., 2017). Therefore, we addressed whether social norms can be leveraged to encourage counterspeech.

With three online experiments (N_{total} = 1,948), we investigated whether perceptions of counterspeech as being normative for one's ingroup predicted own counterspeech endorsement and action and whether it did so over and above other candidate counterspeech predictors. We moreover tested whether counterspeech by ingroup members increased these perceptions. In the third experiment, we further extended our research scope and tested whether prior counterspeech in general, rather than neutral speech, increased participants' pro-counterspeech attitudes and behavior. In sum, our studies shed light on the psychological mechanisms underlying bystander interventions against online hate speech in general and on the role of ingroup norms in particular.

The Need for Counterspeech

Hate speech can be defined as communication that disparages or attacks its target based on central identity characteristics such as ethnicity, nationality, religion, or gender (United Nations, 2020). It is associated with a multitude of adverse effects, such as increased anxiety and depression in its victims (Tynes et al., 2008), prejudice against them (Soral et al., 2018; Weber et al., 2020), and even offline violence: in Germany, anti-refugee rhetoric on the social networking service (SNS) Facebook led to higher offline crime rates against refugees (Müller & Schwarz, 2021).

Fortunately, bystanders can effectively discourage further hate speech by engaging in counterspeech (Hangartner et al., 2021; Schieb & Preuss, 2016). We define counterspeech as any direct rejection of the transgression, such as openly criticizing the hate comment or expressing solidarity with the victim. It not only curbs future hate speech by the perpetrator (Bilewicz et al., 2021; Hangartner et al., 2021; Munger, 2017) but it might also positively shift online discourses on a broader level by inspiring further bystanders to speak up themselves (Cary et al., 2020; Garland et al., 2020; Miškolci et al., 2018; but see e.g.,

Leonhard et al., 2018 for conflicting results). Counterspeech against anti-Romani comments led to further pro-Romani comments on Facebook (Miškolci et al., 2018). Similarly, organized counterspeech on the SNS X (formerly known as Twitter) was positively associated with more counterspeech and less hate speech (Garland et al., 2020). Crucially, counterspeech is especially effective when done by multiple counterspeakers (Schieb & Preuss, 2016). To effectively leverage counterspeech in inspiring further bystander interventions and promoting a healthy discourse, it is essential to understand the mechanisms behind its influence.

Ingroup Norms Explaining Behavior

We proposed that prior counterspeech can prompt bystanders to intervene by presenting counterspeech as socially normative. Social norms define which attitudes and behaviors are common and endorsed in relevant groups or situations (Tankard & Paluck, 2016). Although not central in major bystander intervention frameworks (Halmburger et al., 2016; Latané & Darley, 1970), social norms have been highlighted as a potential but understudied tool against online hate speech (Bilewicz & Soral, 2020; Rudnicki et al., 2023). If social norms motivate bystanders to engage in counterspeech, norms of their relevant ingroups should be especially influential.

Ingroup norms strongly influence peoples' online behavior through cognitive and strategic processes that are heightened by characteristics of online environments (Klein et al., 2007; Reicher et al., 1995; Turner et al., 1987). According to Self-Categorization Theory, individuals try to adhere to the perceived norms of salient ingroups (Turner et al., 1987). The Social Identity Model of Deindividuation Effects further posits that in environments where individual traits are less visible, like many online spaces, stronger identification with group identities leads to greater adherence to ingroup norms (Reicher et al., 1995; Spears & Postmes, 2015). Strategic motives could further promote norm adherence as social identity performance (Klein et al., 2007). In the context of counterspeech, people could speak up for utilitarian reasons to mobilize further in- or outgroup members. If they perceive counterspeech as normative for their ingroup, people could also employ it as a form of virtue-signaling to consolidate their position within their ingroup or to show outgroup members that their ingroup endorses counterspeech (Saab et al., 2015). Lebanese Muslim participants, for instance, showed increased rejection of sectarian hate speech and increased endorsement of counterspeech when their religious leaders condemned the former (Siegel & Badaan, 2020). However, it remains unclear how well findings from a sectarian Lebanese context generalize and whether perceived ingroup norms, when directly assessed, are a driving mechanism for subsequent bystander counterspeech.

We thus hypothesized and tested that

The perception that counterspeech is representative of one's ingroup predicts one's own counterspeech endorsement and proliferation. This effect should remain above and beyond other candidate pro-counterspeech and anti-hate speech predictors (*H-Inspiration*).

Furthermore, the perception that counterspeech is representative of one's ingroup could be heightened by ingroup members engaging in counterspeech. Actions by ingroup members,

especially punishment of deviant behavior, strongly affect perceptions of ingroup norms (Tankard & Paluck, 2016). For instance, in virtual reality settings, bystanders reacted differently to the inaction of ingroup members compared to the inaction of outgroup members. Football club fans were less likely to intervene in a fight if other fans of their club, instead of unaffiliated bystanders, did not intervene (Rovira et al., 2021). We hypothesize that these results could be explained by the passive ingroup members signaling an ingroup norm of non-intervention.

We therefore predicted and tested that

The perception that counterspeech is representative of one's ingroup positively mediates an association between prior counterspeech by an ingroup member, compared to a non-ingroup member, and own pro-counterspeech behavior (*H-Mediation*).

The Present Research

In our three studies, participants interacted with mock social media content featuring hate speech and varied bystander responses. Across all studies, we investigated the role of ingroup norms in different scenarios by varying the identity of previous counterspeakers. Study 3 further expanded the scope of our investigation and examined whether prior counterspeech, compared to neutral speech, inspired more bystander counterspeech.

Pre-Registrations

Hypotheses, sample size determination via a priori power simulations, and confirmatory analyses were pre-registered (see *Supplement*). Regarding *H-Inspiration*, we pre-registered confirmatory tests for the effect of norm perceptions on bystander counterspeech endorsement and proliferation for all studies, and for Study 3, we also preregistered the candidate predictors against which we tested the incremental predictive power of norm perceptions. Moreover, we pre-registered confirmatory analyses for *H-Mediation* for all studies. The complete study materials, data, analysis scripts, and supplementary materials at https://osf.io/2b4dp/?view_only=1b7a07b2f4ef4b6faf542963b095561e

Study 1

In Study 1, we examined whether the perception that counterspeech is representative of one's ingroup was associated with more pro-counterspeech attitudes and behavior (*H-Inspiration*) against homophobic hate speech. Moreover, we examined the incremental predictive explanatory value of the perception that counterspeech is representative of one's ingroup against the following candidate predictors: own victimization (Obermaier et al., 2021), pro-target group attitudes (Weber et al., 2020), Right-Wing Authoritarianism (Altemeyer, 1981), and political orientation (Downs & Cowan, 2012)[3]. We also tested whether the perception that counterspeech is representative of one's ingroup mediated a positive association between prior counterspeech by a student from the participants' own

[3] Moreover, we tested if participants were differently affected by a human counterspeaker compared to an automated counterspeech entity. Prior research had shown that effects could even be observed for counterspeech by accounts who did not represent real human beings (Munger, 2017).

university, compared to a student from a competing university, and subsequent pro-counterspeech attitudes and behavior (*H-Mediation*).

Methods

Participants and Design

We collected data from students at a German university who received course credit for their participation. Of the 862 participants who saw the experimental manipulation, we excluded from the analyses anyone who failed the attention check or was not a student at the university, resulting in 673 observations. The age and gender distributions of the participants are displayed in *Table 1*. Sixty-seven participants did not indicate that they were heterosexual.

Participants were randomly assigned to conditions in a 2 (*Group*: ingroup vs. outgroup intervener) x 2 (*Nature*: human vs. non-human) between-subjects design[4].

Table 1

Age and Gender Distributions

		Study 1 (*N* = 673)		**Study 2** (*N* = 505)		**Study 3** (*N* = 866)	
		Mean	SD	Mean	SD	Mean	SD
Age		22.3	3.1	24.0	3.9	45.9	24.0
		N	%	N	%	N	%
Gender	female	368	54.7	268	53.1	480	55.4
	male	290	43.1	233	46.1	330	38.1
	n.a.	14	2.1	1	0.2	55	6.4
	other	1	0.1	3	0.6	1	0.1

Procedure

Once participants consented, they viewed a news headline on a mock social media site resembling Facebook. The headline discussed a competition between their own university and a rival one. A student from the participants' university commented on this headline, using a homophobic slur, which we blurred and marked as *homophobic content*[5]. Beneath this, a counterspeech reply read: “Watch your language. There is no place for homophobia.”

We manipulated the counter-speaker's affiliation, either affiliated with the participant's university (ingroup) or the rival one (i.e., another large university located in the same city; outgroup). This was highlighted using an introductory text, profile pictures, and usernames

[4] Effects of human compared to non-human interveners are not further investigated in this paper.

[5] We refrained from presenting actual homophobic content to minimize potential negative consequences for the participants.

that contained respective university initials. Moreover, we varied the nature of the counter-speaker as a student (human) or a chatbot (non-human) created by students from the respective universities.

Post exposure, participants indicated how they would respond in real life to a similar situation and completed additional attitudinal and demographic questions. They also had to recall the counter-speaker's university affiliation as an attention check.

Measures

If not stated otherwise, response options range from 1 (*strongly disagree*) to 7 (*strongly agree*).

Dependent Measures

Support. *Support* for the counterspeech comment was measured with six items (e.g., "I approve of the commenter's behavior," α = 0.87), partly adapted from Kutlaca et al. (Kutlaca et al., 2020).

Likes & Comments. Participants could indicate whether they would "like" any of the displayed comments (similar to the Facebook functionality), and they could write a comment themselves. Two independent raters coded each response as counterspeech (yes/no), and disagreements were resolved by the authors (see *Supplement* for coding instructions).

Ingroup Representativeness

The representativeness of counterspeech for one's ingroup was measured using six items (α = 0.84). Two items ("The message feels like it came from 'my people'." and "The message above reflects my group's values.") were adapted from Wolsko et al. (2016).

Control Variables

Political Orientation. Participants indicated their political orientation on one item (1 = "left" and 7 = "right).

Right-Wing Authoritarianism. Right-wing authoritarianism was measured with six items using the Very Short Authoritarianism Scale (Bizumic & Duckitt, 2018). Cronbach's α was low, however, at 0.64.

Pro-Target Group Attitudes. Attitudes towards LGB* rights were measured by asking participants to indicate how strongly they agreed with the statement, "People should be free to live their own life as they wish regardless of their sexual orientation."

Manipulation Checks

Closeness to the Intervener. Perceived personal closeness to the intervener was indicated by choosing between six pictograms that showed two circles where one circle represented the intervener and one the participant (0 = no overlap, 6 = complete overlap; adapted from Aron et al., 1993; Schubert & Otten, 2002).

Closeness to and between Universities. Perceived personal closeness to one's own university, the rival university, and between the universities were measured in the same way as closeness to the intervener, with circles representing the respective universities and, where applicable, the participant.

Results

Descriptives

The prevalence of likes for the prior counterspeech comment and of one's own counterspeech is displayed in *Table 2*.

Table 2

Distributions of Outcome Variables

	Study 1 (*N* = 673)		**Study 2** (*N* = 505)		**Study 3** (*N* = 866)	
	Mean	SD	Mean	SD	Mean	SD
Support	5.42	1.29	5.83	1.31	5.26	1.55
Counter Prob.					3.35	1.97
	Sum	%	Sum	%	Sum	%
Likes	264	39	322	64	193	45
Counterspeech 1	171	25	69	14	115	13
Counterspeech 2					145	17

Note. Response options for metric variables ranged from 1-7. *Likes* denotes indicated likes for a prior reply to hate speech. *Counterspeech 1* denotes own counterspeech. *Counterspeech 2* denotes own counterspeech against a second hate speech comment in Study 3 (see further below).

Descriptive statistics for social norm variables and alternative predictors are listed in *Table 3*.

Table 3

Distributions of Mediators and Alternative Predictors

	Study 1		**Study 2**		**Study 3**	
	Mean	SD	Mean	SD	Mean	SD
Social Norms						
Ingroup Representativeness	5.00	1.22	5.00	1.02	4.53	1.16
Ingroup Norm					4.92	1.04
Ingroup Identification	3.62	1.33	4.17	1.52	5.08	1.37
Alternative Counterspeech Predictors						
Pro-Target Att.	6.66	0.90	5.05	1.09	5.08[1]	1.15[1]
					4.85[2]	1.32[2]
Anti-Hate Att.			5.28	1.10	4.80	1.04
Political Orient.	3.76	1.01	3.20	1.08	3.40	1.08
Efficacy			4.31	1.60	4.02	1.61

	Study 1		Study 2		Study 3	
	Mean	SD	Mean	SD	Mean	SD
Right-Wing Authoritarianism	2.94	0.91				
Individualizing Moral Foundations					5.82	0.77
Solidarity Norms					5.06	1.20

[1]Czechs; [2]Faroe Population

Manipulation Check

First, we tested whether the experimental manipulation achieved the desired effect that participants feel closer to the ingroup counterspeaker. A Welch two-sample t-test between the ingroup and the outgroup condition yielded no differences in closeness to the counterspeaker ($M_{Ingroup}$ = 3.76 (*SD* = 1.55); $M_{Outgroup}$ = 3.69 (*SD* = 1.47); $t(669) = -0.58$, $p = .563$). Identification with one's own university was significantly positively correlated with identification with the rival university, $r(670) = 0.34$, $p < .001$). Similarly, identification with one's university was significantly positively correlated with the perception that both universities are close to each other, $r(670) = 0.20$, $p < .001$. In summary, participants did not feel closer to a counterspeaker from their own university, and their own perceived distance to the other university and distance between universities even decreased the more participants identified with their own university.

Mediation Hypothesis

The manipulation checks indicated that the manipulation was not effective. Consequently, we did not find any effects of the manipulation on the outcome variables or perceptions of ingroup representativeness (*see Supplement*).

Ingroup Representativeness on Outcomes

However, Pearson's product-moment correlation for *Support* and point-biserial correlations for the other outcomes yielded that *Ingroup Representativeness* was positively associated with all outcomes (*Support*: $r(671) = 0.71$, $p < .001$; *Likes*: $r(671) = 0.38$, $p < .001$; *Comments*: $r(671) = 0.20$, $p < .001$).

Ingroup Representativeness versus Other Predictors

We explored whether Ingroup Representativeness predicted Support, Likes, and Comments when adding Right-Wing Authoritarianism, Pro-LGB* Attitudes, Own Sexual Orientation, and Political Orientation as additional predictors in a linear regression and logistic regressions, respectively. Ingroup Representativeness remained a significant predictor for all three dependent variables (Support: $\beta = 0.66$, SE = 0.03, $p < .001$; Likes: $\beta = 0.86$, SE = 0.86, $p < .001$; Comments: $\beta = 0.50$, SE = 0.11, $p < .001$; see Table S.1 in Supplement for full results).

Discussion

In line with *H-Inspiration*, we found that the perception that counterspeech is representative of one's ingroup predicted own pro-counterspeech attitudes and hypothetical behavior against homophobic hate speech. This association remained when controlling for established pro-counterspeech and anti-hate speech predictors. However, we did not find that our participants contrasted their own university and the competing

university as ingroup and outgroup. Instead, the students who most strongly identified with their own university also most strongly identified with the competing university and perceived people from the two universities to be more similar. These findings suggest that one overarching student identity could have been salient rather than two distinct university identities. Consequently, we did not observe an effect of the group affiliation manipulation on our counterspeech measures or *Ingroup Representativeness* and, therefore, no support for *H-Mediation*.

Study 2

To overcome limitations of Study 1, in Study 2, we selected a different context, a different target group, and included measures of actual, rather than hypothetical behavior. First, we contrasted a student ingroup counterspeaker with an unaffiliated counterspeaker to avoid participants categorizing both interveners in a superordinate *student* category. In addition, we chose Czechs as a target group whose perception is less affected by political orientation than sexual minorities. Moreover, we aimed to measure actual instead of hypothetical counterspeech by adding a cover story, which implied that comments would be visible to other participants. Bystander intervention in hypothetical and real scenarios can be predicted by different dispositions (Baumert et al., 2013).

We tested *H-Inspiration* and *H-Mediation* for the dependent variable *Support*. Due to power concerns, we did not pre-register confirmatory but exploratory analyses regarding *Likes* and *Own Counterspeech*. Moreover, we once more explored whether the predicted effect of ingroup representativeness remained robust after controlling for other theoretically relevant explanatory variables, this time the effects of pro-target group (Obermaier et al., 2021) and anti-hate speech attitudes (Weber et al., 2020), for the perceived efficacy of individual anti-hate speech measures (Lumsden & Morgan, 2017; Obermaier, 2022), and for political orientation (Downs & Cowan, 2012).

Methods

Participants and Design

Participants were recruited from a German university in return for a small monetary compensation. Of the 629 participants who saw the stimulus material, we analyzed data from 505 subjects who passed the comprehension tests. The age and gender distributions of the final sample can be found in *Table 1*. None of the participants indicated that they had a Czech nationality. Twelve participants did not indicate that they lived in Germany.

We randomly assigned participants to one of two experimental conditions (*Group*: ingroup affiliation vs. no affiliation).

Procedure

As a cover story, participants were informed that they would interact with individuals from various German institutions for a virtual communication study. After consenting, they selected a username and profile picture showcasing their university initials. They then viewed three mock news headlines on a platform resembling Facebook. The initial two

headlines had neutral content. The third addressed a rise in COVID-19 cases in Czechia[6]. Beneath it, there was a hate speech comment by an anonymous user targeting Czechs and a counterspeech reply. This reply either came from someone whose profile picture displayed the participants' university initials (ingroup) and who rejected the hate speech and mentioned that such behavior was not normative for their university or from an unaffiliated anonymous user who only voiced his individual disagreement with the hate speech without referencing any group identity (unaffiliated).

Participants could "like" or comment on each headline after being instructed that their responses might be shown to others. After the headlines, they answered two comprehension questions regarding the hate speech comment: its tone towards Czechs and whether the subsequent comment agreed with it. Participants then completed attitudinal and demographic measures before being debriefed.

Measures

If not stated otherwise, response options range from 1 (*strongly disagree*) to 7 (*strongly agree*).

Dependent Variables

Support (α = 0.92)*, Likes,* and *Comments* were measured as in Study 1 (see *Table 2* for summary statistics). Due to a data collection error, the comments of the first ten participants could not be analyzed. For a more economic measure, one item for *Support* with the least favorable psychometric properties was dropped (*see Supplement* for Log File).

Ingroup Representativeness

Ingroup Representativeness (α = 0.78) was measured as in Study 1. Compared to Study 1, we replaced the rather unspecific term "my group" with "[university name] students" to reduce ambiguity regarding the relevant group.

Ingroup Identification

The participants' identification with their university was indicated with the same closeness measure as in Study 1 and four further items adapted from Doosje et al. (1995). The answers were z-transformed and combined into one index (α = 0.87).

Pro-Target Group Attitudes

Attitudes towards Czechs were measured with four items (e.g., "In general, the Czech population is friendly"; α = 0.94) adapted from Cuddy and colleagues (2009).

Anti-Hate Speech Attitudes

We used four items to measure rejection of group-based hate (e.g., "Posting hate comments against other groups of people in virtual space can be as bad as physical violence."), that were combined into a heterogeneous index (α = 0.55).

Efficacy

Two items, adapted from van Zomeren et al. (2013), measured the perceived efficacy of participants in rejecting hate speech (*r* = 0.83).

[6] That pandemic was highly relevant in the spring of 2021 when the study was conducted.

Political Orientation

Participants indicated their political orientation as in Study 1.

Manipulation Check: Intervener Group

As a manipulation check, the belonging of the intervener to the participants' university was measured using six pictograms that showed two circles with varying degrees of overlap (adapted from Aron et al., 1993; Schubert & Otten, 2002).

