Encouraging User-Initiated Counter Speech for Governing Hate Speech on Social Media Platforms

The Effectiveness of a Short Perspective-Taking Encouragement Through a Causal Chain of Dehumanization, Empathy and Threat Perceptions

Name: Sarah-Luisa Simon
Student ID:
Course: Master's Thesis in OD & D (MAN-MTHODA)
Supervisor: Dr. Stefan Schembera
Second Examiner: Dr. Claudia Groß
Institute: Radboud University Nijmegen
Due Date Master Thesis: 15.08.2022

Acknowledgement

Before you lies my master thesis for the specialization "Organisational Design & Development" of the Business Administration master program at Radboud University. This thesis marks the end of a challenging yet highly educational, insightful and interesting process. A couple of years ago when I was still busy with my former education at a different faculty, I would have never imagined that one day, I will hand in a thesis for a business-related program. Yet, despite all the challenges associated with taking on a new educational direction, I am extremely glad that I got the opportunity to do so, to form new academic experiences and thus to expand my professional horizon. Furthermore, I am grateful that I got the opportunity to combine the newly acquired knowledge from the Business Administration program with the psychological insights from my previous education in order to explore the interesting and highly important topic of how to contribute to the fight of hate speech on Social Media Platforms. I am proud that I have reached the moment to submit the present graduation project and hope that you will be as enthusiastic about the topic as I am.

Furthermore, I would like to take the present moment to express my gratitude to a number of people that supported me in the process of conducting the master thesis at hand. First of all, a word of gratitude to Dr. Stefan Schembera, who took on the role as supervisor for the present graduation project. He invested many hours to provide me with support and valuable feedback which improved the quality of the present master thesis tremendously and helped me to grow as an academic. Second of all, I would like to thank Dr. Claudia Groß, who acted as a second reader for my master thesis. She provided me with helpful feedback on my research proposal and offered swift support when I had a question regarding an ethical aspect of my project. Third of all, I want to express my gratitude to my friends and family for their patience and emotional support within the present process as well as to the participants of the present study – without you this project would not have been possible.

Sarah-Luisa Simon
Nijmegen, August 2022

Abstract

The study at hand investigated the effectiveness of the provision of a short perspective-taking encouragement as a strategy for increasing social media users' intentions to counterargue against hateful content on Social Media Platforms (SMPs). In addition, the underlying mechanism of potential effectiveness was examined. In order to do so, 102 participating social media users were randomly assigned to either the experimental- or the control condition. Participants in the experimental condition received a short perspective-taking encouragement, meaning that they were asked to visualize what targets of hate speech may be feeling, thinking and experiencing, whereas participants in the control condition received a control task. Afterwards, all participants were presented with the same fictional hate speech post as well as the same questionnaires assessing their threat perceptions, meaning the extent to which they perceived the hate speech incident in question as being (potentially) harmful to those targeted by it and their counter speech intentions. Furthermore, all participants received the same questionnaire investigating their dehumanization of these hate speech targets, meaning the extent to which they denied human attributes to these victims and their empathy with the victims. The obtained data was analyzed with the SPSS PROCESS Macro. Results demonstrated that participating social media users in the experimental condition displayed significantly higher intentions to counterargue against the hate speech post than participants in the control task. Significant effects were serially mediated by dehumanization, empathy and threat perceptions. More specifically, receiving the perspective-taking encouragement led participating social media users in the experimental condition to dehumanize the targets of the hate speech post to a lower extent than participants in the control condition which in turn enhanced their empathy with these victims. This facilitated empathy with the targets, on the other hand, gave rise to increased threat perceptions, ultimately enhancing counter speech intentions. These results suggest that the provision of a short perspective-taking encouragement is an effective strategy for enhancing social media users counter speech intentions towards hate speech. Furthermore, findings suggest that this effectiveness is due to a causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions. Practical and theoretical implications as well as future research directions are discussed.

Keywords

Hate Speech Governance; Social Media; Counter Speech Intentions; Dehumanization; Threat Perceptions; Empathy; Perspective-Taking Encouragement; Cyberhate; Strategies

Table of Contents

Introduction

In today's modern information society, the usage of social media platforms (SMPs) such as Facebook or Twitter is among the most popular online activities with the amount of people utilizing these networks on a daily basis for communicating with friends, accessing entertaining content or important information and participating in public debates and dialogue all across the globe only being expected to increase in the coming years (Statista Research Development, 2022a, 2022b). Despite the associated benefits, the growth in popularity of these networks also carries individual and societal risks as they appear to provide a space in which cyberhate can flourish (Rieger, Schmitt, & Frischlich, 2018). In the recent years, a dramatic increase of hate speech, defined as "norm-transgressing communication that may derogate and defame individuals or targeted social groups" (Rieger et al., 2018, p. 461), could be witnessed on SMPs across the globe (Brandwatch & Ditch the Label, 2021; Landesanstalt für Medien NRW, 2021). A collaborative research between Brandwatch and Ditch the Label (2021) on the evolution of online hate speech in the United States and the United Kingdom showed a 38% increase of such incidents between mid-2020 and mid-2021. Similarly, in Germany an increase in online hate speech between 2020 and 2021 became apparent. As such, the number of respondents who have ever encountered cyberhate in their life rose from 73% to 76% with the percentage of people indicating to be exposed to such incidents "(very) frequently" having increased from 34% to 39% (Landesanstalt für Medien NRW, 2020, 2021). While the most common types of hate speech occurring on SMPs are racist speech directed at ethnic- or religious minorities as well as gendered cyberhate targeting women and the LGBTQIA+ community in the form of, among others, insults, violent- or sexualized rhetoric, intimidation, death- or rape threats, shaming and discreditation, the dramatic increase of cyberhate is not limited to these types (Castaño-Pulgarín, Suárez-Betancur, Tilano Vega, & Herrera López, 2021; Guo & Johnson, 2020; Jane, 2017). In this regard, other examples include jihadist speech as well as political- and racial cyberhate around the COVID-19 pandemic (De Smedt, De Pauw, & Van Ostaeyen, 2018; Klausen, Tschaen Barbieri, Reichlin-Melnick, & Zelin, 2012; Tahmasbi et al., 2021; Uyheng & Carley, 2021).

The rising prevalence of the various forms of hate speech on SMPs constitutes a serious contemporary problem with possible offline consequences occurring on both the individual and the societal level (Leonhard, Rueß, Obermaier, & Reinemann, 2018). Starting with the individual level, research suggests a high prevalence of hate speech to inflict extensive psychological damage on those targeted by such speech and to increase the risk of them experiencing physical harm (Nemes, 2002; Ring, 2013; Tsesis, 2002). Psychologically

speaking, repeated exposure to hateful content may result in self-hatred and internalization of the promoted messages, meaning in members of the derogated or defamed group starting to believe the accusations made (Ring, 2013). This is especially true when hate speech occurs based on immutable characteristics, meaning traits that are fixed and thus impossible to change as for instance ethnicity (Delgado, as cited in Ring, 2013, p. 30). Furthermore, exposure to such content typically leads to distress, fear and humiliation in the victims concerned (Nemes, 2002). Cyberhate may also increase the threat of physical harm towards its targets. As such, regular exposure to hateful content targeting certain individuals and social groups such as ethnic minorities may promote racist beliefs or even create a climate in which hostility towards the victims appears somewhat tolerated. This, on the other hand, increases the probability of offline, physical attacks such as hate crimes (Nemes, 2002; Ring, 2013; Tsesis, 2002).

Besides posing harm to the targets' human dignity, meaning to their essence and identity, including their physical and psychological integrity, a high prevalence of hate speech on SMPs may result in harmful offline consequences on the societal level by not only interfering with the platforms' ability to foster each individual's active participation in democracy but also by exerting negative effects on societal peace. The ability of SMPs to foster active participation in democracy is based on their ability to provide each individual with the opportunity and freedom to participate in ongoing public debates and dialogue, the exchange of different, often competing opinions and claims which, in turn, enables the inclusion of diverse input into political decision making (Citron & Norton, 2011; Gimmler, 2001). While hate speech does not objectively remove this opportunity and freedom, the fear, humiliation, distress, internalization and self-hatred they experience as a result of exposure to hateful content may have a silencing effect on those targeted by hate speech, meaning that it decreases their likelihood to make use of their freedom and the provided opportunities (Citron & Norton, 2011; Ring, 2013). This, on the other hand, closes or at the very least reduces existing discourse about discrimination such as workplace discrimination based on for instance gender or ethnicity, thus limiting the inclusion of these issues in political decision making, ultimately interfering with the victims' active participation in democracy (Citron & Norton, 2011; Nemes, 2002; Ring, 2013). Continuing with the negative effects of cyberhate on societal peace, "online media can serve as a catalyst in complex radicalization or polarization processes" (Rieger et al., 2018, p. 462), suggesting that a high prevalence of for instance jihadist hate speech on SMPs may contribute to the rise of major contemporary societal problems such as extremism and terrorism (Sageman, 2004). Another way in which

online hate may threaten societal peace is through the generation and amplification of social tensions during critical events such as for instance the COVID-19 pandemic (Edwards et al., 2021).