Credibility check: Audience Size

Participants indicated how many other people they expected to be able to see their potential comments.

Results

See *Tables 1*, *2*, and *3* for descriptive results.

Manipulation and Credibility Checks

On average, participants expected 80 others to be able to see their potential comments (*SD* = 20). No participants indicated that their potential comments would not have been visible to anyone else.

We tested whether the affiliation manipulation had worked as planned by computing a two-sided Welch two-sample t-test. Results showed that participants considered the counter speaker to belong to their university, irrespective of the condition ($M_{Ingroup}$ = 4.65 (*SD* = 1.35) vs. $M_{Outgroup}$ = 4.78 (*SD* = 1.50); *t*(452) = -0.95, *p* = .345)).

Mediation Hypothesis

Consistent with the non-significant manipulation checks, we did not find any effects of the manipulation on the outcome variables or perceptions of ingroup representativeness (*see Supplement*).

Ingroup Representativeness on Outcomes

Pearson's product-moment correlation for *Support* and point-biserial correlations for the other outcomes yielded that *Ingroup Representativeness* was positively associated with *Support* and *Likes* (*Support*: *r*(503) = 0.32, *p* < .001; *Likes*: *r*(503) = 0.22, *p* < .001; *Comments*: *r*(493) = 0.08, *p* = .074).

Ingroup Representativeness versus Other Predictors

Using linear and logistic regressions, we explored whether Ingroup Representativeness explained variance of Support, Likes, and Comments in addition to Efficacy, Anti-Hate Speech Attitudes, Pro-Target Group Attitudes, and Political Orientation. *Ingroup Representativeness* had a significant positive association with *Support* (β = 0.16, *SE* = 0.05, *p* = .002) and *Likes* (β = 0.31, *SE* = 0.11, *p* = .008) but not for *Comments* (β = 0.04, *SE* = 0.16, *p* = .786; *see Table S.2 in Supplement for full results*).

Discussion

Study 2 again supported *H-Inspiration*, showing that perceiving counterspeech as representative of one's ingroup predicted one's own pro-counterspeech attitudes and behavior against hate speech, this time against Czechs. This perception explained support and likes for counterspeech over and above established attitudinal predictors. As the

incremental predictive power was insufficient, we did not find the same effect for own counterspeech.

As the manipulation checks showed, unexpectedly and similar to Study 1, participants associated both the ingroup counterspeaker and the unaffiliated counterspeaker equally with their own university, meaning that we did not successfully create a meaningful ingroup outgroup distinction between our experimental conditions. In line with this, we again did not find an effect of the manipulation on the outcome variables or counterspeech representativeness and thus no support for *H-Mediation*. Since no identifying markers were provided for the counterspeaker in the unaffiliated condition, participants possibly had no indication that he belonged to a different group than themselves. Another factor could have been the insufficient salience and importance of one's own university as a social group by the subjects and in the German context in general. Participants only indicated identification with their own university at .17 points above the midpoint on a seven-point scale. As a result, participants could have had little motivation to properly discern ingroup members from unaffiliated members. We addressed this shortcoming again in Study 3.

Study 3

In the first two studies, we consistently found that perceived counterspeech endorsement of one's ingroup was related to own counterspeech endorsement. However, we did not find an effect of counterspeaker affiliation on that perception or own counterspeech endorsement and proliferation. This could either have been due to an insufficient distinctiveness between the counterspeakers' social identities or an insufficient effect of counterspeech on the assessed outcomes. In other words, either our experimental manipulations did not cause participants to perceive the ingroup and the outgroup counterpeakers differently, or prior counterspeech, in general, could simply not have affected further bystanders.

Supporting the former, participants showed a general tendency to associate any counterspeaker with their social ingroup both in the context of different university affiliations and of a missing affiliation of the outgroup counterspeaker. In Study 3, we, therefore, contrasted two other groups, namely students and retirees, which do not share a direct superordinate group to the same extent as observed in Study 1, have few members that belong to both groups and are regularly salient and relevant as distinct groups in everyday life.

Conversely, it can be questioned whether counterspeech has an effect at all motivating further counterspeech. The evidence for such an inspiring effect of prior counterspeech stems from observational and self-report studies in real-world online contexts (Cary et al., 2020; Miškolci et al., 2018). The reported effects could have been due to other reasons than the counterspeech's appeal. Instead, counterspeech replies in online contexts could increase the visibility of hate speech comments via platform algorithms that promote posts with more engagement. Increased counterspeech could thus be the result of an increased audience size instead of actual behavioral change. Additional factors like biased recall could have driven the effects in self-report studies. Controlled, hypothetical settings often failed to find positive effects of counterspeech on further bystanders (Leonhard et al., 2018;

Obermaier et al., 2021). For instance, observing counterspeech against anti-refugee hate speech in fictitious Facebook posts did not impact subsequent bystander counterspeech (Leonhard et al., 2018). Studies in hypothetical settings could have failed to detect effects on subsequent bystander counterspeech due to their assessment of behavioral intentions rather than actual behaviors. These two types of outcomes are generally determined by different psychological aspects (Banyard & Moynihan, 2011; Baumert et al., 2013; A. L. Brown et al., 2014). We, therefore, broadened our research scope by investigating whether counterspeech replies more strongly inspired subsequent bystander endorsement and proliferation of counterspeech. We predicted that in a controlled setting that measured actual behaviors instead of behavioral intentions,

Bystanders who see prior counterspeech against hate speech, compared to neutral speech, engage in more pro-counterspeech behavior (*H-Emulation*).

We also tested *H-Inspiration*, that the perception of pro-counterspeech ingroup norms explains pro-counterspeech attitudes and behavior and that it does so over and above other candidate predictors - solidarity norms (Kunst et al., 2021), moral foundation endorsement (Wilhelm et al., 2020), individual effic*acy* (Lumsden & Morgan, 2017; Obermaier, 2022), attitudes towards hate speech and attitudes towards the target group (Cowan & Hodge, 1996; Weber et al., 2020), and political orientation (Downs & Cowan, 2012). For that, we applied an ingroup norm measure that assessed how representative, in general, people perceive counterspeech for their respective ingroups. Moreover, we again tested *H-Mediation* as in the previous studies but not for own counterspeech comments due to power concerns (see pre-registration in *Supplement*).

As a further extension of our research scope, we added two outcome measures to examine how bystander reactions to hate speech in subsequent encounters and predictions to engage in counterspeech outside of the study are affected by witnessing ingroup and outgroup counterspeech. In addition to Czechs, we assessed reactions to hate speech against people from the Faroe Islands, a Danish territory, as another group whose endorsement is generally not associated with political orientation.

Methods

Participants & Design

Initially, 1135 participants saw the experimental manipulation. Two hundred sixty-nine participants were excluded for incorrectly answering comprehension questions, resulting in a sample of 866 participants. Age and gender distributions can be found in *Table 1*. Participants were recruited via a panel provider and participated for a small financial compensation. Four hundred fifty-nine of our participants were students and 407 were retired with no overlap between groups. No participants indicated a Czech or Danish nationality. Germany was the primary place of residence of 813 participants; two lived outside of Germany, and 51 did not answer the question. Participants were randomly assigned to one of four conditions in a 2(*Group*: ingroup vs. outgroup) X 2(*Response Type*: neutral vs. counterspeech) design.

Procedure

Participants were informed that they were part of a study on inter-generational online communication. After giving consent, they chose a username, identified their occupation (student, employed, retired, other), and received an occupation-specific profile picture. They then viewed four news headlines with comments that were ostensibly posted by other users. Three news headlines were the same as in Study 2[7]. In addition, participants saw a headline about whale hunting on the Faroe Islands. The headlines about Czechia and the Faroe Islands had been replied to with hate speech comments against the respective populations, while neutral replies accompanied the two other headlines.

First, three news headlines were displayed in a random sequence, including one of the headlines with a hate speech reply. This reply came with a response from a student or a retiree, which constituted either the ingroup or the outgroup condition, depending on the participant's occupation. This response was either counterspeech (counterspeech condition) or a neutral comment not addressing the hate speech (neutral condition).

Subsequently, the second headline with a hate speech reply was displayed but was not followed by any further replies. The presentation order of the headlines with hate speech replies against Czech and Faroese people was balanced between conditions. Participants could "like" comments or respond to them. Afterward, they answered questions about their perceptions and attitudes before being debriefed.

Measures

If not stated otherwise, response options range from 1 (*strongly disagree*) to 7 (*strongly agree*).

Dependent Variables

Support (α = 0.90)*, Likes,* and *Comments* were measured as in Study 2. We assessed participants' own counterspeech comments (henceforth called *Comments 1* and *Comments 2* for replies to the first and second hate speech comment, respectively). Also, participants indicated their *Counterspeech Probability,* the likelihood with which they would engage in counterspeech against hate speech if they encountered it in their daily lives with three items (e.g., "How likely is it that you write a comment that criticizes the hate comment?", α = 0.90, see *Table 2* for summary statistics).

Ingroup Representativeness

Ingroup Representativeness was measured as in the previous studies (α = 0.79), with the group of reference being students for student participants and retirees for retired participants (e.g., "In general, students/retirees would act similarly").

Ingroup Norm

In addition to our measure of ingroup representativeness of a concrete reaction to hate speech, we also measured how representative counterspeech is seen in general to achieve a similar abstraction level to other candidate predictors. Participants indicated how representative, in general, they perceived counterspeech to be for their respective ingroups.

[7] As a slight variation, the headline talking about Czechia did not mention COVID-19 but multi-resistant germs.

They were told to imagine someone using counterspeech against online hate speech and asked to indicate how representative such behavior would be for their respective group, using the same scale as for *Ingroup Representativeness* (α = 0.79).

Ingroup Identification

Participants' identification with their ingroup was indicated with only the six pictograms used in Study 2 (α = 0.91).

Pro-Target Group Attitudes

Attitudes towards the hate speech target groups were measured as in Study 2 (both αs = 0.97).

Anti-Hate Speech Attitudes

Participants indicated how much they rejected group-based hate with four items. Albeit showing low internal consistency (α = 0.43), we combined the variables into a heterogenous index.

Efficacy

Individual efficacy was measured as in Study 2 (α = 0.93).

Moral Foundations Endorsement (Individualizing)

Endorsement of the individualizing moral foundations (*Care/Harm* and *Fairness/Cheating*) was measured with six items each, using the Moral Foundations Questionnaire (Graham et al., 2013) (α = 0.87).

Solidarity Citizenship Norms

Participants answered seven items (α = 0.91) to indicate how strongly they associate showing solidarity with others with being a good citizen (Kunst et al., 2021).

Political Orientation

Participants indicated their political orientation as in the previous studies.

Credibility check: Audience Size

Participants indicated how many other people they expected to be able to see their potential comments.

Results

On average, participants expected 48 others to be able to see their potential comments (*SD* = 56). Only 3 participants indicated that their potential comments would not be visible to anyone else.

H-Emulation: Counterspeech versus Neutral Speech

First, we conducted linear and logistic regressions for the five dependent variables *Support*, *Likes*, *Comments 1*, *Comments 2*, and *Counterspeech Probability* with the predictors *Group* (0 = ingroup intervener, 1 = outgroup intervener), *Response Type* (0 = counterspeech, 1 = neutral speech), and their interaction. P-values were adjusted for quintuple testing. Results are displayed in *Table 4*. We observed a significant negative main effect of *Response Type* on *Support*, *Likes*, and *Comments 2*, meaning that in the ingroup condition, neutral speech was associated with smaller values on these outcomes. We did not observe any main effects for *Group* or interactions between the two predictors.

We moreover computed contrasts to explore the overall effects of the different predictors using the marginaleffects package (Arel-Bundock, 2023). That is, we computed the average difference between the two values of *Response Type* and of *Group* for the predicted value of the criterion and estimated standard errors using the delta method. Results yielded a negative overall effect of *Response Type* for all outcomes except for *Counterspeech Probability* (*see Supplement*). When we controlled for *Occupation* (0 = student, 1 = retiree) of the participant and *Target* (0 = Czechs; 1 = Faroe Islanders), we found that *Likes* and *Comments 1* were affected by the bystander's occupation and the targeted group's occupation. Nevertheless, contrasts corroborated our findings of an overall effect of neutral speech vs. counterspeech for both outcomes (*see Supplement*).

Table 4

H-Emulation Regression Results

	Support		Likes		Comments 1		Comments 2		Count. Probability	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Inter-cept	0.246***	0.070	-0.147	0.136	-1.404***	0.170	-1.238***	0.162	0.006	0.071
Group	-0.002	0.097	-0.085	0.195	-0.633	0.276	-0.430	0.250	-0.094	0.100
Resp. Type	**-0.473*** **	**0.094**	**-1.742*** **	**0.236**	**-0.485**	**0.257**	**-0.544***	**0.246**	**-0.024**	**0.097**
Group X Resp. Type	0.000	0.135	0.254	0.336	0.124	0.421	0.415	0.370	0.216	0.138
Num. Obs.	835		866		866		866		839	
R2	0.052		0.118		0.015		0.008		0.001	

* p < 0.05, ** p < 0.01, *** p < 0.001

R^2 values are computed for linear regressions using adjusted R^2 and for logistic regressions using Tjur's R^2.

H-Inspiration: Ingroup Representativeness on Outcome Variables

Furthermore, we tested the effect of *Ingroup Representativeness* on *Support*, *Likes*, and *Counterspeech Probability*. As outlined in our pre-registration, we did not include *Comments 1* and *Comments 2* in these analyses due to power concerns. For *Likes*, we conducted a logistic regression; for the other two outcomes, we conducted linear regressions. For each regression, we used *Group* (0 = outgroup intervener, 1 = ingroup intervener), *Ingroup Representativeness*, and their interaction as predictors. The results of the logistic

regressions are displayed in *Table 5*. P-values for the predictors were adjusted for triple testing.

Table 5

H-Inspiration Regression Results

	Support		Likes		Count. Probability	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	0.135	0.068	-0.110	0.153	0.031	0.074
Ingr. Rep.	**0.324*** **	**0.066**	**0.317***	**0.152**	**0.199** **	**0.072**
Group	**-0.391*** **	**0.099**	-0.318	0.233	-0.067	0.108
Ingr. Rep. X Group	**0.289***	**0.100**	**0.578***	**0.253**	0.010	0.108
Num.Obs.	398		398		398	
R2	0.183		0.068		0.030	

* p < 0.05, ** p < 0.01, *** p < 0.001

R^2 values are computed for linear regressions using adjusted R^2 and for logistic regressions using Tjur's R^2.

Ingroup Representativeness significantly predicted the three dependent variables. Moreover, we observed significant positive interactions between *Ingroup Representativeness* and *Group* for *Support* and *Likes*, meaning that *Ingroup Representativeness* had a bigger effect on the outcomes in the ingroup than in the outgroup condition. Finally, *Group* negatively predicted *Support*. Adding *Occupation* and *Topic*, as well as their interactions as control variables, yielded a significant negative interaction between *Ingroup Representativeness* and the Faroe scenario (β = -1.21, *SE* = 0.49, *p* = .039) for *Likes*. Nevertheless, contrasts confirmed an overall positive effect of *Ingroup Representativeness* on *Likes* (*see Supplement*).

H-Inspiration: Ingroup Norms versus Other Predictors

We also tested if Ingroup Norm predicted Support, Likes, Comments 1, Comments 2, and Counterspeech Probability in addition to Solidarity Norms, Moral Foundation Endorsement, Individual Efficacy, Attitudes Towards Hate Speech, Attitudes Towards the Target Group, and Political Orientation. We used linear regressions for Support and Counterspeech Probability and logistic regressions for the other dependent variables. We also included Occupation and Target as control variables. P-values for the predictors were adjusted for quintuple testing.

Ingroup Norm predicted all outcome variables except for *Likes* which were only predicted by the target group (*Support*: β = 0.28, *SE* = 0.09, *p* = .001; *Likes*: β = 0.07, *SE* = 0.13, *p* = .580; *Comments 1*: β = 0.33, *SE* = 0.12, *p* = .007; *Comments 2*: β = 0.36, *SE* = 0.11, *p* = .002; *Counterspeech Probability*: β = 0.13, *SE* = 0.04, *p* = < .001). Other established predictors did not consistently explain variance across outcome variables (see *Table S.3* in *Supplement* for full results).

H-Mediation: Ingroup Representativeness as a Mediator

Following, we tested if counterspeech by an ingroup member (vs. an outgroup member) affected perceptions of how representative counterspeech is for one's ingroup using the participants in the counterspeech condition. A multiple linear regression with *Group* (0 = outgroup intervener, 1 = ingroup intervener)[8] as the predictor and *Ingroup Representativeness* as the criterion yielded a significant positive effect for *Group* (β = 0.81, *SE* = 0.18, $p < .001$). No other predictors were significant ($p > .152$) The overall regression significantly predicted *Ingroup Representativeness* ($F(7, 390) = 11.05$, $p < .001$, $R^2_{adjusted}$ = 0.15). Adding *Occupation* and *Topic*, as well as their interactions as control variables, did not change the pattern of the results (*see Supplement*).

For each of the three dependent variables, *Support*, *Likes*, and *Counterspeech Probability*, we used the mediation package (Tingley et al., 2014) to compute the indirect effect of *Group* on the respective dependent variable via *Ingroup Representativeness* using bootstrapping with 10,000 iterations. Again, analyses were only conducted for participants in the counterspeech condition and *Group* was coded as 0 = outgroup intervener and 1 = ingroup intervener. P-values for the predictors were adjusted for triple testing.

For each dependent variable, we observed the pattern of a significant positive average causal mediation effect, a negative average direct effect, and a resulting non-significant total effect, indicating an inconsistent mediation (*Support*: *ACME* = 0.57, $p < .001$, 95%CI[0.41, 0.77]; *ADE* = -0.63, $p < .001$, 95%CI[-0.90, -0.30]; *Total* = -0.05, $p = .848$, 95%CI[-0.33, 0.28]; *Likes*: *ACME* = 0.11, $p < .001$, 95%CI[0.07, 0.15]; *ADE* = -0.06, $p = .226$, 95%CI[-0.16, 0.04]; *Total* = 0.05, $p = .409$, 95%CI[-0.06, 0.14]; *Counterspeech Probability*: *ACME* = 0.31, $p < .001$, 95%CI[0.14, 0.51]; *ADE* = -0.13, $p = .540$, 95%CI[-0.55, 0.29]; *Total* = 0.18, $p = .346$, 95%CI[-0.21, 0.56]).

Exploratory Analyses

We explored whether the mediation could still be observed after controlling for *Outgroup Representativeness* and *Hate Speech Attitudes*, finding that the indirect effect remained for all three outcome variables. Further analyses for sequential ignorability (Imai et al., 2010) indicated that another unobserved variable correlated with the mediator and the outcome variable might cause the observed inconsistent mediation for *Counterspeech Probability* (*see Supplement*) but not *Support* and *Likes*.

We further explored the effect of *Ingroup Representativeness* on *Comments 1* and *Comments 2* using logistic regressions with *Group* (0 = outgroup intervener, 1 = ingroup intervener), *Ingroup Representativeness*, and their interaction as predictors. Neither *Comments 1* nor *Comments 2* were predicted by any of the variables (*see Supplement)*.

Discussion

We observed that counterspeech responses, compared to neutral responses, led to more support and likes for the response and to more own counterspeech against the same hate speech and consecutive hate speech. It did not increase the self-reported probability of engaging in counterspeech outside the study. We also confirmed that perceptions of pro-

[8] We reversed the coding of *Group* to facilitate interpretation of the results.

counterspeech ingroup norms predict pro-counterspeech attitudes and behavior and that they do so over and above other candidate predictors.

In addition, we found that counterspeech by an ingroup member positively affected the perception of pro-counterspeech ingroup norms. This led to a positive indirect mediation effect of ingroup counterspeech on support and likes for counterspeech and willingness to engage in counterspeech outside of the study. However, a parallel direct mechanism attenuated this effect, leading to an inconsistent mediation. We, therefore, conducted exploratory tests to determine whether our findings could be explained by potential confounding variables correlated with *Ingroup Representativeness* and the outcome variables. The indirect effect of ingroup norms remained after controlling for pro-counterspeech norms of the outgroup and for a general rejection of hate speech. Additional sensitivity analyses indicated no issues for counterspeech support and likes. However, they yielded that the association between ingroup norms and the indicated probability of speaking up outside of the study was likely a statistical artifact (Imai et al., 2010).

Chapter Discussion

Bystander counterspeech can be a powerful tool against online hate speech (Garland et al., 2020), especially if done by a multitude of bystanders (Schieb & Preuss, 2016). Here, we addressed the question of what motivates bystander counterspeech. In this line of studies, we investigated the influence of prior counterspeech and pro-counterspeech ingroup norms on decisions to speak up. We observed in Study 3 that prior counterspeech against hate speech, as opposed to a neutral response, significantly increased subsequent bystander counterspeech against the same hate speech and even different hate comments. Moreover, across the three studies, we consistently found that perceptions of a pro-counterspeech ingroup norm predicted pro-counterspeech attitudes and behavior across multiple hate speech instances and social groups over and above other counterspeech and hate speech predictors. We also found that these perceptions positively mediated the relationship between prior ingroup counterspeech and pro-counterspeech attitudes and behavior, but only if the ingroup and outgroup were clearly distinct - as was the case in Study 3.

Prior Counterspeech Versus Neutral Speech and Pro-Counterspeech Outcomes

Our findings align with previous studies that assessed behavior on social media (Miškolci et al., 2018), showing that prior counterspeech, compared to neutral replies, increases subsequent counterspeech endorsement and generation by bystanders. We extend these findings by showing that prior counterspeech indeed exerts a causal effect by inspiring further bystanders to speak up.

However, we did not find an effect on the self-rated likelihood to engage in counterspeech outside of a study setting. This finding is consistent with studies that assessed hypothetical subsequent bystander counterspeech (Leonhard et al., 2018; Obermaier et al., 2021) in controlled settings. It thus seems plausible that the heterogeneous effects of prior counterspeech could be due to the nature of the outcome variables. Behavioral intentions and actual behavior often diverge in the context of bystander interventions and are predicted by different dispositions (Baumert et al., 2013). This notion was further supported by our findings that pro-counterspeech ingroup norms positively predict counterspeech

endorsement and own actual counterspeech but not on participants' intentions to engage in counterspeech. Our findings underline the necessity of appropriate outcome measures when assessing the actual effects of online counterspeech.

Ingroup Norms and Pro-Counterspeech Outcomes

Suggesting ingroup norms as an important psychological mechanism of counterspeech, we found that perceptions of counterspeech as representative of one's ingroup were consistently linked to increased counterspeech endorsement and engagement across different hate speech targets and ingroups. We moreover confirmed exploratory findings in Studies 1 and 2 with preregistered hypothesis tests in Study 3 showing this effect persisted even after controlling for an array of established predictors such as political orientation (Downs & Cowan, 2012) or attitudes toward the hate speech target group (Cowan & Hodge, 1996; Weber et al., 2020).

As hypothesized, Study 3 yielded evidence that perceptions of counterspeech as representative of one's ingroup positively mediated the association between counterspeech from an ingroup member and one's own endorsement of counterspeech. These results extend previous research, which suggested an important role of social norms without directly assessing them (Siegel & Badaan, 2020). We are the first to show that counterspeech can indeed motivate others to speak up specifically through its effect on ingroup norm perceptions.

However, this effect occurred only when ingroups and outgroups were highly distinct, as in Study 3 and not in Studies 1 and 2, and even then, the counterspeaker's identity did not have an overall effect on behavior. Participants showed a general tendency to see counterspeakers as ingroup members. While people can be seen as outgroup members based on minimal differences (Tajfel, 1974), especially in online contexts (Tepper, 1997), our findings suggest that people who oppose hate speech are often sorted into one's ingroup. In Study 1, participants perceived no difference in closeness to a counterspeaker from their own compared to a rival university, likely due to a perceived overarching *student* identity. In Study 2, participants assumed that an unaffiliated commenter belonged to their university just as firmly as a commenter who was clearly affiliated with it through his profile picture and comment content. Only when groups were strongly distinct, students and retirees, did participants make a distinction between the commenters' group identities.

Nevertheless, we did not observe a significant total increase in counterspeech endorsement after seeing counterspeech from an ingroup member rather than an outgroup member, even in that case. The positive mediation effect of perceived pro-counterspeech ingroup norms was suppressed by a negative direct effect of ingroup intervention on counterspeech endorsement. Two conflicting goals of pro-counterspeech behavior could cause these incongruent effects (Klein et al., 2007; Saab et al., 2015). In addition to the utilitarian goal of fighting hate speech, counterspeech endorsement could be performed to consolidate one's identity. After seeing counterspeech from an ingroup member, participants could have been motivated to signal their ingroup affiliation by showing that they individually endorse counterspeech. On the contrary, after seeing counterspeech from an outgroup member, participants could have been motivated to consolidate their ingroup's position relative to

the outgroup by showing that their own ingroup also endorses counterspeech (Saab et al., 2015). Thus, while ingroup counterspeech could have increased the likelihood of counterspeech endorsement for individual identity consolidation, it could have decreased the likelihood of counterspeech endorsement for group identity consolidation towards the outgroup. Further research could directly assess people's motivations for counterspeech endorsement to test these speculations.

Implications

We found that counterspeech inspires further bystanders to also engage in counterspeech against the same hate speech and even against subsequent hate comments. Our findings thus suggest counterspeech as a potentially powerful tool to combat online hate speech. In the context of online hate speech, the phrase "Do not feed the troll" is frequently invoked, implying that counterspeech is futile. Our findings contradict this notion, showing that counterspeech can serve an important purpose – not necessarily by changing the hate speaker's mind but by nudging other bystanders to speak up, too.

Our findings, moreover, show that perceptions of pro-counterspeech ingroup norms can be a powerful motivator for bystander counterspeech. Higher perceptions of pro-counterspeech ingroup norms were consistently associated with more own counterspeech and counterspeech endorsement. These findings highlight the importance of extending existing psychological models explaining bystander interventions (Halmburger et al., 2016; Latané & Darley, 1970) and suggest leveraging ingroup norm perceptions to combat online hate speech effectively. We find that ingroup norms could perhaps be leveraged through prior counterspeech. Our studies showed that even ostensible outgroup members who engage in counterspeech can be as successful at shaping pro-counterspeech ingroup norms as ingroup members – as long as groups are not too distinct. However, it is crucial to consider the motives and audience at play when bystanders decide whether to engage in counterspeech. In Study 3, we found that a positive effect of ingroup norms was attenuated through concurrent mechanisms, possibly through a person's desire to engage in counterspeech to enhance either their own standing within their group or their group's standing towards other groups (Klein et al., 2007; Saab et al., 2015). Our results thus indicate that a person's reason to engage in counterspeech over and above purely utilitarian reasons should be considered when trying to leverage ingroup norms through prior counterspeech.

Limitations

While our studies encompassed many aspects of real social media settings, they were limited in two core aspects. Users interacted with our materials only once and with assumed strangers. Interactions on real SNS are characterized by the possibility of multiple encounters and interactions with important and known others such as friends, family, or colleagues. The effect of social norm adherence could be increased when people expect multiple interactions with others (Ellingsen & Johannesson, 2008; Xiao & Houser, 2009) and when others' evaluations can also have offline consequences.

Also, seeing a single counterspeech comment could simply have been too little to affect people's behavior substantially. Rather, perceptions of ingroup norms could be affected

more strongly by multiple observations of hate and counterspeech – either by multiple commenters or across an extended time period. Thus, future studies addressing the longer-term effects of repeated encounters would be pronouncedly valuable.

Conclusion

Our findings showed the significant influence of social norms on counterspeech against online hate speech. The perception that counterspeech is representative of one's ingroup may serve as a strong predictor for one's own engagement in counterspeech. Prior counterspeech can positively shape these norm perceptions. These findings not only present a promising path for further research on bystander interventions beyond the factors that are outlined by existing major frameworks (Halmburger et al., 2016; Latané & Darley, 1970) but also support efforts by civic society to confront online hate speech through the active use of counterspeech (Ley, 2018).

Chapter D – Counterspeech on Twitter: Effects of Ethnicity and Status

Across the globe, the prevalence of online hate speech, specifically on social media, poses significant societal challenges. Hate speech can be defined as communication that attacks or disparages its target based on central identity characteristics such as race, religion, gender, or other identity factors (United Nations, 2020). Online hate speech causes a plethora of harmful consequences, such as the marginalization of vulnerable groups (Bilewicz & Soral, 2020), the deterioration of communication norms (Garland et al., 2020), and even offline violence (Müller & Schwarz, 2021). The widespread presence and proliferation of hate speech regularly leaves its victims and sympathetic bystanders feeling powerless to counter the tide of online hate (Coles & Lane, 2023; Nadim & Fladmoe, 2021).

Efforts by social media providers and governments to reduce hate speech have so far been centered around its deletion. However, in addition to raising censorship concerns (Strossen, 2020), deletion has not been able to keep up with the sheer amount of online hate (FRA, 2023) and does not discourage transgressors from further offenses (Jiménez Durán, 2022). More recently, user-generated counterspeech has been highlighted as a potentially powerful supplement to deletion. Counterspeech can be defined as communication that openly confronts and rejects a transgression. It has been suggested as a useful way for members of victimized groups to make their voices heard (Ozalp et al., 2020) and for unaffected bystanders to support victims of hate speech (Leets, 2002; Sasse et al., 2023). To effectively combat online hate speech, it is moreover necessary to reach transgressors and bystanders and sustainably change their behavior.

To comprehensively understand whether hate speech victims and bystanders can positively impact transgressors, it is vital to look at their influence not just on the average transgressor but also on the radical ones who are responsible for the vast majority of online hate speech (Evkoski et al., 2022; Goel et al., 2023). However, there is limited empirical evidence of how counterspeaker identity affects both groups. Research investigating its impact has been limited to highly offensive transgressors (Munger, 2017), potentially missing varying effects for the average transgressor on social media. Field studies that encompass a broader sample of transgressors, in turn, considered counterspeech by anonymous individuals (Bilewicz et al., 2021; Hangartner et al., 2021), not capturing how counterspeaker identity may modulate their effectiveness.

Bystanders, in addition to transgressors, play a pivotal role in the dynamics of hate and counterspeech. Their participants in hate speech can greatly amplify its presence and exacerbate its impact (Bilewicz & Soral, 2020). Conversely, when bystanders actively oppose hate speech, they can substantially attenuate its negative effects (Garland et al., 2022). However, the current literature only yields indirect evidence of how counterspeaker identity might modulate its impact on further bystanders. A previous survey study found that hearing of an ingroup authority rejecting hate speech entices people also to reject such posts and endorse counterspeech efforts (Siegel & Badaan, 2020). Nevertheless, direct investigations of such effects in online environments remain scarce.

In this study, we therefore asked: Can members of the victimized group and bystanders use counterspeech to influence other bystanders and transgressors effectively? Looking at transgressors, we are interested in the average transgressor as well as radical ones.

For a counterspeaker from the victimized group, prior research suggests they might successfully influence the average transgressor, while their effect on radical transgressors and bystanders remains uncertain. One pathway through which counterspeech can positively influence its recipients is by fostering empathy and perspective-taking, which have been linked to hate speech rejection (Bilewicz & Soral, 2020; Cowan & Khatchadourian, 2003; Wachs et al., 2022). This mechanism may be especially potent for transgressors wishing to avoid feelings of guilt and self-disappointment (Chaney & Sanchez, 2018). On X, transgressors reduced further hate speech posts after being confronted by anonymous users, highlighting the harm caused by their posts (Hangartner et al., 2021). A counterspeaker belonging to the victimized group could potentially leverage this mechanism more effectively than an anonymous account. Members of marginalized groups are considered more credible than majority group members when communicating that an action is harmful to their group (Crosby & Monin, 2013). In contrast, individuals who speak up merely on the victimized group's behalf without being targeted can be seen as self-interested do-gooders (Kutlaca et al., 2020). Thus, the impact of empathy-based counterspeech by a member of the victimized group could be even bigger than that of other counterspeakers. However, their influence may be attenuated for more radical transgressors who are less open to the perspectives of marginalized groups (Stone, 2011) or concerned about appearing prejudiced (Crosby & Monin, 2013). Similarly, it remains uncertain whether such a counterspeaker equally influences bystanders who may feel less guilty than the transgressor since they did not harm the victimized group themselves.

A counterspeaker who does not belong to the victimized group but rather to the transgressor’s ingroup could positively influence bystanders and more radical transgressors via an effect on ingroup norms (Munger, 2017; Siegel & Badaan, 2020). People are motivated to conform to the perceived norms of their ingroups, especially when they are highly identified with their group (Turner et al., 1987). These norms are commonly inferred by observing the behavior of ingroup members (Tankard & Paluck, 2016). Ingroup counterspeakers can thus reduce subsequent hate speech by communicating that hate speech is antinormative for their group. For example, Munger (2017) found on the social networking service X (formerly known as Twitter) that counterspeech from high-status in-group members substantially reduced subsequent racist slurs by transgressors. However, the study investigated only a subsample of users who regularly used highly offensive language, including racial and misogynistic slurs. It remains unclear whether its findings generalize to the vast majority of less extreme transgressors. It appears plausible that the less frequent anti-outgroup slur use indicates lower identification with one’s ingroup, reducing the impact of its norms (Turner et al., 1987). It is equally uncertain how a counterspeaker who shares the transgressor’s group influences further bystanders. As mentioned above, such counterspeakers risk being perceived as do-gooders (Kutlaca et al., 2020). In contrast, in their general communication on social media, bystanders are inspired particularly by ingroup compared to outgroup members (Seering et al., 2017), yet results are less conclusive for similar effects on bystander counterspeech (Obermaier et al., 2021; Siegel & Badaan, 2020).

Complementing the focus on counterspeaker identity, it is further necessary to consider how social status modulates their impact. High-status individuals often exert greater influence in many online interactions (Aral & Walker, 2012; Seering et al., 2017). Highlighting the importance of status for counterspeech success, research found that bystander attitudes towards hate speech were only shifted by anti-hate speech statements voiced by an authority figure but not by ones without such endorsement (Siegel & Badaan, 2020). The importance of status was further underscored for extreme transgressors, who were only affected by high-status confronters (Munger, 2017; Siegel & Badaan, 2020). Moreover, our re-analysis of data from Munger's (2017) study revealed that low-status counterspeakers could even cause backlash effects, increasing the use of racial slurs among transgressors with a prior propensity to slur use (*see Supplement*). It is thus crucial to consider the modulating effects of counterspeaker status for their effectiveness in achieving behavioral change in transgressors and bystanders alike.

We conducted a pre-registered and high-powered (N = 481) field experiment on the social media platform X to experimentally test the effect of counterspeech. We varied the counterspeaker's group affiliation and status, testing how they impacted transgressor and bystander behavior. Our research addressed the following research questions:

1. Can counterspeech by members of the transgressor's group and the victimized group discourage transgressors from using the racist slur *n*gger* (n-word) for the next week, two weeks, and/or month?
2. Does the influence of the counterspeakers vary between transgressors who frequently used the slur prior to receiving counterspeech and transgressors who used it less often?
3. Do the counterspeakers affect bystanders' reactions differently?
4. Does the social status of a counterspeaker modulate their impact?

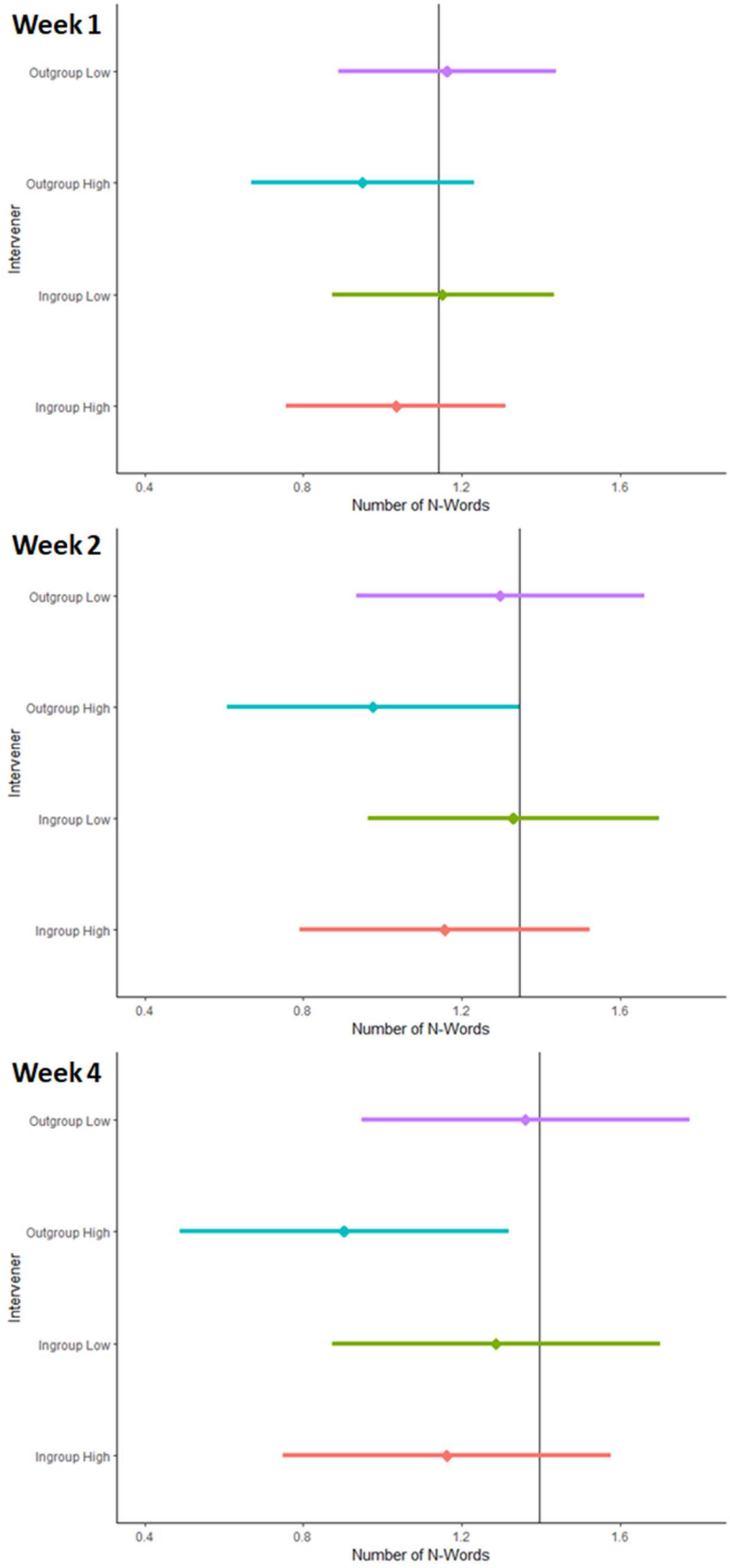
In a field study on X, we posted counterspeech in response to White transgressors who insulted other users with the n-word. The counterspeech comment invited the transgressors to consider the implications of their language for their targets. We varied the counterspeaker's ethnicity: White (transgressor's *ingroup*) or a Person of Color (victimized *outgroup*). In line with prior research (Munger, 2017; Paluck et al., 2016), we manipulated the counterspeaker's status through their number of followers (under 10 for low status and over 400 for high status). Moreover, we added a control condition in which we did not post counterspeech. To measure effects on transgressors, we then assessed how often the transgressors used the n-word in the following week, two-week, and four-week intervals. To analyze the effects of counterspeech on bystanders, we assessed likes for hate speech and counterspeech comments, as well as pro-counterspeech and pro-hate speech responses. In doing so, we extended prior research on identity and status effects of counterspeech (Munger, 2017) by considering the whole population of transgressors and by analyzing bystander reactions to the hate and counterspeech comments.

Results

Overall Effects of Group Affiliation and Status

We computed multiple linear OLS regressions to test whether ingroup or outgroup counterspeakers with high or low social status significantly influenced transgressors in our sample. The dependent variable was the number of n-words used by the transgressor in the given time period and the predictor was the counterspeech condition with a dummy-coded predictor *no counterspeech* as the baseline to which the *ingroup-high*, *ingroup-low*, *outgroup-high*, and *outgroup-low* conditions were compared. We controlled for the number of n-words used by the transgressor in the two months before the counterspeech (*Prior Slurs*) and for the logarithmic follower count of the transgressor. We found that counterspeech in the *outgroup-high* condition predicted a decrease of the n-word use compared to *no counterspeech* from 1.14 to 0.95 n-words after one week (b = -0.19, 95% CI [-0.47, 0.09], p = .184), from 1.35 to 0.98 after two weeks (b = -0.37, 95% CI [-0.74, 0.00], p = .050), and from 1.4 to 0.9 after four weeks (b = -0.49, 95% CI [-0.91, -0.08], p = .020). That is, while n-word use increased in the *no counterspeech* condition over time, it remained constant in the *outgroup-high* condition. No other condition led to a significant reduction of n-word use. Results for all conditions are displayed in *Figure 1*. Values can decrease across time points when tweets containing the n-word are deleted.

Figure 1.


Week 1
Outgroup Low
Outgroup High
Ingroup Low
Ingroup High
Intervener
0.4
0.8
1.2
1.6
Number of N-Words
Week 2
Outgroup Low
Outgroup High
Ingroup Low
Ingroup High
Intervener
0.4
0.8
1.2
1.6
Number of N-Words
Week 4
Outgroup Low
Outgroup High
Ingroup Low
Ingroup High
Intervener
0.4
0.8
1.2
1.6
Number of N-Words

The predicted total number of n-words posted in the time period after controlling for racism score and logarithmic follower count are depicted on the x-axis. Intervener identity is depicted on the y-axis. The vertical line represents the predicted number of n-words in the no counterspeech condition. Decreases between weeks represent n-word deletions. *Week 1* denotes the one-week time interval, and *Week 2* and *Week 4* the other two intervals.

Differential Effects for Racist Transgressors

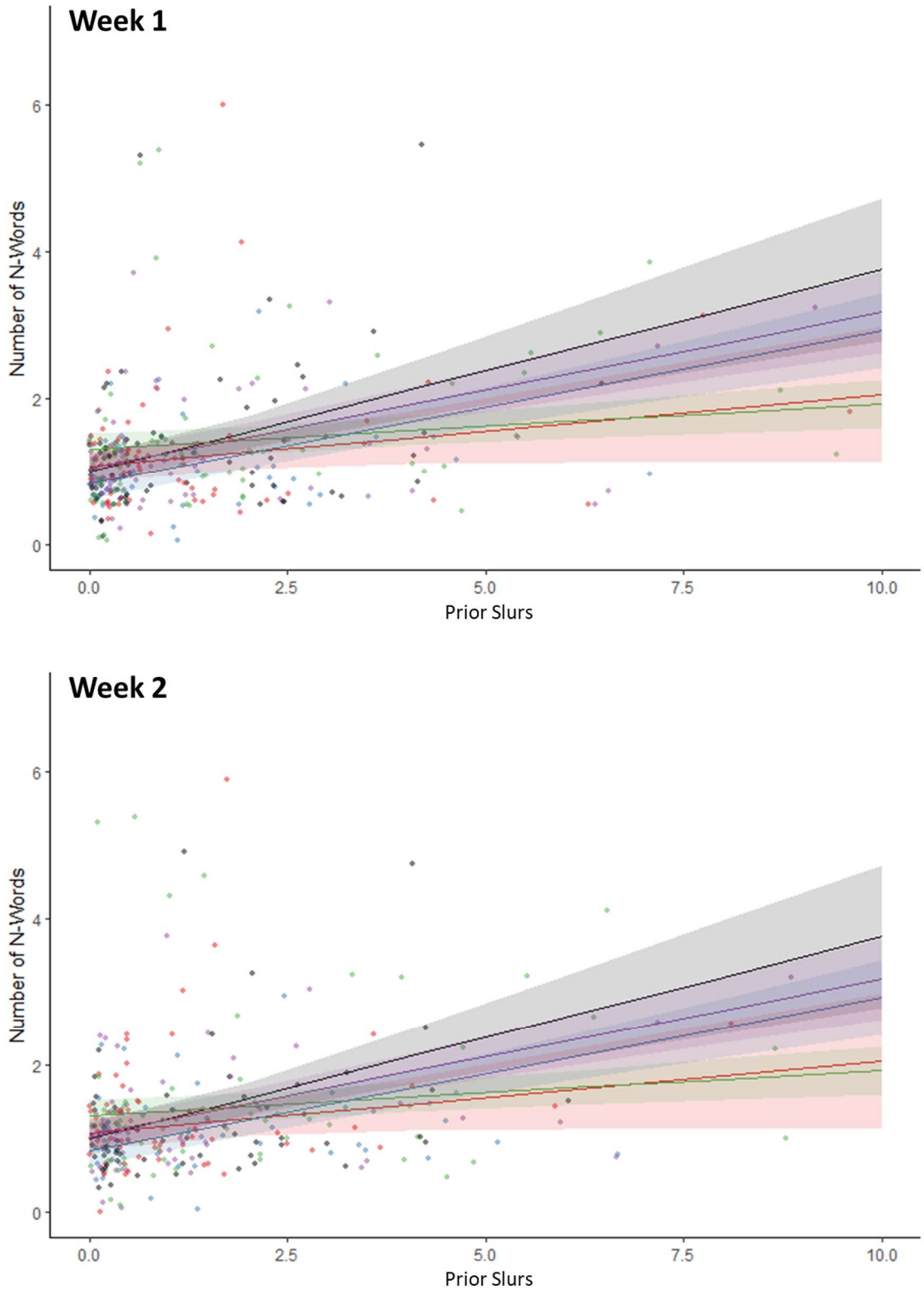
Next, we tested whether the effect of the different counterspeakers varied depending on the transgressor's prior racism. We added an interaction term between the counterspeech condition and *Prior Slurs* to the regressions. We found that ingroup accounts decreased the use of hate speech after one week for people who had displayed high prior racism but not for those who had displayed low prior racism (for *ingroup-high status*: *b* = -0.18, 95% CI [-0.32, -0.03], *p* = .018; for *ingroup-low status*: *b* = -0.21, 95% CI [-0.33, -0.10], *p* = < .001). This was also observed for the *ingroup-low* condition after two weeks (*b* = -0.18, 95% CI [-0.33, -0.03], *p* = .016). Moreover, we observed that the median transgressor used more n-words in the week after being addressed by the *ingroup-low* condition compared to *no counterspeech* (*b* = 0.31, 95% CI [0.02, 0.61], *p* = .039).

We explored the effect of the different conditions on more and less racist transgressors by computing marginal effects (Arel-Bundock, 2023) at the median *Prior Slurs* score (1.00) as well as one standard deviation (4.48) and two standard deviations (7.96) above it.

Turning to the ingroup intervener, we found that the *ingroup-high* status account did not have an effect on the median transgressor after one week (*b* = -0.11, 95% CI [-0.38, 0.16], *p* = .424), but more racist transgressors were affected (*Prior Slurs*$_{+1SD}$: *b* = -0.72, 95% CI [-1.30, -0.14], *p* = .014; *Prior Slurs*$_{+2SD}$: *b* = -1.34, 95% CI [-2.39, -0.28], *p* = .013). We observed the same pattern for the *ingroup-low* condition after one week (*Prior Slurs*$_{Median}$: *b* = 0.10, 95% CI [-0.17, 0.37], *p* = .474, *Prior Slurs*$_{+1SD}$: *b* = -0.65, 95% CI [-1.11, -0.18], *p* = .006; *Prior Slurs*$_{+2SD}$: *b* = -1.39, 95% CI [-2.20, -0.58], *p* = < .001); and after two weeks (*Prior Slurs*$_{Median}$: *b* = 0.09, 95% CI [-0.27, 0.45], *p* = .634, *Prior Slurs*$_{+1SD}$: *b* = -0.54, 95% CI [-1.16, 0.07], *p* = .083; *Prior Slurs*$_{+2SD}$: *b* = -1.17, 95% CI [-2.25, -0.10], *p* = .032).

Turning to the outgroup intervener, the median transgressor and transgressors whose *Prior Slurs* were one standard deviation above the mean were positively affected by the *outgroup-high* condition after two weeks (*Prior Slurs*$_{Median}$: *b* = -0.38, 95% CI [-0.74, -0.02], *p* = .039, *Prior Slurs*$_{+1SD}$: *b* = -0.71, 95% CI [-1.36, -0.07], *p* = .030). This did not apply to the most racist transgressor (*Prior Slurs*$_{+2SD}$: *b* = -1.05, 95% CI [-2.19, 0.09], *p* = .072). This pattern persisted after four weeks (*Prior Slurs*$_{Median}$: *b* = -0.47, 95% CI [-0.88, -0.05], *p* = .027, *Prior Slurs*$_{+1SD}$: *b* = -0.77, 95% CI [-1.52, -0.03], *p* = .042; *Prior Slurs*$_{+2SD}$: *b* = -1.08, 95% CI [-2.40, 0.24], *p* = .109).

Figure 2.



The predicted total number of n-words posted in the time period after controlling for logarithmic follower count are depicted on the y-axis. The racism score is depicted on the x-axis. Extreme values are not displayed for easier interpretability. The *ingroup-high* condition is red, *ingroup-low* green, *outgroup-high* blue, *outgroup-low* purple, and *no counterspeech* is black.

Effects on Bystanders

Finally, we investigated the effects of the different counterspeakers on bystander behavior. We conducted logistic regressions with dummy-coded counterspeech conditions (excluding the *no counterspeech* condition and with *ingroup-high* as the baseline condition) and *Prior Slurs* as well as the transgressors' logarithmic follower count as control variables. The analyses yielded no difference for different counterspeech conditions on likes for the counterspeech comment (*p*s > .24), bystander comments supporting counterspeech (*p*s > .10), or bystander use of the n-word (*p*s > .07). However, logistic regression analysis with the *ingroup-high* condition as the baseline yielded that pro-hate speech comments by bystanders were substantially more frequent at 13.4% of all exchanges in the *outgroup-high* condition (*b* = 2.66, 95% CI [0.98, 5.58], *p* = .012) and the no counterspeech condition with 9.2% (*b* = 2.23, 95% CI [0.50, 5.16], *p* = .038) compared to 1.1% in the ingroup-high condition.

Replication of Prior Research

As reported above, Munger (2017) found an effect for high-status ingroup counterspeakers and not for outgroup or low-status counterspeakers in a field experiment on Twitter. However, that study only considered the most offensive quartile of transgressors, which was pre-selected based on their use of offensive words such as slurs, insults, and sexual language. They were only included if their offensiveness exceeded that of 75% of a random sample of Twitter accounts. We checked whether we could replicate the study's findings. Therefore, we added *Prior Offensiveness* (*see Methods*) and its interaction with the counterspeech conditions as a predictor to the regression analyses. We applied the same coding scheme for missing values used by Munger (2017). We found that the *ingroup-high* condition positively interacted with *Prior Offensiveness* when predicting n-word use after one week (*b* = -0.03, 95% CI [-0.05, -0.01], *p* = .006), indicating that the high-status ingroup counterspeaker's positive influence increased with increasing offensive word use of the transgressor. This resulted in a significant reduction of n-word use by the *ingroup-high* condition for users whose offensiveness was in the top quartile as indicated by the marginal effect (*b* = -0.08, 95% CI [-0.13, -0.03], *p* = .003, *see Supplement for full analyses*). Moreover, we replicated the backfire effect of the *outgroup-low* condition for highly racist and offensive transgressors that our reanalysis of the prior study's data had yielded for every time interval (*see Supplement for analyses*).

Discussion

A growing body of research suggests that user-generated counterspeech can serve as an effective tool against online hate speech that can potentially be used by both hate speech targets and bystanders (Bilewicz et al., 2021; Hangartner et al., 2021; Ozalp et al., 2020; Siegel & Badaan, 2020). Our pre-registered field experiment on the social media platform X/Twitter found that counterspeech against racist slurs can indeed exert a long-lasting positive impact. However, the influence of counterspeech by bystanders and members of the victimized group highly depended on their status and the transgressor's racist disposition. For the median transgressor, counterspeech by a high-status member of the victimized group effectively reduced racial slur use over a month. Highly racist transgressors,

on the other hand, were positively impacted by counterspeakers from their ethnic ingroup. We also found that counterspeech by a high-status member of the transgressor's ingroup substantially reduced subsequent bystander hate speech. Our findings demonstrate that counterspeech can be an effective tool for hate speech targets and bystanders alike. However, it is not a one-size-fits-all solution - the social identity and status of the counterspeaker can have very different effects depending on the audience.

The majority of users in our sample who had seldomly or not at all used the n-word in the two months prior to posting their transgression were only influenced by a high-status member of the targeted outgroup. Counterspeech from that user significantly suppressed further n-word use compared to transgressors receiving no counterspeech and this difference increased across time points. Our results resonate with findings that empathy-evoking messages are particularly effective (Hangartner et al., 2021). The counterspeech message we applied can be classified as empathy-based counterspeech as it invites the transgressor to consider the perspective of the target of the hate speech (Hangartner et al., 2021). Such counterspeech being voiced by a member of the victimized group could, therefore, be especially successful because they are perceived as more credible and knowledgeable when it comes to evaluating transgressions against their group (Crosby & Monin, 2013). However, in our field study, we could not directly assess this assumed mechanism. Future research in controlled environments could further explore whether empathy is indeed why counterspeech by a high-status member of the victimized group is effective.

Our findings contradicted our prior reasoning that a high-status counterspeaker of the transgressor's ingroup, rather than the victimized group, would wield the most influence. We had been informed by a prior study employing a similar design that had been conducted on a highly offensive subsample of Twitter users (Munger, 2017). The transgressor's offensiveness seems to modulate counterspeaker influence since we were able to replicate the findings of this prior study for the most offensive users in our study. Using offensive language, including outgroup-derogating content, could indicate a decreased level of empathetic concern and a higher endorsement of one's ingroup, which would attenuate the effect of a counterspeaker of the victimized group compared to that of an ingroup member (Crosby & Monin, 2013; Stone, 2011). Our findings thus highlight the need for comprehensive samples to obtain generalizable insights into the effects of online counterspeech.

In addition, this study's results reaffirm that high-status members in a social media environment can be highly influential actors, with markers such as social connectedness indicating one's relevance (Paluck et al., 2016; Seering et al., 2017). In our factorial experimental design, we contrasted users with less than ten followers to those with more than four hundred. While our research highlights the importance of status for effective counterspeech, it remains unclear what influence a peer counterspeaker could exert. Prior studies have shown that a user's status can have a linear effect on their influence (Seering et al., 2017), suggesting an attenuated effect for peer counterspeech compared to high-status accounts. Encouragingly, findings for anonymous counterspeakers showed that a smaller number of 55 followers can already be sufficient to ensure counterspeech effects

(Hangartner et al., 2021). Future research could investigate boundary conditions, identifying a minimum number of followers necessary to address transgressors and bystanders effectively.

We found that transgressors who had displayed high levels of prior racism, but not the median transgressor, were influenced by counterspeakers sharing their ethnic ingroup. The results suggest that a different mechanism might be at play for more racist users. They were likely influenced by ingroup norm perceptions (Tankard & Paluck, 2016; Turner et al., 1987). Through their public rejection of slur use, ethnic ingroup members could communicate that such behaviors are not tolerated within their group. Surprisingly, the low-status ingroup counterspeaker had a more sustained effect than the high-status confronter, reducing slur use for two weeks instead of one. The similarity between counterspeakers and transgressors might explain this finding. Individuals are most influential in shaping ingroup norms when they are perceived as prototypical for their ingroup (Klein et al., 2007). Transgressors whose prior slur use was one standard deviation above the median had a median of 26.5 followers, placing them closer to the low-status account than the high-status one. Future research could further explore this interaction between social status and ingroup membership on counterspeech effectiveness.

Moreover, high-status counterspeakers from the transgressor's ethnic ingroup successfully suppressed subsequent bystander hate. Exploratory analyses yielded that counterspeech from a high-status ingroup member led to less pro-hate speech bystander responses than no counterspeech or even counterspeech by a high-status member of the victimized group. The latter led to the highest number of pro-hate speech bystander comments across all conditions. Our findings align with prior research that found that high-status ingroup members are especially effective at addressing bystanders (Seering et al., 2017; Siegel & Badaan, 2020). However, unlike previous research (Miškolci et al., 2018), we did not observe that counterspeech motivated further bystander counterspeech. This might have been caused by the nature of the online context in which we conducted our study. While prior research found effects for heterogeneous bystander audiences of online news readers, our interactions were often only seen by the transgressor's and his target's followers. Hate speakers on social media are disproportionately connected with a homogeneous group of like-minded individuals, which could decrease the likelihood of bystander counterspeech relative to a more varied audience (Goel et al., 2023; Mathew, Dutt, et al., 2019). Counterspeech in such homogenous contexts may be especially important to avoid radicalizing feedback loops (Bilewicz & Soral, 2020; Schieb & Preuss, 2016). It may, however, also risk more hostile backlash for members of the victimized group.

Our research has several limitations. Since we exclusively looked at racist hate speech, it remains unclear whether the same effect can be found for other hate speech domains (Mathew, Saha, et al., 2019). It seems plausible that counterspeakers can also leverage empathy and social norms in other contexts. For instance, ingroup norm-based counterspeech positively influenced extreme transgressors who had posted sectarian hate (Siegel & Badaan, 2020). While that study applied counterspeech from an anonymous user who merely mentioned ingroup norms, it appears likely that an ingroup speaker could have a similar effect. Moreover, empathy has been identified as a driver for bystander

interventions in other related domains, such as cyberbullying (Van Cleemput et al., 2014; van der Ploeg et al., 2017). In contrast, prior research found that delayed counterspeech by female users did not affect misogynist transgressors (Whiley et al., 2023), contradicting our findings that outgroup members can wield positive influence. However, the counterspeaker in that context more closely resembled our low-status counterspeaker. This leaves the question unanswered as to whether the observed differences were caused by methodological differences or dynamics that set apart misogynistic hate speech. Future research could explore the generalizability of our findings across different hate speech contexts.

Our findings can form the basis for more systematic explorations into how counterspeech effects generalize to other online contexts. Other social media platforms characterized by more tightly-knit or pre-selected communities, such as Reddit, could enhance the effect of group identity and status. By contrast, its effect could be reduced in more heterogeneous settings, such as comment sections at news outlets. As mentioned earlier, counterspeech could have inspired further bystanders to speak up in addition to suppressing further hate speech in a more heterogeneous audience. Future research could explore whether our findings generalize to other hate speech dimensions and contexts.

Another limitation pertains to the context in which we found these results. We were fortunate to use X's free Academic Research API during our data collection. However, we had to stop data collection before reaching our target sample size due to indications of tweet filtering by X in the last weeks and the ultimate discontinuation of the Academic Research API. Unfortunately, the current trend is to limit academic access across social media platforms, making counterspeech studies like ours increasingly difficult.

Our findings can inform people who wish to stand up against racist hate speech online. Firstly, our results yielded that social status is crucial for successful counterspeech. Counterspeech by a member of the victimized group with close to no followers had no effect. High-status counterspeakers wielded more influence. Counterspeech by a high-status member of the victimized group reduced subsequent racist transgressions by the median transgressors. However, we also found tentative evidence that they also faced more hostile bystander reactions than high-status counterspeakers who shared the transgressor's ethnic ingroup. Thus, high-status members of the victimized group who employ empathy-eliciting counterspeech are most effective at reducing overall racist slur use, but they also pay the highest price for their moral courage. Moreover, we found that bystanders who share the same ethnicity as the hate speaker are uniquely positioned to influence the most radical hate speakers who are responsible for the vast majority of online hate speech (Evkoski et al., 2022; Goel et al., 2023). Tentative results, moreover, suggest that the high-status ingroup member can discourage further bystanders from endorsing hate speech. Thus, high-status members of the same ethnic group as the transgressor are most suitable to address bystanders in our setting, in addition to the most radical hate speakers.

In summary, our field study on Twitter/X revealed that counterspeech can be a successful countermeasure against racist hate speech but that its effectiveness highly depends on the counterspeaker's social identity and status and the transgressor's predispositions. Effects

for the median transgressor, highly racist transgressors, and for bystanders varied by the ethnicity of counterspeakers as well as their number of followers. Our findings highlight the practical importance of targeted counterspeech and the methodological importance of comprehensive samples to evaluate which counterspeech and counterspeakers can make an impact in the real world.

Materials and Methods

The ethics committee of the Bergische University Wuppertal approved this study. Data, analyses, replication materials, and the preregistration can be found at https://osf.io/3udqb/?view_only=ecdec805dfb14d7e9cbe8cab2e80faa9.

Data was collected between July 27th, 2022, and May 16th, 2023. Our sample consisted of English-speaking X users with publicly accessible accounts who had used the n-word in a disparaging manner in a tweet and directly addressed another user. To detect such users, we captured all tweets containing the n-word posted in the last 24 hours using the now-defunct Academic Research API via the academictwitteR package (Barrie & Ho, 2021). Considering only the last 24 hours, we ensured that our responses were posted in close temporal proximity to the transgression. We then manually checked whether tweets had been composed in English, contained the n-word in a disparaging manner, and addressed another user. Moreover, tweets were excluded if their authors were identifiably less than 18 years old, female, not white, if they appeared to be friends with the target, or if their self-description on X indicated an innocuous use of the n-word. If their tweet was included in the study, users were randomly assigned to one of the four counterspeech conditions 2 (*ethnicity*: White/Person of Color) X 2 (*popularity*: high/low) or a non-counterspeech baseline condition.

Our final sample consisted of 481 users ($N_{control}$ = 101, $N_{ingroup\text{-}high\ status}$ = 95, $N_{outgroup\text{-}high\ status}$ = 91, $N_{ingroup\text{-}low\ status}$ = 95, $N_{outgroup\text{-}low\ status}$: 99). The final sample was substantially smaller than the initial size of 800 that we had pre-registered. This was due to X's ownership change during the data collection period and resulting changes, as well as the eventual discontinuation of the Academic Research API. Therefore, we had to stop data collection in May 2023.

Of the initial 481 transgressors included in our study, nine accounts were inaccessible after one week. That number increased to 115 after two weeks and 124 after four weeks. Overall, rates ranged from 21% for *outgroup-low* to 32% for *outgroup-low*. The number of inaccessible accounts did not differ significantly between the counterspeech and control conditions (*See Supplement*). We could not assess whether users had voluntarily deleted their accounts, restricted access to them, or whether X had deleted them. As we had pre-registered, we did not exclude accounts whose tweets became inaccessible to us from analyses. Instead, we coded missing values as the confirmed number of n-words that could potentially have been visible to other X users during the time period. That is, if we could not assess a transgressor's tweets after one week, we coded their n-word use as zero. If we could not assess their tweets in the subsequent time periods, we coded their n-word use for that period as the same as for the prior time period (i.e., if a transgressor had posted two n-words in the two-week period, his score for the four-week period was also coded as two).

In the counterspeech conditions, transgressors received a reply from an account that had a light-skinned profile picture and was named “Greg” (White ethnicity cues) or a dark-skinned profile picture with the name “Rasheed” (Person of Color cues). Moreover, the account either had more than 400 followers (high status) or ten or fewer followers (low status). Each of our accounts had existed for more than a year at the start of data collection and followed multiple non-political accounts. Throughout the year, we had actively posted and shared non-political content from each account to give it more resemblance to a real user. In each condition, we posted the same counterspeech: *@[subject] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language.* After the counterspeech, we measured the subsequent behavior of the addressed user and other users who commented on the same conversation over a month. We collected the number of n-words posted by the addressed transgressors after a one-week, a two-week, and a four-week period starting with the n-word to which we had replied.

At the time of the counterspeech, we collected the number of followers of the transgressor account and the number of tweets they had posted prior to the counterspeech. We also assessed prior offensiveness and slur use of the transgressor. We assessed prior offensiveness to replicate findings by Munger (2017). For that, we collected all transgressor tweets of the previous two months and counted how many words from an offensiveness dictionary created by Munger (2017) that contained racist and sexist slurs, swear words, and sexually explicit language had been used (*see Supplement*). As done by Munger (2017), we computed a reference offensiveness score from the 400 most recent tweets by 450 randomly sampled X accounts that were older than six months as a reference. We also computed a more specific offensiveness score that contained the racist words in the dictionary used by Munger (2017) and additional racist slurs derived from the English Wiktionary (2022). Finally, we counted the overall number of n-words used by the transgressor. If the addressed account had not existed two months prior, we collected how many times, on average, a person used the n-word from the moment the account was created and extrapolated how often that would have been over two months.

After one week, we collected the bystander replies to the initial transgression tweet and coded whether the replies were (1) counterspeech or supporting the counterspeech comment, (2) hate speech or supporting the hate speech comment. We assessed the sum of favorites, replies, and retweets that the hate and the counterspeech comments received. We further measured if the transgressor had replied to a counterspeech comment in support or rejecting the counterspeech comment.

The regression analyses for overall and differential effects were pre-registered as confirmatory analyses, the regression analyses regarding bystander reactions as exploratory analyses, and the exploration of marginal effects as well as the replication of prior findings regarding highly offensive transgressors were decided on post-hoc (*see Pre-Registration*). For our statistical analyses, we used R version 4.3.2 (2023-10-31 ucrt). Marginal effects were computed using the `marginaleffects` package (Arel-Bundock, 2023). Metric predictors were median-centered. As pre-registered, we excluded outliers using the `car::outlierTest()` function (Fox & Weisberg, 2019). The function computes the Bonferroni-adjusted probability to observe a data point if residuals are normally distributed.

The observation is excluded from the analysis if the probability is below 5%. As robustness checks, we computed the regression analyses with robust covariance matrices, without missing cases, and with imputed instead of recoded data for missing values (*see Supplement*).

Overall Discussion

Hate speech is a severe problem in online spaces, harming communities and societies across the globe (United Nations, 2020). It not only inflicts immense harm on its direct victims (Keighley, 2022; Tynes et al., 2008) but also extends its adverse effects to bystanders (Hsueh et al., 2015; Müller & Schwarz, 2021) and undermines entire discourses (Bilewicz & Soral, 2020; Garland et al., 2022). Its damaging influence operates through two mechanisms: the erosion of established anti-hate speech norms (Alvarez-Benjumea, 2022) and the desensitization of its audience to hateful language (Soral et al., 2018).

Counterspeech is a promising countermeasure against hate speech that can be applied broadly, maintain free speech, and possibly improve online discourse by fostering sustainable behavioral changes in bystanders and transgressors. However, the current empirical evidence cannot sufficiently answer whether regular counterspeech inspires bystanders to speak up and dissuades users from further offenses (Cepollaro et al., 2023; Rudnicki et al., 2023; Windisch et al., 2022). For bystanders, research assessing counterspeech effects on real-world behavior showed promising results (Miškolci et al., 2018), but controlled experiments did not observe a causal link between counterspeech and increased bystander intervention (Alvarez-Benjumea & Winter, 2018; Leonhard et al., 2018). Moreover, evidence for the average transgressor is limited to anonymous counterspeakers (Bilewicz et al., 2021; Hangartner et al., 2021), but studies with more extreme transgressor samples found that counterspeaker identity can substantially alter counterspeech effects (Munger, 2017).

It further remains unclear through which mechanisms counterspeech may exert its influence. While prior studies suggested social norms as a mediator (Munger, 2017; Siegel & Badaan, 2020), their effect was not directly captured and evidence stemmed from research on extreme transgressor subsamples. The perceived severity of hate speech, which can motivate bystander counterspeech and suppress subsequent transgressions, could also serve as a mediator (Chaney & Sanchez, 2018; Leonhard et al., 2018). However, it remains uncertain whether counterspeech can influence perceptions of bystanders and transgressors. If counterspeech can successfully leverage social norms and severity perceptions, it can potentially act as a potent antidote against hate speech by directly countering two of the latter's critical effects - the breakdown of conversational norms and the desensitization of its audience (Bilewicz & Soral, 2020).

In this dissertation, I examined counterspeech effects through a multi-method investigation encompassing a longitudinal study in a mock social media forum, cross-sectional vignette-based studies, and a field study on social media. This approach allowed me to isolate counterspeech effects on subsequent behavior and possible mediators in controlled experiments and further confirm their relevance in a real-world social media setting. My empirical investigation focused on bystander counterspeech in Chapters B and C and extended its scope to bystander and transgressor hate speech in Chapter D.

I found that counterspeech not only emboldens bystanders to speak out against hate speech but that it can also reduce subsequent hate speech among both bystanders and transgressors. Across studies, counterspeech shaped collective and group-specific social

norms and impacted how the severity of hate speech was perceived. Subsequently, I will provide a detailed discussion of these findings, highlighting the congruencies and incongruencies across my studies and shedding light on the potential of counterspeech as a tool for fostering inclusive online interactions.

Chapter	Design	Counterspeech Impact	Collective Norms	Ingroup Norms	Severity Assessments
B	Controlled experiment: Longitudinal interaction in mock social media forum	Increase in bystander counterspeech	Positive mediation for bystander counterspeech		Positive mediation for bystander counterspeech
C	Three controlled experiments: Interactions with social media vignettes	Increase in bystander counterspeech (Exp. 3)		Positively predicts bystander counterspeech endorsement Positive mediation only in Exp. 3	
D	Field study: User interactions on Twitter	Partial decrease in bystander and transgressor hate speech	No indication of collective norms' effect	Ingroup counterspeech suppressed hate speech by bystanders and racist transgressors	Outgroup counterspeech suppressed hate speech by median transgressors

RQ.1: Counterspeech Effects

The first overarching question of my dissertation was

Does counterspeech have a positive impact on bystanders and transgressors?

Across all chapters, I investigated whether counterspeech inspired further bystanders to speak up. In Chapter D, I also investigated whether it suppressed hate speech by bystanders and transgressors on social media.

Effects on Bystanders

For bystanders, I found in Chapter B that counterspeech inspired subsequent counterspeech against online hate speech for almost a week. Similarly, I observed in Chapter C that counterspeech from in- and outgroup members alike inspired bystanders to speak up

against the same hate speech and a subsequent hate comment. In the social media setting of Chapter D, counterspeech significantly reduced further bystander hate speech when voiced by a high-status ingroup counterspeaker.

In a variety of settings and across different groups, my research consistently showed that counterspeech has a positive effect on further bystanders. This aligns with previous studies conducted in real-world contexts, which identified a positive association between counterspeech and increased bystander interventions (Cary et al., 2020; Miškolci et al., 2018). Importantly, my findings in controlled settings provide evidence that this association is caused by the direct influence of counterspeech on further bystander behavior. These results are particularly encouraging, given that the impact of counterspeech is amplified as more bystanders choose to speak up (Schieb & Preuss, 2016, 2018). Thus, my findings indicate that counterspeech can foster inclusive online discourse and serve as an effective tool against online hate speech by mobilizing further bystander intervention.

My research also sheds light on why previous controlled studies found no effects of counterspeech (Alvarez-Benjumea & Winter, 2018; Leonhard et al., 2018). First, I found that counterspeech influenced actual bystander behavior but not intentions to intervene. These findings align with research showing that peoples' predictions whether they will speak out are often inaccurate (Crosby & Wilson, 2015; Kawakami et al., 2009) and that behavioral intentions and actual interventions can be motivated by different factors (Baumert et al., 2013; Goodwin et al., 2020). Second, I observed relatively small effect sizes across studies, which I could only detect due to high-powered research designs. Previous research might not have had sufficient statistical power to observe such effects. My findings thus underscore the necessity of selecting appropriate outcome variables and employing high-powered designs to capture counterspeech effects accurately.

I had further hypothesized that counterspeech effects could have been absent in prior research due to its cross-sectional designs. Such one-shot interactions could have limited the effect of reputational concerns that can drive bystander intervention (Van Bommel et al., 2012; Ziegele et al., 2020). Contrary to this prediction, I observed that counterspeech inspired bystanders both in the longitudinal setting of Chapter B and in the cross-sectional setting of Chapter C. Encouragingly, these results suggest that counterspeech can impact bystanders even when they are anonymous and do not anticipate future interactions with their audience.

Boundary Conditions for Bystanders

Across all studies, I found a consistent positive effect of counterspeech on bystanders. However, the nature of its impact varied between studies. In controlled environments, counterspeech inspired bystanders to also speak up, whereas, in the field study, they were instead discouraged from posting further hate speech. This discrepancy may be attributed to differences in the bystanders' relationships with the victims and transgressors. Participants had no connection to the transgressor or victim in most controlled studies and only a very superficial one in Chapter C Exp. 1. In contrast, counterspeech in the field study primarily reached the transgressor's and the victim's followers. Previous research suggests that a shared group affiliation with a victim increases bystander motivation to intervene

(DeSmet et al., 2016; Levine et al., 2005; Liebst et al., 2019). Counterspeech effects could thus be attenuated by the already heightened motivation of the victim's followers to defend it. Conversely, sharing a group identity with the transgressor decreases the motivation to openly oppose them (Kutlaca et al., 2020; Packer, 2014). This might lower the likelihood that the transgressor's followers overtly oppose the hate comment. Counterspeech could instead influence less public behavior, prompting them to refrain from affirming the hate comment.

In Chapter D, I identified possible boundary conditions for counterspeech effectiveness, observing that only a white counterspeaker with many followers impacted bystander behavior but not a member of the victimized group or low-status accounts. Prior research found that members of the disadvantaged group who speak out against discrimination can be seen as self-interested, which can decrease their influence (Drury & Kaiser, 2014). In addition, the behavior of low-status users is copied less frequently by other users than that of peers or high-status users (Seering et al., 2017). Nevertheless, it is important to consider that my exploratory results regarding bystander behavior in Chapter D were driven by few observations and should, therefore, be interpreted with caution. Additionally, the diminished impact of low-status counterspeakers might be caused by platform mechanics rather than the persuasive impact of bystander status. During the time of data collection, Twitter regularly reduced the visibility of posts from accounts with few followers, either by downranking or completely hiding their comments. This practice could have substantially reduced counterspeech effectiveness, offering an alternative explanation for the decreased impact of low-status accounts.

Having employed peer counterspeakers who are not overtly members of the victimized group, m*y* studies in controlled settings cannot provide further information on group and status effects*.* In the controlled experiments, counterspeech was delivered by individuals who did not explicitly identify with the victimized groups—neither as being overweight, Black, Jewish, or homeless in Chapter B nor as gay, Czech, or from the Faroe Islands in Chapter C. Moreover, since the counterspeakers were framed as fellow study participants, they were peers rather than low- or high-status users.

Given these constraints, the empirical evidence provided by my dissertation cannot definitively determine whether counterspeaker status and group affiliation moderate counterspeech effects on further bystander behavior. It seems plausible that at least a peer status is necessary (Hangartner et al., 2021). However, more research is needed to conclusively investigate status and social identity as possible boundary conditions.

Effects on Transgressors

For transgressors, I found in Chapter D that counterspeech could reduce further hate speech for a month. However, not every counterspeaker was equally successful. Only a high-status counterspeaker from the victimized group reduced racist slur use for the median transgressor. In addition, white counterspeakers were effective for the more racist transgressors in my study. Immediate transgressor reactions in the same conversation remained unaffected by counterspeech.

My results provide additional evidence that counterspeech can be used to influence transgressor behavior over time (Bilewicz et al., 2021; Hangartner et al., 2021; Munger, 2017; Siegel & Badaan, 2020), though it may not affect their immediate responses (Miškolci et al., 2018; Munger, 2017). The discrepancy between short and long-term effects could be caused by the increased impact that counterspeech can exert when its targets have time to ruminate and reflect on their actions (Chaney & Sanchez, 2018). My findings challenge the widespread assumption about the futility of engaging in online hate speech: While counterspeech might not affect the transgressors' immediate reactions, it can cause even more desirable sustainable behavioral change.

Boundary Conditions for Transgressors

In line with extant research, I found that the counterspeaker's social identity and status modulate counterspeech effectiveness (Munger, 2017; Siegel & Badaan, 2020). Moreover, I identified differences between the average and specifically racist transgressors. Both high-status outgroup and ingroup counterspeakers can impact different types of transgressors, suggesting different pathways of counterspeech influence. I will discuss these mechanisms below. These results indicate that prior research on extreme transgressor samples may not accurately capture the dynamics of typical bystander counterspeech. My findings, therefore, underscore the importance of considering both the counterspeaker's characteristics and the transgressor's dispositions to accurately assess counterspeech effectiveness.

Summary

Answering the question of *whether* counterspeech has a positive impact on bystanders and transgressors, my dissertation provides consistent evidence that counterspeech positively impacts bystanders across both controlled experiments and real-world social media interactions. Its effectiveness may be attenuated for counterspeakers who are either low-status or are part of the victimized group. However, there is only tentative evidence for these identity effects. Counterspeech can also reduce hate speech among transgressors. Depending on the recipient, both ingroup counterspeakers, regardless of their status, and high-status outgroup counterspeakers can have a positive impact. Notably, my research indicates that counterspeech can have a long-lasting positive impact, causing sustained behavioral changes among both bystanders and transgressors for up to a month after the intervention.

RQ2: Mechanisms

The second central question of my dissertation was

RQ2: Which mechanisms mediate counterspeech effects on transgressors and bystanders?

I investigated counterspeech impact on perceptions of social norms, divided into collective and ingroup norms, and perceptions of hate speech severity. I used controlled environments to accurately measure the causal impact of counterspeech on these proposed mediators and their respective effects on subsequent bystander counterspeech. Moreover, I tested whether their effect also emerged in real-life interactions on social media.

Mediator: Collective Norms

In Chapter B, I found that the positive effect of counterspeech on subsequent bystander counterspeech was mediated by perceptions of pro-counterspeech forum norms. People who saw a handful of other forum users speak out against hate speech perceived counterspeech as more normative for the forum, even more so when they repeatedly saw others speak up. Their perception of pro-counterspeech forum norms, in turn, predicted their own engagement in counterspeech. However, in Chapter D, a single counterspeech comment did not have the same overall effect on bystanders or transgressors.

My results show that counterspeech can influence bystander behavior through its effect on collective forum norms. Collective norms have been shown to exert a powerful influence on peoples' behaviors in online environments, guiding them to align their actions with what they perceive as normative within a given conversation space (Matias, 2019; Rajadesingan et al., 2020). Potentially leveraging this effect, counterspeech by organized groups positively affected broader online discourse (Garland et al., 2020). However, prior research did not directly assess counterspeech effects on collective conversation norms. My research established a causal link between counterspeech and subsequent perceptions of collective pro-counterspeech norms, which, in turn, embolden bystanders to actively counter online hate speech. These findings suggest counterspeech as a direct antidote for the deleterious effect of hate speech on online conversation norms (Alvarez-Benjumea, 2022; Bilewicz & Soral, 2020).

Boundary Conditions for Collective Norms

In Chapter B's controlled environment, counterspeech influenced bystander behavior via collective norms, a pattern not replicated in Chapter D's real-world setting. If the counterspeech comment had affected transgressors or bystanders through collective norm perceptions, any counterspeaker should have caused behavioral change – or at least both high-status counterspeakers. However, only a high-status ingroup counterspeaker proved successful, indicating an influence of ingroup norms instead of collective norms, as I will discuss below. Unlike Chapter D, Chapter B featured multiple instances of counterspeech. Simulation studies showed that even a handful of counterspeech comments can already be substantially more influential than a single one (Schieb & Preuss, 2016, 2018). This suggests that a minimum number of counterspeakers might be required to affect collective norm perceptions. Supporting this hypothesis, organized counterspeech was more successful at shifting discourse on Twitter than individual counterspeech (Garland et al., 2022). Alternatively, the difference between the controlled and social media settings might stem from participants' pre-existing norm perceptions. Twitter users who are familiar with the platform's discourse may be more resistant to norm shifts than those in a novel mock social media setting. In summary, collective norm perceptions could either have remained unaffected on social media because not enough users engaged in counterspeech or because the bystanders' norm perceptions were too entrenched. My results cannot definitively answer this question, calling for further research varying the number of counterspeakers in a social media context to clarify the dynamics at play.

Mediator: Ingroup Norms

I moreover looked at the role of ingroup norms as a more specific social norm mediator of counterspeech effects on bystander and transgressor behavior. I directly measured the mechanism for bystander counterspeech in Chapter C and tested whether I could find support for it in real life in Chapter D. In Chapter C, I found that ingroup norms predicted pro-counterspeech attitudes and behaviors for bystanders across three different experiments. However, an ingroup counterspeaker positively affected ingroup norm perceptions in only one of the three studies and this did not translate into more counterspeech compared to counterspeech by an outgroup counterspeaker. Supporting the role of ingroup norms on social media, I found that high-status ingroup counterspeakers, but not outgroup counterspeakers, suppressed bystander transgressions in Chapter D. Moreover, ingroup counterspeakers, regardless of their social status, reduced subsequent slur use by racist transgressors.

These results highlight the potential role of ingroup norms driving counterspeech effectiveness for bystanders and transgressors. While previous studies had hinted at their utility for influencing bystanders and transgressors (Bilewicz et al., 2017; Munger, 2017; Siegel & Badaan, 2020), ingroup norms had not been assessed directly prior to this dissertation. My research in controlled settings confirmed that counterspeech can have a causal effect on pro-counterspeech ingroup norm perceptions, which, in turn, motivated bystander endorsement of counterspeech. Moreover, I found evidence for their impact in real-life social media interactions. My findings expand upon earlier research focused on extreme transgressors (Munger, 2017; Siegel & Badaan, 2020) by demonstrating that ingroup norms do not affect the median transgressor but significantly affect the more racist ones.

Boundary Conditions for Ingroup Norms

Counterspeech effectively leveraged ingroup norms only for a subsample of ingroups in Chapter C and a subsample of transgressors in Chapter D. Both findings emphasize the importance of distinct and meaningful group categories. Ingroup norms exert influence when individuals consider a group personally relevant and see themselves as part of it (Tankard & Paluck, 2016). As discussed in Chapter C, counterspeech affected ingroup norm perceptions only when students and retired people were contrasted as relevant identities. When specific university affiliation, a less meaningful ingroup in the German context, was contrasted with either a rival university or unaffiliated counterspeakers, participants perceived little to no distinction between counterspeakers. Further illustrating this dynamic, in Chapter D, ingroup counterspeech only discouraged racist transgressors but not the median transgressor. Moreover, it discouraged the, likely more racist, bystanders who were willing to post further hate but did not encourage the, likely less racist, bystanders prone to counterspeech. Racism is associated with seeing ethnicity as a meaningful and important group category to whose norms one wants to conform (Merriam-Webster, 2024). For less racist individuals, white ethnic identity may play a subordinate role as a category, and they are less motivated to act prototypically for their ethnicity. Taken together, these results suggest that ingroup norms are influenced by counterspeech from an ingroup member only if the group category is meaningful to the receiver.

Identifying a meaningful group to be leveraged by counterspeech can be complex. Individuals belong to a plethora of different social groups and it is unclear which ones are salient and important at a given point (Turner & Reynolds, 2011). This is further complicated by peoples' propensity to form new, personally meaningful groups even on the slightest premises (Tajfel, 1974; Tepper, 1997). A usually meaningful identity dimension in contexts of hate and counterspeech could be the identity characteristic targeted by the hate comment. Prior research indicated that this dimension can be leveraged for extreme transgressors (Munger, 2017; Siegel & Badaan, 2020). Users who posted racist hate speech were swayed by ethnic ingroup norms (Munger, 2017), and users who posted sectarian hate speech were influenced by religious ingroup norms (Siegel & Badaan, 2020). The same mechanism might also apply to bystanders (Siegel & Badaan, 2020). My findings suggest the same for radical transgressors and bystanders. Furthermore, my results indicate that meaningful identity dimensions unrelated to the hate speech content can also be leveraged for less polarized bystanders. However, more research beyond this dissertation is needed to comprehensively determine which group identities and identity dimensions can be most effectively leveraged to combat online hate speech.

Mediator: Hate Speech Severity

Moreover, I investigated if perceptions of hate speech severity mediate counterspeech impact for transgressors and bystanders. In Chapter B, I found that participants who saw other users speak out against hate speech perceived it as a more severe transgression. This perception positively predicted their likelihood to post counterspeech themselves. Results from Chapter D further suggest that severity perceptions can influence transgressors in real life. In that study, different accounts posted empathy-based counterspeech in response to hate speech, stressing its harmful impact on its victims. That counterspeech was effective for the median transgressor when a high-status counterspeaker from the victimized group voiced it. As I discussed in Chapter D, people consider members of the victimized group more credible than others in assessing whether a transgression is harmful (Crosby & Monin, 2013). Thus, empathy-based counterspeech by a member of the targeted group could exert its effect by shaping hate speech severity perceptions. Notably, I did not observe that the more racist and offensive transgressors in my sample were affected by the outgroup counterspeaker. Instead, they were only affected by ingroup counterspeech.

Prior research identified severity assessments as a critical motivator for bystander interventions against cyberbullying (Bastiaensens et al., 2014; Koehler & Weber, 2018) and online hate speech (Leonhard et al., 2018). However, these studies focused on the objective severity of hate speech (Bastiaensens et al., 2014; Koehler & Weber, 2018; Leonhard et al., 2018). For instance, researchers would vary whether participants saw hate speech containing offensive insults against the target group or more extreme hate speech containing dehumanization and calls for violence (Leonhard et al., 2018). However, severity perceptions can be dynamic and substantially fluctuate depending on contextual factors, such as the victimized group (Cowan & Hodge, 1996; Obermaier et al., 2023) or the way hate speech is presented (Schmid, 2023; Schmid et al., 2022). My findings demonstrate that counterspeech can leverage this malleability of severity evaluations, reshaping audience perceptions and inspiring further counterspeech. This way, counterspeech can serve as an

antidote to the desensitizing influence of online hate speech (Bilewicz & Soral, 2020; Soral et al., 2018).

For transgressors, prior research on social media suggests that severity perceptions can successfully influence the average transgressor through empathy-based counterspeech by anonymous accounts (Bilewicz et al., 2021; Hangartner et al., 2021). My findings confirm this pattern for counterspeech by non-anonymous users. The median transgressor is likely more susceptible to feeling empathy for the outgroup than highly racist and offensive transgressors (Avenanti et al., 2010; Forgiarini et al., 2011) and, therefore, more open to empathy-based counterspeech. However, since my research in Chapter D did not directly measure severity perceptions, this interpretation needs empirical confirmation.

Boundary Conditions for Hate Speech Severity

Bystander behavior in Chapter D remained unaffected by outgroup counterspeakers, suggesting that they were less influenced by severity perceptions than transgressors or bystanders in the controlled setting of Chapter B. The discrepancy between bystanders could once more have been caused by their association with the victim and the transgressor. As mentioned earlier, most bystanders in the field study likely either followed the victim or the transgressor, while those in the controlled study had no association with either party. Given that people generally show increased empathy towards ingroup members (Gutsell & Inzlicht, 2012; Tarrant et al., 2009), counterspeech impact could be attenuated for the victim's followers, who may already perceive hate speech against their associate as severe. As I discussed above, the transgressor's followers are likely less open to empathy and, therefore, less likely to reduce their hate speech due to changed severity perceptions (Avenanti et al., 2010; Forgiarini et al., 2011). Exploring the influence of outgroup counterspeakers on a more heterogeneous audience could provide deeper insights into its impact on severity evaluations.

I observed that only the outgroup counterspeaker with a high status, defined by follower count, but not the one with a low status, affected transgressors in Chapter D. Followers are often viewed as a marker of source credibility (Son et al., 2020). Potentially, the low-status account lacked the perceived credibility to effectively communicate the harm associated with racial slurs, attenuating the influence that members of the victimized group usually wield in shaping these perceptions (Crosby & Monin, 2013). Alternatively, platform affordances could have driven the effect, with the counterspeaker's low status translating into less counterspeech visibility. Nonetheless, without conclusive evidence to support these speculations, further research is needed to clarify whether the influence of counterspeaker status is driven by psychological or circumstantial factors.

Summary

Taken together, the results illuminate mechanisms that mediate counterspeech effects on transgressors and bystanders. My studies in controlled settings yielded that counterspeech positively influenced perceptions of collective forum norms, specific ingroup norms, and hate speech severity assessments. I, moreover, observed indications for the impact of ingroup norms and severity assessments in real-life social media interactions.

I found tentative evidence suggesting that counterspeech can positively influence perceptions of collective forum norms, motivating subsequent bystander counterspeech in Chapter B. In a controlled environment, just a few bystander comments were enough to significantly shift perceptions toward a pro-counterspeech norm. However, I did not find this effect for solitary counterspeech comments in the social media setting of Chapter D. This incongruency could either stem from the number of counterspeakers or the entrenched nature of norm perceptions on the real-life social media platform.

My studies revealed more robust evidence for the impact of ingroup norms both in controlled settings in Chapter C and real-life interactions in Chapter D. In Chapter C, I demonstrated that pro-counterspeech ingroup norm perceptions positively mediate the impact of ingroup counterspeech on counterspeech endorsement by bystanders. My investigation on Twitter further showed that both bystanders and racist transgressors are positively affected by counterspeech from a white counterspeaker. However, both chapters underscored the relevance of the group identity as a boundary condition.

Finally, my results consistently highlighted hate speech severity assessments as an additional mediator. In the controlled context of Chapter B, I established that perceptions of hate speech severity mediated the positive impact of counterspeech, leading to increased bystander intervention. While my investigation in Chapter D did not replicate this effect for bystanders, it indicated that transgressors were influenced by hate speech severity assessments. These findings suggest that audience characteristics can substantially modulate the effect of severity perceptions.

Going Beyond This Dissertation

Additional Moderators

Effects of hate speech and counterspeech can be highly context-dependent. For instance, young audiences consider hate speech as less harmful than old ones (Schmid et al., 2022); men reject it less than women (Cowan & Khatchadourian, 2003); homophobic hate speech is considered less civil than misogynistic speech (Obermaier et al., 2023); and some conversation spaces tolerate toxic speech while others ban it (Rajadesingan et al., 2020). Therefore, I approached counterspeech impact from multiple angles to avoid contextual artifacts. My multi-method investigation of counterspeech effects revealed its robust impact across multiple dimensions. Counterspeech positively affected German students and retirees as well as English-speaking Twitter users. Ingroup norms were effective when they were based on life stage as well as ethnicity. Bystanders were positively influenced in ephemeral one-shot exchanges as well as long-term, multi-week forum interactions. Counterspeech had an impact against hate speech that targeted groups based on nationality, ethnicity, class, religion, and body type. However, my studies also revealed two factors that consistently yielded differences between controlled environments and real-world settings.

Status Effects

My controlled experiments found that peer counterspeech inspired bystander counterspeech and positively affected perceptions of collective and ingroup norms as well as hate speech severity. Conversely, my investigation on social media suggested that status

may modulate a counterspeaker's impact on bystanders and the median transgressor, resulting in ineffectual counterspeech from low-status users. It remains uncertain whether status exerts an actual persuasive effect on counterspeech on social media platforms or whether it merely affects its algorithmically determined visibility. To unravel this, future research should explore the influence of counterspeaker status in naturalistic online environments and confirm its causal effect by manipulating status in controlled settings.

Audience Effects

Moreover, I observed discrepancies between bystander effects in controlled settings and in the field study that I attributed to different bystander audiences. In controlled environments, counterspeech prompted bystanders to engage in counterspeech, whereas, on social media, it primarily reduced bystander hate speech. Moreover, collective norms and severity assessments motivated bystanders in controlled experiments but probably did not affect their behavior on Twitter. These findings highlight the importance of the context for counterspeech effectiveness. Future research should further explore counterspeech impacts across diverse real-life bystander groups, such as news article audiences or participants in large discussion threads, for a more comprehensive assessment of its effects.

Mediator Interactions

On a different note, my dissertation provides robust evidence that perceptions of social norms and hate speech severity mediate the positive effect of counterspeech on bystanders and transgressors. Future research could explore the circumstances under which they might amplify or attenuate each other's effects. My results also showed that the mechanisms' impact was modulated by their recipients. The median user seemed more affected by severity assessments and polarized individuals were more susceptible to their respective ingroup norms. To enable targeted counterspeech interventions, future research could further explore the recipient characteristics that determine which mechanism proves effective and confirm their interplay in controlled settings. Given my findings on their significant impact, understanding the interplay between severity and social norm assessments could open a route to more effective strategies for combating online hate.

Automated Application

In addition, for counterspeech to effectively complement automated deletion, it must also be scalable, including the potential for automation. My research showed how the identity and status of an ostensibly human counterspeaker can significantly influence outcomes on social media. Therefore, it remains uncertain how automated counterspeech by non-human agents would impact its audience. Encouragingly, some research suggests that artificial agents can have a similar effect on norm perceptions as human ones (Nass & Moon, 2000; Xu & Lombard, 2017). However, in another survey study, I found that a substantial number of people would be very reactant to social media tools that aim to increase their propensity to engage in counterspeech (Cypris et al., in preparation). Therefore, my research offers promising insights into a possible application of automated counterspeech against online hate speech, but further research is essential to validate its effectiveness and gather additional support for scalable solutions.

Recommendations

For people who wish to combat online hate speech, this dissertation's results offer two core recommendations:

If you see something, say something (if you feel up to it): Counterspeech positively affects bystanders and transgressors.

Counterspeech can positively affect transgressors and bystanders, if not directly, then over time. However, this should not translate into feelings of a moral obligation to openly oppose hate speech. Counterspeakers might have a very reasonable desire to avoid personal consequences, especially if they themselves come from a marginalized group.

Remind people around you that online hate speech is at least as harmful and hurtful as its offline counterpart. It is commendable to be morally courageous and openly oppose it.

This dissertation identified social norms and perceptions of hate speech severity as two central mechanisms through which counterspeech can influence its audience. These mechanisms do not necessarily need to be leveraged via counterspeech. Instead, talking to friends, family, and peers could also have a substantial impact. Online hate speech is still widely trivialized and opposing it perceived as futile. These misconceptions may necessitate a broader change in societal perceptions, starting with each of us.

References

ADL. (2021). *Online hate and harassment: The american experience 2021*.

ADL. (2023). *Online hate and harassment: The american experience 2023*.

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

Altemeyer, R. A. (1981). *Right-wing authoritarianism*. University of Manitoba Press.

Alvarez-Benjumea, A. (2022). Uncovering hidden opinions: Social norms and the expression of xenophobic attitudes. *European Sociological Review*, jcac056. https://doi.org/10.1093/esr/jcac056

Alvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Amichai-Hamburger, Y. (2017). *Internet psychology: The basics* (1st ed.). Routledge.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, *337*(6092), 337–341. https://doi.org/10.1126/science.1215842

Arel-Bundock, V. (2023). *marginaleffects: Marginal effects, marginal means, predictions, and contrasts* [Manual]. https://vincentarelbundock.github.io/marginaleffects/

Aron, A., Aron, E. N., & Smollan, D. (1993). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596. https://doi.org/10.1037/0022-3514.63.4.596

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35.

Avenanti, A., Sirigu, A., & Aglioti, S. M. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, *20*(11), 1018–1022. https://doi.org/10.1016/j.cub.2010.03.071

Bandura, A. (2002). Social cognitive theory in cultural context. *Applied Psychology*, *51*(2), 269–290. https://doi.org/10.1111/1464-0597.00092

Banyard, V. L., & Moynihan, M. M. (2011). Variation in bystander behavior related to sexual and intimate partner violence prevention: Correlates in a sample of college students. *Psychology of Violence*, *1*(4), 287–301. https://doi.org/10.1037/a0023544

Barberá, P. (2015). Birds of the same feather tweet together. Bayesian ideal point estimation using twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/doi:10.1093/pan/mpu011

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barrie, C., & Ho, J. C. (2021). academictwitteR: an R package to access the twitter academic research product track v2 API endpoint. *Journal of Open Source Software*, *6*(62), 3272. https://doi.org/10.21105/joss.03272

Bastiaensens, S., Pabian, S., Vandebosch, H., Poels, K., Cleemput, K. V., DeSmet, A., & Bourdeaudhuij, I. D. (2016). From normative influence to social pressure: How relevant others affect whether bystanders join in cyberbullying. *Social Development*, *25*(1), 193–211. https://doi.org/10.1111/sode.12134

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, *31*, 259–271. https://doi.org/10.1016/j.chb.2013.10.036

Baumert, A., Halmburger, A., & Schmitt, M. (2013). Interventions against norm violations: Dispositional determinants of self-reported and real moral courage. *Personality and Social Psychology Bulletin*, *39*(8), 1053–1068. https://doi.org/10.1177/0146167213490032

Baumert, A., Li, M., Sasse, J., & Skitka, L. (2020). Standing up against moral violations: Psychological processes of moral courage. *Journal of Experimental Social Psychology*, *88*, 103951. https://doi.org/10.1016/j.jesp.2020.103951

Bennett, J. E., & Sekaquaptewa, D. (2014). Setting an egalitarian social norm in the classroom: Improving attitudes towards diversity among male engineering students. *Social Psychology of Education*, *17*(2), 343–355. https://doi.org/10.1007/s11218-014-9253-y

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When authoritarians confront prejudice. differential effects of sdo and rwa on support for hate-speech prohibition: When authoritarians confront prejudice. *Political Psychology*, *38*(1), 87–99. https://doi.org/10.1111/pops.12313

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, *47*(3), 260–266. https://doi.org/10.1002/ab.21948

Bizumic, B., & Duckitt, J. (2018). Investigating right wing authoritarianism with a very short authoritarianism scale. *Journal of Social and Political Psychology*, *6*(1), 129–150. https://doi.org/10.5964/jspp.v6i1.835

Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, *79*(6), 993–997. https://doi.org/10.1037/0021-9010.79.6.993

Bor, A., & Petersen, M. B. (2021). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, 1–18. https://doi.org/10.1017/S0003055421000885

Bowes-Sperry, L., & O'Leary-Kelly, A. M. (2005). To act or not to act: The dilemma faced by sexual harassment observers. *Academy of Management Review*, *30*(2), Article 2. https://doi.org/10.5465/amr.2005.16387886

Brady, W. J., Crockett, M., & Van Bavel, J. J. (2019). *The mad model of moral contagion: The role of motivation, attention and design in the spread of moralized content online* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/pz9g6

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Brown, A. L., Banyard, V. L., & Moynihan, M. M. (2014). College students as helpful bystanders against sexual violence: Gender, race, and year in college moderate the impact of perceived peer norms. *Psychology of Women Quarterly*, *38*(3), 350–362. https://doi.org/10.1177/0361684314526855

Buerger, C. (2021). #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, *7*(4), 205630512110638. https://doi.org/10.1177/20563051211063843

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Cary, L. A., Axt, J., & Chasteen, A. L. (2020). The interplay of individual differences, norms, and group identification in predicting prejudiced behavior in online video game interactions. *Journal of Applied Social Psychology*, *50*(11), 623–637. https://doi.org/10.1111/jasp.12700

Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*(1), e12890. https://doi.org/10.1111/phc3.12890

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (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*, 1201–1213. https://doi.org/10.1145/2818048.2819963

Chaney, K. E., & Sanchez, D. T. (2018). The endurance of interpersonal confrontations as a prejudice reduction strategy. *Personality and Social Psychology Bulletin*, *44*(3), 418–429. https://doi.org/10.1177/0146167217741344

Chekol, M. A., Moges, M. A., & Nigatu, B. A. (2023). Social media hate speech in the walk of Ethiopian political reform: Analysis of hate speech prevalence, severity, and natures. *Information, Communication & Society*, *26*(1), 218–237. https://doi.org/10.1080/1369118X.2021.1942955

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the*

2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17, 1217–1230. https://doi.org/10.1145/2998181.2998213

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015.

Citron, D. K. (2014). *Hate crimes in cyberspace*. Harvard University Press. https://doi.org/10.4159/harvard.9780674735613

Coles, S. M., & Lane, D. S. (2023). Making the impossible possible? Framing confrontations of racism on social media as norm-setting. *New Media & Society*, 14614448231208707. https://doi.org/10.1177/14614448231208707

Council of the European Union. (2008). *COUNCIL FRAMEWORK DECISION 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN

Cowan, G., & Hodge, C. (1996). Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, *26*(4), 355–374. https://doi.org/10.1111/j.1559-1816.1996.tb01854.x

Cowan, G., & Khatchadourian, D. (2003). Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychology of Women Quarterly*, *27*(4), 300–308. https://doi.org/10.1111/1471-6402.00110

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Crosby, J. R., & Monin, B. (2013). How the opinions of racial minorities influence judgments of discrimination. *Basic and Applied Social Psychology*, *35*(4), 334–345. https://doi.org/10.1080/01973533.2013.803963

Crosby, J. R., & Wilson, J. (2015). Let's not, and say we would: Imagined and actual responses to witnessing homophobia. *Journal of Homosexuality*, *62*(7), 957–970. https://doi.org/10.1080/00918369.2015.1008284

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., Bond, M. H., Croizet, J.-C., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H.-J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez-Bailón, R., Morales, E., Moya, M., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33. https://doi.org/10.1348/014466608X314935

Cypris, N. F., Engelmann, S., Sasse, J., Grossklags, J., & Baumert, A. (2022). Intervening against online hate speech: A case for automated counterspeech. *IEAI Research Brief, Technical University of Munich.*

Cypris, N. F., Willing, H., Sasse, J., & Baumert, A. (in preparation). *Profiling Ethics*.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784–803. https://doi.org/10.1037/0022-3514.90.5.784

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4, Pt.1), Article 4, Pt.1. https://doi.org/10.1037/h0025589

Delgado, R., & Stefancic, J. (2014). Hate speech in cyberspace. *Wake Forest Law Review*, *49*(2), 319–344.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, *57*, 398–415. https://doi.org/10.1016/j.chb.2015.12.051

Dhorida, A. (2017, November 21). Unsocial media: The real toll of online abuse against women. *Amnesty Insights*. https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4

Dillon, K. P., & Bushman, B. J. (2015). Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, *45*, 144–150. https://doi.org/10.1016/j.chb.2014.12.009

Doosje, B., Ellemers, N., & Spears, R. (1995). Perceived intragroup variability as a function of group status and identification. *Journal of Experimental Social Psychology*, *31*(5), 410–436.

Downs, D. M., & Cowan, G. (2012). Predicting the importance of freedom of speech and the perceived harm of hate speech. *Journal of Applied Social Psychology*, *42*(6), 1353–1375. https://doi.org/10.1111/j.1559-1816.2012.00902.x

Dreißigacker, A., Müller, P., Isenhardt, A., & Schemmel, J. (2024). Online hate speech victimization: Consequences for victims' feelings of insecurity. *Crime Science*, *13*(1), 4. https://doi.org/10.1186/s40163-024-00204-y

Drury, B. J. (2013). *Confronting for the greater good: Are confrontations that address the broad benefits of* [Doctoral Dissertation, University of Washington]. https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/22767/Drury_washington_0250E_11507.pdf?sequence=1&isAllowed=y

Drury, B. J., & Kaiser, C. R. (2014). Allies against sexism: The role of men in confronting sexism. *Journal of Social Issues*, *70*(4), 637–652. https://doi.org/10.1111/josi.12083

Ellemers, N., Van Der Toorn, J., Paunov, Y., & Van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366. https://doi.org/10.1177/1088868318811759

Ellingsen, T., & Johannesson, M. (2008). Anticipated verbal feedback induces altruistic behavior. *Evolution and Human Behavior*, *29*(2), 100–105. https://doi.org/10.1016/j.evolhumbehav.2007.11.001

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 572–595). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195126013.003.0029

Espelage, D., Green, H., & Polanin, J. (2012). Willingness to intervene in bullying episodes among middle school students: Individual and peer-group influences. *The Journal of Early Adolescence*, *32*(6), 776–801. https://doi.org/10.1177/0272431611423017

Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., & Kralj Novak, P. (2022). Retweet communities reveal the main sources of hate speech. *PLOS ONE*, *17*(3), e0265602. https://doi.org/10.1371/journal.pone.0265602

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), Article 6868. https://doi.org/10.1038/415137a

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517. https://doi.org/10.1037/a0023304

Forgiarini, M., Gallucci, M., & Maravita, A. (2011). Racism and the empathy for pain on our skin. *Frontiers in Psycholog*, *2*. https://doi.org/10.3389/fpsyg.2011.00108

Forsa. (2023). *Forsa-Befragung zur Wahrnehmung von Hassrede*. https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html

Forscher, P. S., Cox, W. T. L., Graetz, N., & Devine, P. G. (2015). The motivation to express prejudice. *Journal of Personality and Social Psychology*, *109*(5), 791–812. https://doi.org/10.1037/pspi0000030

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

FRA. (2023). *Online Content Moderation – Current challenges in detecting hate speech*. European Union Agency for Fundamental Rights. doi:10.2811/923316

Frey, D., Peus, C., Brandstätter, V., Winkler, M., Fischer, P., Bierhoff, H.-W., & Frey, D. (2006). Zivilcourage. In *Handbuch der Sozialpsychologie und Kommunikationspsychologie* (pp. 180–186). Hogrefe. https://www.zora.uzh.ch/id/eprint/98092/

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the fourth workshop on online abuse and harms* (pp. 102–112). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.13

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11*(1), 3. https://doi.org/10.1140/epjds/s13688-021-00314-6

Goel, V., Sahnan, D., Dutta, S., Bandhakavi, A., & Chakraborty, T. (2023). Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS Nexus*, *2*(3), pgad041. https://doi.org/10.1093/pnasnexus/pgad041

Goldenberg, A., & Gross, J. J. (2020). Digital emotion contagion. *Trends in Cognitive Sciences*, S1364661320300279. https://doi.org/10.1016/j.tics.2020.01.009

Goodwin, R., Graham, J., & Diekmann, K. A. (2020). Good intentions aren't good enough: Moral courage in opposing sexual harassment. *Journal of Experimental Social Psychology*, *86*, 103894. https://doi.org/10.1016/j.jesp.2019.103894

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Greitemeyer, T., Fischer, P., Kastenmüller, A., & Frey, D. (2006). Civil courage and helping behavior: Differences and similarities. *European Psychologist*, *11*(2), 90–98. https://doi.org/10.1027/1016-9040.11.2.90

Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: Neural evidence of an empathy gap. *Social Cognitive and Affective Neuroscience*, *7*(5), 596–603. https://doi.org/10.1093/scan/nsr035

Halmburger, A., Baumert, A., & Schmitt, M. (2015). Anger as driving factor of moral courage in comparison with guilt and global mood: A multimethod approach. *European Journal of Social Psychology*, *45*(1), 39–51. https://doi.org/10.1002/ejsp.2071

Halmburger, A., Baumert, A., & Schmitt, M. (2016). Everyday heroes: Determinants of moral courage. In S. T. Allison, G. R. Goethals, & R. M. Kramer (Eds.), *Handbook of Heroism and Heroic Leadership* (1st Edition, p. 20). Routledge.

Han, S.-H., & Brazeal, L. M. (2015). Playing nice: Modeling civility in online political discussions. *Communication Research Reports*, *32*(1), 20–28. https://doi.org/10.1080/08824096.2014.989971

Han, S.-H., Brazeal, L. M., & Pennington, N. (2018). Is civility contagious? Examining the impact of modeling in online political discussions. *Social Media + Society*, *4*(3), 205630511879340. https://doi.org/10.1177/2056305118793404

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), e2116310118. https://doi.org/10.1073/pnas.2116310118

Henson, B., Fisher, B. S., & Reyns, B. W. (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), 505–527. https://doi.org/10.1177/1077801219835050

Hickey, D., Schmitz, M., Fessler, D., Smaldino, P. E., Muric, G., & Burghardt, K. (2023). Auditing elon musk's impact on hate speech and bots. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 1133–1137. https://doi.org/10.1609/icwsm.v17i1.22222

Hogg, M. A., & Rinella, M. J. (2018). Social identities and shared realities. *Current Opinion in Psychology*, *23*, 6–10. https://doi.org/10.1016/j.copsyc.2017.10.003

Hogg, M. A., & Turner, J. C. (1987). Intergroup behaviour, self-stereotyping and the salience of social categories. *British Journal of Social Psychology*, *26*(4), 325–340. https://doi.org/10.1111/j.2044-8309.1987.tb00795.x

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100. https://doi.org/10.1207/s15324834basp1101_6

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors?: online comments on prejudice expression. *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Hulk, F. C. (2018, July 12). *Don't feed the trolls, and other hideous lies*. The Verge. https://www.theverge.com/2018/7/12/17561768/dont-feed-the-trolls-online-harassment-abuse

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1). https://doi.org/10.1214/10-STS321

Jagayat, A., Boparai, G., & Choma, B. L. (2021). *Mock Social Media Website Tool (1.0)* (1.0) [Computer software]. https://docs.studysocial.media

Jiménez Durán, R. (2022). The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4044098

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*(5911), 276–278. https://doi.org/10.1126/science.1164951

Kayser, D. N., Greitemeyer, T., Fischer, P., & Frey, D. (2010). Why mood affects help giving, but not moral courage: Comparing two types of prosocial behaviour. *European Journal of Social Psychology*, *40*(7), 1136–1157. https://doi.org/10.1002/ejsp.717

Keighley, R. (2022). Hate hurts: Exploring the impact of online hate on lgbtq+ young people. *Women & Criminal Justice*, *32*(1–2), 29–48. https://doi.org/10.1080/08974454.2021.1988034

Keipi, T., Räsänen, P., Oksanen, A., Hawdon, J., & Näsi, M. (2018). Exposure to online hate material and subjective well-being: A comparative study of American and Finnish youth. *Online Information Review*, *42*(1), 2–15. https://doi.org/10.1108/OIR-05-2016-0133

Klein, O., Spears, R., & Reicher, S. (2007). Social identity performance: Extending the strategic side of side. *Personality and Social Psychology Review*, *11*(1), 28–45. https://doi.org/10.1177/1088868306294588

Koehler, C., & Weber, M. (2018). ”Do I really need to help?!” Perceived severity of cyberbullying, victim blaming, and bystanders’ willingness to help the victim. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *12*(4). https://doi.org/10.5817/CP2018-4-4

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do “Good Citizens” fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, *18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

Kutlaca, M., Becker, J., & Radke, H. (2020). A hero for the outgroup, a black sheep for the ingroup: Societal perceptions of those who confront discrimination. *Journal of Experimental Social Psychology*, *88*, 103832. https://doi.org/10.1016/j.jesp.2019.103832

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn’t he help?* Appleton-Century Crofts.

Lee, E.-J. (2004). Effects of visual representation on social influence in computer-mediated communication. *Human Communication Research*, *30*(2), 234–259. https://doi.org/10.1111/j.1468-2958.2004.tb00732.x

Lee, E.-J. (2007a). Character-based team identification and referent informational influence in computer-mediated communication. *Media Psychology*, *9*(1), 135–155. https://doi.org/10.1080/15213260709336806

Lee, E.-J. (2007b). Deindividuation effects on group polarization in computer-mediated communication: The role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication*, *57*(2), 385–403. https://doi.org/10.1111/j.1460-2466.2007.00348.x

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues*, *58*(2), 341–361. https://doi.org/10.1111/1540-4560.00264

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication | Media*, *7*(4), 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

Lepoutre, M. (2017). Hate speech in public discourse: A pessimistic defense of counterspeech. *Social Theory and Practice*, *43*(4), 851–883. https://doi.org/10.5840/soctheorpract201711125

Levine, M., & Crowther, S. (2008). The responsive bystander: How social group membership and group size can encourage as well as inhibit bystander intervention. *Journal of Personality and Social Psychology*, *95*(6), 1429–1439. https://doi.org/10.1037/a0012634

Levine, M., & Manning, R. (2013). Social identity, group processes, and helping in emergencies. *European Review of Social Psychology*, *24*(1), 225–251. https://doi.org/10.1080/10463283.2014.892318

Levine, M., Prosser, A., Evans, D., & Reicher, S. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social Psychology Bulletin*, *31*(4), 443–453. https://doi.org/10.1177/0146167204271651

Ley, H. (2018). *#ichbinhier: Zusammen gegen Fake News und Hass im Netz: [#iamhere: Together against Fake News and Hate in the Net]*. DuMont.

Li, M., Sasse, Julia, Halmburger, A., & Baumert, A. (2021). Standing up against moral transgressions: An integrative perspective on the socio-psychological antecedents and barriers to moral courage. *Under Review*.

Li, R., Gordon, S., & Gelfand, M. J. (2017). Tightness–looseness: A new framework to understand consumer behavior. *Journal of Consumer Psychology*, *27*(3), 377–391. https://doi.org/10.1016/j.jcps.2017.04.001

Liebst, L. S., Philpot, R., Bernasco, W., Dausel, K. L., Ejbye-Ernst, P., Nicolaisen, M. H., & Lindegaard, M. R. (2019). Social relations and presence of others predict bystander intervention: Evidence from violent incidents captured on CCTV. *Aggressive Behavior*, *45*(6), 598–609. https://doi.org/10.1002/ab.21853

Lopez-Sanchez, M., & Müller, A. (2021). On simulating the propagation and countermeasures of hate speech in social networks. *Applied Sciences*, *11*(24), 12003. https://doi.org/10.3390/app112412003

Lu, S., & Luqiu, L. R. (2023). When will one help? Understanding audience intervention in online harassment of women journalists. *Journalism Practice*, 1–19. https://doi.org/10.1080/17512786.2023.2201582

Lumsden, K., & Morgan, H. (2017). Media framing of trolling and online abuse: Silencing strategies, symbolic violence, and victim blaming. *Feminist Media Studies*, *17*(6), 926–940. https://doi.org/10.1080/14680777.2017.1316755

Machackova, H., Dedkova, L., & Mezulanikova, K. (2015). Brief report: The bystander effect in cyberbullying incidents. *Journal of Adolescence*, *43*, 96–99. https://doi.org/10.1016/j.adolescence.2015.05.010

Majed, R. (2021). In defense of intra-sectarian divide: Street mobilization, coalition formation, and rapid realignments of sectarian boundaries in lebanon. *Social Forces*, *99*(4), 1772–1798. https://doi.org/10.1093/sf/soaa076

March, E. (2019). Psychopathy, sadism, empathy, and the motivation to cause harm: New evidence confirms malevolent nature of the Internet Troll. *Personality and Individual Differences*, *141*, 133–137. https://doi.org/10.1016/j.paid.2019.01.001

March, E., & Steele, G. (2020). High esteem and hurting others online: Trait sadism moderates the relationship between self-esteem and internet trolling. *Cyberpsychology, Behavior, and Social Networking*, cyber.2019.0652. https://doi.org/10.1089/cyber.2019.0652

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 173–182. https://doi.org/10.1145/3292522.3326034

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2019). Temporal effects of unmoderated hate speech in gab. *arXiv:1909.10966 [Cs]*. http://arxiv.org/abs/1909.10966

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 369–380.

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

Merriam-Webster. (2024). *Definition of racism*. Merriam-Webster.Com Dictionary. https://www.merriam-webster.com/dictionary/racism

Meta. (2024, February 24). *Hate speech | transparency center*. https://transparency.fb.com/policies/community-standards/hate-speech/

Meyer, G. (2014). Zivilcourage und ihr Kern. In *Mut und Zivilcourage: Grundlagen und gesellschaftliche Praxis*. Verlag Barbara Budrich.

Meyers, C., Leon, A., & Williams, A. (2020). Aggressive confrontation shapes perceptions and attitudes toward racist content online. *Group Processes & Intergroup Relations*, *23*(6), 845–862. https://doi.org/10.1177/1368430220935974

Mikal, J. P., Rice, R. E., Kent, R. G., & Uchino, B. N. (2014). Common voice: Analysis of behavior modification and content convergence in a popular online community. *Computers in Human Behavior*, *35*, 506–515. https://doi.org/10.1016/j.chb.2014.02.036

Ministerium für Inneres und Sport des Landes Sachsen-Anhalt. (2021). *Verfassungsschutzbericht des Landes Sachsen-Anhalt für das Jahr 2021*. Ministerium für Inneres und Sport des Landes Sachsen-Anhalt. https://mi.sachsen-anhalt.de/fileadmin/Bibliothek/Politik_und_Verwaltung/MI/MI/3._Themen/Verfassungsschutz/VSB_ST_2021_Endfassung_01.pdf

Miškolci, J., Kováčová, L., & Rigová, E. (2018). Countering hate speech on facebook: The case of the roma minority in slovakia. *Social Science Computer Review*, *38*(2), 128–146. https://doi.org/10.1177/0894439318791786

Molina, R. G., & Jennings, F. J. (2018). The role of civility and metacommunication in facebook discussions. *Communication Studies*, *69*(1), 42–66. https://doi.org/10.1080/10510974.2017.1397038

Moor, L., & Anderson, J. R. (2019). A systematic literature review of the relationship between dark personality traits and antisocial online behaviours. *Personality and Individual Differences*, *144*, 40–55. https://doi.org/10.1016/j.paid.2019.02.027

Morris, M. W., Hong, Y., Chiu, C., & Liu, Z. (2015). Normology: Integrating insights about social norms to understand cultural dynamics. *Organizational Behavior and Human Decision Processes*, *129*, 1–13. https://doi.org/10.1016/j.obhdp.2015.03.001

Mullen, B., & Smyth, J. M. (2004). Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine*.

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167. https://doi.org/10.1093/jeea/jvaa045

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795. https://doi.org/10.1177/1461444816670923

Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social Science Computer Review*, *39*(2), 245–258. https://doi.org/10.1177/0894439319865518

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Ng, Y.-L., Song, Y., & Huang, Y. (2022). Supportive and uncivil expressions in discussions on out-groups by in-group members in anonymous computer-mediated communication. *Telematics and Informatics*, *69*, 101785. https://doi.org/10.1016/j.tele.2022.101785

Niesta Kayser, D., Frey, D., Kirsch, F., Brandstätter, V., & Agthe, M. (2016). Zivilcourage. In H.-W. Bierhoff & D. Frey (Eds.), *Soziale Motive und soziale Einstellungen* (Vol. 2, pp. 255–275). Hogrefe. https://epub.ub.uni-muenchen.de/56801/

Niesta Kayser, D., Greitemeyer, T., Fischer, P., & Frey, D. (2010). Why mood affects help giving, but not moral courage: Comparing two types of prosocial behaviour. *European Journal of Social Psychology*, *40*(7), Article 7. https://doi.org/10.1002/ejsp.717

Noel, J. G., Wann, D. L., & Branscombe, N. R. (1995). Peripheral ingroup membership status and public negativity toward outgroups. *Journal of Personality and Social Psychology*, *68*, 127–137.

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, *41*(1), 100–126. https://doi.org/10.1111/j.1745-6606.2006.00070.x

Nosko, A., Wood, E., & Molema, S. (2010). All about me: Disclosure in online social networking profiles: The case of FACEBOOK. *Computers in Human Behavior*, *26*(3), 406–418. https://doi.org/10.1016/j.chb.2009.11.012

Obermaier, M. (2022). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, 146144482211254. https://doi.org/10.1177/14614448221125417

Obermaier, M., Fawzi, N., & Koch, T. (2015). Bystanderintervention bei Cybermobbing. *Studies in Communication | Media*, *4*(1), 28–52. https://doi.org/10.5771/2192-4007-2015-1-28

Obermaier, M., Schmid, U. K., & Rieger, D. (2023). Too civil to care? How online hate speech against different social groups affects bystander intervention. *European Journal of Criminology*, *20*(3), 817–833. https://doi.org/10.1177/14773708231156328

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

Osswald, S., Frey, D., & Streicher, B. (2011). Moral Courage. In E. Kals & J. Maes (Eds.), *Justice and Conflicts* (pp. 391–405). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19035-3_24

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society*, *6*(2), 205630512091685. https://doi.org/10.1177/2056305120916850

Packer, D. J. (2014). On not airing our dirty laundry: Intergroup contexts suppress ingroup criticism among strongly identified group members. *British Journal of Social Psychology*, *53*(1), 93–111. https://doi.org/10.1111/bjso.12017

Paluck, E. L., & Shepherd, H. (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology*, *103*(6), 899–915. https://doi.org/10.1037/a0030015

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571. https://doi.org/10.1073/pnas.1514483113

Perry, D. G., & Bussey, K. (1979). The social learning theory of sex differences: Imitation is alive and well. *Journal of Personality and Social Psychology*, *37*(10), 1699–1712. https://doi.org/10.1037/0022-3514.37.10.1699

Pluta, A., Mazurek, J., Wojciechowski, J., Wolak, T., Soral, W., & Bilewicz, M. (2023). Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports*, *13*(1), 4127. https://doi.org/10.1038/s41598-023-31146-1

Postmes, T., Spears, R., Sakhel, K., & de Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, *27*(10), 1243–1254. https://doi.org/10.1177/01461672012710001

Postmes, T., & Turner, F. M. (2015). Deindividuation, Psychology of. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 38–41). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.24015-4

Quinn, K., Epstein, D., & Moon, B. (2019). We care about different things: Non-elite conceptualizations of social media privacy. *Social Media + Society*, *5*(3), 205630511986600. https://doi.org/10.1177/2056305119866008

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 557–568.

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2021). Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, *42*(9), 1100–1111. https://doi.org/10.1080/01639625.2020.1722337

Reicher, S., & Levine, M. (1994). On the consequences of deindividuation manipulations for the strategic communication of self: Identifiability and the presentation of social identity. *European Journal of Social Psychology*, *24*(4), 511–524. https://doi.org/10.1002/ejsp.2420240408

Reicher, S., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198. https://doi.org/10.1080/14792779443000049

Rendsvig, R. K. (2014). Pluralistic ignorance in the bystander effect: Informational dynamics of unresponsive witnesses in situations calling for intervention. *Synthese*, *191*(11), 2471–2498. https://doi.org/10.1007/s11229-014-0435-0

Rosenberg, M. J. (2009). The conditions and consequences of evaluation apprehension. In *Artifacts in behavioral research: Robert rosenthal and ralph L. Rosnow's classic books* (pp. 211–263). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195385540.003.0007

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. M. (2016, November). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer- Mediated Communication*. https://doi.org/10.17185/duepublico/42132

Rovira, A., Southern, R., Swapp, D., Campbell, C., Zhang, J. J., Levine, M., & Slater, M. (2021). Bystander affiliation influences intervention behavior: A virtual reality study. *SAGE Open*, *11*(3), 215824402110400. https://doi.org/10.1177/21582440211040076

Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. (2023). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*, *42*(5), 527–544. https://doi.org/10.1080/0144929X.2022.2027013

Saab, R., Tausch, N., Spears, R., & Cheung, W.-Y. (2015). Acting in solidarity: Testing an extended dual pathway model of collective action by bystander group members. *British Journal of Social Psychology*, *54*(3), 539–560. https://doi.org/10.1111/bjso.12095

Sabini, J., Siepmann, M., & Stein, J. (2001). Target article: "the really fundamental attribution error in social psychological research." *Psychological Inquiry*, *12*(1), 1–15. https://doi.org/10.1207/S15327965PLI1201_01

Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, *37*(1), 2166719. https://doi.org/10.1080/08839514.2023.2166719

Sasse, J., Cypris, N. F., & Baumert, A. (2023). Online moral courage. In C. Cohrs, N. Knab, & G. Sommer (Eds.), *Handbook of Peace Psychology*. https://doi.org/10.17192/es2022.0074

Sasse, J., Halmburger, A., & Baumert, A. (2022). The functions of anger in moral courage—Insights from a behavioral study. *Emotion*, *22*(6), 1321–1335. https://doi.org/10.1037/emo0000906

Sasse, J., Li, M., & Baumert, A. (2022). How prosocial is moral courage? *Current Opinion in Psychology*, *44*, 146–150. https://doi.org/10.1016/j.copsyc.2021.09.004

Schaake, M. (2020, September 29). How democracies can claim back power in the digital world. *MIT Technology Review*. https://www.technologyreview.

com/2020/09/29/1009088/democracies-power-digitalsocial-media-governance-tech-companies-opinion

Schieb, C., & Preuss, M. (2016). *Governing hate speech by means of counter speech on facebook*. 25.

Schieb, C., & Preuss, M. (2018). Considering the elaboration likelihood model for simulating hate and counter speech on facebook. *Studies in Communication | Media*, *7*(4), 580–606. https://doi.org/10.5771/2192-4007-2018-4-580

Schmid, U. K. (2023). Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*, 14614448231198169. https://doi.org/10.1177/14614448231198169

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, 146144482210911. https://doi.org/10.1177/14614448221091185

Schubert, T. W., & Otten, S. (2002). Overlap of self, ingroup, and outgroup: Pictorial measures of self-categorization. *Self and Identity*, *1*(4), 353–376. https://doi.org/10.1080/152988602760328012

Seering, J., Kraut, R., & Dabbish, L. (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 - CSCW '17*, 111–125. https://doi.org/10.1145/2998181.2998277

Sharma, A., & Kaushal, R. (2023). Detecting hate speech in hindi in online social media. *2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, 1–5. https://doi.org/10.1109/ICCT56969.2023.10075749

Siegel, A. A., & Badaan, V. (2020). #no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, *114*(3), 837–855. https://doi.org/10.1017/S0003055420000283

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, *49*(2), 147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x

Son, J., Lee, J., Oh, O., Lee, H. K., & Woo, J. (2020). Using a Heuristic-Systematic Model to assess the Twitter user profile's impact on disaster tweet credibility. *International Journal of Information Management*, *54*, 102176. https://doi.org/10.1016/j.ijinfomgt.2020.102176

Song, J., & Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. *Computers in Human Behavior*, *78*, 273–282. https://doi.org/10.1016/j.chb.2017.10.008

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Soral, W., Liu, J., & Bilewicz, M. (2020). Media of contempt: Social media consumption predicts normative acceptance of anti-muslim hate speech and islamoprejudice. *International Journal of Conflict and Violence (IJCV)*, 1-13 Pages. https://doi.org/10.4119/IJCV-3774

Soral, W., Świderska, A., Puchała, D., & Bilewicz, M. (2023). Desensitization to hate speech: Examination using heart rate measurement. *Aggressive Behavior*, ab.22118. https://doi.org/10.1002/ab.22118

Spears, R., & Postmes, T. (2015). Group identity, social influence, and collective action online: Extensions and applications of the side model. In S. S. Sundar (Ed.), *The Handbook of the Psychology of Communication Technology* (1st ed., pp. 23–46). Wiley. https://doi.org/10.1002/9781118426456.ch2

Staub, E. (2015). *The Roots of Goodness and Resistance to Evil: Inclusive Caring, Moral Courage, Altruism Born of Suffering, Active Bystandership, and Heroism*. Oxford University Press.

Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on tiktok. *Social Media + Society*, *9*(3), 20563051231194586. https://doi.org/10.1177/20563051231194586

Stone, J. (2011). Thanks for asking: Self-affirming questions reduce backlash when stigmatized targets confront prejudice. *Journal of Experimental Social Psychology*.

Strossen, N. (2020). *Hate: Why we should resist it with free speech, not censorship* (First issued as an Oxford University Press paperback). Oxford University Press.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, 321–326.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton University Press. http://www.jstor.org/stable/j.ctt7tbsw

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In W. G. Austin & S. Worchel (Eds.), *Psychology of intergroup relations* (2nd ed). Nelson-Hall Publishers.

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change: Vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Tankard, M. E., & Paluck, E. L. (2017). The effect of a supreme court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science*, *28*(9), 1334–1344. https://doi.org/10.1177/0956797617709594

Tarrant, M., Dazeley, S., & Cottom, T. (2009). Social categorization and empathy for outgroup members. *British Journal of Social Psychology*, *48*(3), 427–446.

Täuber, S., & van Zomeren, M. (2013). Outrage towards whom? Threats to moral group status impede striving to improve via out-group-directed outrage: The role of

outrage focus on moral motivation. *European Journal of Social Psychology*, *43*(2), 149–159. https://doi.org/10.1002/ejsp.1930

Tepper, M. (1997). Usenet communities and the cultural politics of information. In *Internet culture* (pp. 39–54). Routledge.

TikTok. (2022, December 8). *Countering hate on TikTok*. TikTok. https://www.tiktok.com/safety/en/countering-hate/

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). **mediation**: *R* package for causal mediation analysis. *Journal of Statistical Software*, *59*(5). https://doi.org/10.18637/jss.v059.i05

Toribio-Flórez, D., Saße, J., & Baumert, A. (2023). "Proof under reasonable doubt": Ambiguity of the norm violation as boundary condition of third-party punishment. *Personality and Social Psychology Bulletin*, *49*(3), 429–446. https://doi.org/10.1177/01461672211067675

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* Basil Blackwell.

Turner, J. C., & Reynolds, K. J. (2011). Self-categorization theory. In P. Van Lange, A. Kruglanski, & E. Higgins (Eds.), *Handbook of theories in social psychology* (Vol. 2, pp. 399–417). SAGE Publications Ltd.

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health*, *43*(6), 565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

United Nations. (2020). *United nations strategy and plan of action on hate speech detailed guidance on implementation for united nations field presences*.

Urbaniak, R., Ptaszyński, M., Tempska, P., Leliwa, G., Brochocki, M., & Wroczyński, M. (2022). Personal attacks decrease user activity in social networking platforms. *Computers in Human Behavior*, *126*, 106972. https://doi.org/10.1016/j.chb.2021.106972

Van Bommel, M., Van Prooijen, J.-W., Elffers, H., & Van Lange, P. A. M. (2012). Be aware to care: Public self-awareness leads to a reversal of the bystander effect. *Journal of Experimental Social Psychology*, *48*(4), 926–930. https://doi.org/10.1016/j.jesp.2012.02.011

Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine "helping," "joining in," and "doing nothing" when witnessing cyberbullying: Bystander Behavior in Cyberbullying. *Aggressive Behavior*, *40*(5), 383–396. https://doi.org/10.1002/ab.21534

van der Ploeg, R., Kretschmer, T., Salmivalli, C., & Veenstra, R. (2017). Defending victims: What does it take to intervene in bullying and how is it rewarded by peers? *Journal of School Psychology*, *65*, 1–10. https://doi.org/10.1016/j.jsp.2017.06.002

van der Wilk, A. (2018). *Cyber violence and hate speech online against women* (Women's Rights & Gender Equality). Policy Department for Citizen's Rights and Constitutional Affairs. http://www.europarl.europa.eu/supporting-analyses

Van Houtven, E., Acquah, S. B., Obermaier, M., Saleem, M., & Schmuck, D. (2024). 'You got my back?' severity and counter-speech in online hate speech toward minority groups. *Media Psychology*, 1–32. https://doi.org/10.1080/15213269.2023.2298684

Voelpel, S. C., Eckhoff, R. A., & Förster, J. (2008). David against Goliath? Group size and bystander effects in virtual knowledge sharing. *Human Relations*, *61*(2), 271–295. https://doi.org/10.1177/0018726707087787

Vogels, E. A. (2021). *The State of Online Harassment*. 54.

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Kansok-Dusche, J., Krause, N., & Ballaschk, C. (2022). Associations between witnessing and perpetrating online hate speech among adolescents: Testing moderation effects of moral disengagement and empathy. *Psychology of Violence*, *12*(6), 371–381. https://doi.org/10.1037/vio0000422

Wachs, S., Wettstein, A., Bilz, L., Krause, N., Ballaschk, C., Kansok-Dusche, J., & Wright, M. F. (2021). Playing by the rules? An investigation of the relationship between social norms and adolescents' hate speech perpetration in schools. *Journal of Interpersonal Violence*, 088626052110560. https://doi.org/10.1177/08862605211056032

Wachs, S., & Wright, M. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International Journal of Environmental Research and Public Health*, *15*(9), 2030. https://doi.org/10.3390/ijerph15092030

Weber, M., Viehmann, C., Ziegele, M., & Schemer, C. (2020). Online hate does not stay online – how implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in Human Behavior*, *104*, 106192. https://doi.org/10.1016/j.chb.2019.106192

Whiley, L. A., Walasek, L., & Juanchich, M. (2023). Contributions to reducing online gender harassment: Social re-norming and appealing to empathy as tried-and-failed techniques. *Feminism & Psychology*, *33*(1), 83–104. https://doi.org/10.1177/09593535221104874

Whitney v. California (U.S. 1927). https://supreme.justia.com/cases/federal/us/274/357/

Wiktionary: English ethnic slurs. (2022). *Category: English ethnic slurs—Wiktionary, The free dictionary*. https://en.wiktionary.org/wiki/Category:English_ethnic_slurs

Wilder, D. A. (1990). Some determinants of the persuasive power of in-groups and out-groups: Organization of information and attribution of independence. *Journal of Personality and Social Psychology*, *59*(6), 1202.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research*, *47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Windisch, S., Wiedlitzka, S., Olaghere, A., & Jenaway, E. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, *18*(2). https://doi.org/10.1002/cl2.1243

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, *65*, 7–19. https://doi.org/10.1016/j.jesp.2016.02.005

Wypych, M., & Bilewicz, M. (2024). Psychological toll of hate speech: The role of acculturation stress in the effects of exposure to ethnic slurs on mental health among Ukrainian immigrants in Poland. *Cultural Diversity and Ethnic Minority Psychology*, *30*(1), 35–44. https://doi.org/10.1037/cdp0000522

Xiao, E., & Houser, D. (2009). Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange. *Journal of Economic Psychology*, *30*(3), 393–404. https://doi.org/10.1016/j.joep.2008.12.002

Xu, K., & Lombard, M. (2017). Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior*, *74*, 152–162. https://doi.org/10.1016/j.chb.2017.04.043

Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*(5), 731–751. https://doi.org/10.1177/1461444819870130

Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, *43*(6), 867–876. https://doi.org/10.1016/j.jesp.2006.10.010

Zufall, F., Horsmann, T., & Zesch, T. (2019). From legal to technical concept: Towards an automated classification of German political Twitter postings as criminal offenses. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1337–1347. https://doi.org/10.18653/v1/N19-1135

Acknowledgments

*“*nos esse quasi nanos gigantum umeris insidentes*”* - *Bernard of Chartres*

Ich möchte mich zum Schluss noch bei einigen tollen Menschen bedanken, die mich während meiner Dissertation mit Rat und Tat, intellektueller und emotionaler Unterstützung und Zeit und Zuwendung bedacht haben.

Danke Anna und Julia für eure Hilfe und Anleitung während der Dissertation! Ich hatte und habe wirklich ein riesiges Glück mit euch beiden als Betreuerinnen. Ich habe unglaublich davon profitiert, mit zwei so beeindruckenden Forscherinnen zusammenzuarbeiten und von euch intellektuell und in meinen Fähigkeiten als Wissenschaftler derart gefördert und gefordert zu werden. Aber auch auf menschlicher Ebene bin ich wirklich dankbar für die Zusammenarbeit. Ich hoffe, dass wir das noch auf die eine oder andere Art und Weise in Zukunft fortführen werden!

Danke auch an Jens und Severin! Unsere Zoomcalls und die Treffen in München waren immer eine große Freude und euer Input super bereichernd!

Grüße gehen raus an den Lehrstuhl in Wuppertal! Slieman, Lisa, Carla, Sascha, Axel, Marie, Julia, Leonie, Lisa, Lukas, Lina, Inga und alle anderen: Es ist großartig, wenn man sich jedes Mal drauf freut, zur Arbeit zu kommen, weil die Kolleg:innen so super sind. Das war sogar die Pendelei aus München wert.

I would also like to thank Dani, Aya, Fiona, and Mengyao who I wish I could have spent more time with in person at the Max Planck Institute.

Neben den Schultern zum (intellektuell) Draufsitzen möchte ich mich auch bei den Schultern zum Anlehnen bedanken. Vielen Dank für eure Hilfe und euren Rat und dass ihr einfach immer für mich da seid, Mama, Papa, Mona, Lena, Kerstin und David. Liebe Grüße auch an Oma, Opa, Omi und Opi. Ich bin wirklich glücklich, euch alle in meinem Leben zu haben - in Persona oder in vielen meiner Eigenschaften und Erinnerungen.

Also, thank you so much for sharing this journey with me, Kaley. I have been incredibly lucky to have you in my life. I am immensely grateful for our talks about my dissertation and your reassurance and support throughout this whole process.

Weiterhin möchte ich meinen Freund:innen danken, die mich durch den Weg begleitet haben. Besonders Matthias, Jan Philipp, Maxime, Reichi, Mödder, Marcel und Markus. Danke, dass ihr alle so großartig seid! Also, Slieman, I am exalted that you are not just my colleague from Wuppertal but also my friend from Munich. You are truly a wonderful guy.

Vielen Dank euch allen und all den weiteren Menschen, die mich im während meiner Dissertationszeit unterstützt und geprägt haben!