Given the severity of the problem in terms of harm posed to the victims' human dignity and, due to adversely affecting societal peace and the platforms' ability to foster each individual's active participation in democracy, also for society as a whole, the increasingly widespread recognition of the need for combatting the rising prevalence of hate speech on SMPs by means of governance is not surprising (Wilson & Land, 2021). Considering that the governance of hate speech on SMPs possesses two major problems, addressing this issue, however, is more arduous than it may initially seem. Difficulties inherent to the governance of hate speech include challenges surrounding the design and selection of appropriate governance processes as well as questions regarding which actors should be responsible for and involved in this task (Schwoon, Schembera, & Scherer, 2021). The main difficulty inherent to the selection and design of appropriate governance processes on SMPs lies in the fulfilment and balance of two potentially contradictory moral values crucial in democratic societies, namely 1) the protection of human dignity and society as a whole and 2) freedom of speech (Alkiviadou, 2019; Schwoon et al., 2021; Webb et al., 2015). This means that for a governance process to be regarded as suitable, it is not sufficient for it to only be effective in limiting the prevalence of hateful content on SMPs and thus in protecting targets' psychological and physical integrity as well as these victims' active participation in democracy and societal peace. Instead, any effort to protect the human dignity of those targeted by hate speech and society as a whole needs to be carefully weighed against the important moral value of freedom of expression (Gagliardone, Gal, Alves, & Martinez, 2015; Nemes, 2002; Wilson & Land, 2021). In sum, the core challenge inherent to selecting and designing appropriate governance processes for addressing the rising prevalence of hate speech on SMPs thus is that any governance process to be employed needs to be effective in reducing the prevalence of cyberhate, thus protecting human dignity and society while, at the same time should not be so restrictive as to lightly undermine users' opportunities and freedom to express their opinions in public debates and dialogues (Gagliardone et al., 2015; Nemes, 2002; Wilson & Land, 2021).

Further intensifying the difficulties surrounding the governance of hate speech on SMPs are uncertainties regarding the choice of actors involved in and responsible for this task. In this regard, the increasing digitalization of societies, accompanied by the global expansion of SMPs, limited the capability of state actors to sufficiently govern the

communications occurring on these networks on their own, including hateful content (Banks, 2010; Schwoon et al., 2021). More specifically, the increasingly global scale on which SMPs operate and the thus often global reach of hateful content represents major problems for unilateral state regulation as well as multilateral efforts, creating a need for the engagement of new actors in hate speech governance (Banks, 2010). Unilateral attempts at mounting legislative efforts against this type of content often are impeded by limited jurisdictional reach of national state actors and the conflict that, due to a strong divergence of national legislative frameworks, occurs when states attempt to enforce their laws extraterritorially into other jurisdictions (Banks, 2010). Efforts to create a multilateral governmental regulatory system are undermined by the existence of strong differences regarding the evaluation of the relative importance of the aforementioned moral values, namely the protection of human dignity and society and the ideal of freedom of speech among nation states (Banks, 2010). Thus, while its detrimental individual and societal consequences undoubtedly emphasized the need for combatting the rising prevalence of hate speech on SMPs by means of governance, the aforementioned two main challenges inherent to the governance of hate speech on SMPs make addressing this issue more problematic than it may initially appear. In this regard, it seems governance efforts must address both, 1) challenges pertaining to the fulfilment and balance of the two contradictory moral values in the design and selection of governance processes and 2) the decreasing capability of state actors to execute these tasks.

Faced with the challenge of the decreasing capability of state actors to govern hate speech on SMPs, previous efforts to address the contemporary problem of a rising prevalence of cyberhate showed a trend towards self-governance by the SMP on which the hate speech incident in question occurs (Puppis, 2010; Schwoon et al., 2021). Within self-governance, a commonly utilized process is content moderation, typically describing the formulation of guidelines to which all platform communication must adhere and the removal of content which violates these guidelines, sometimes accompanied by (temporary) removals of the respective user accounts (Grygiel & Brown, 2019; Webb et al., 2015). Detections of guideline violations, in the present case meaning detections of hate speech, mostly rely on a combination of user reporting, algorithmic tools and human review (Grygiel & Brown, 2019). Despite initially appearing promising, SMPs' attempts for self-governance in the form of content moderation increasingly come under scrutiny. In this regard, removing content posted by given users or, in the case of (temporary) account removals, even preventing users from posting content restricts these individuals' freedom to express their opinions in public debates and dialogues, a restriction which becomes increasingly problematic when considering that

the decision of what is deemed hateful and consequently which and whose content and accounts are removed are made by private companies that are not only driven by the interest of preventing harm but also by for instance profit-considerations (Grygiel & Brown, 2019; Langvardt, 2018). Besides the aforementioned problems in terms of freedom of speech, content moderation is also far from optimal when it comes to the protection of human dignity and society. Not only do many incidents of hate speech go unnoticed, for instance in the case of the employment of algorithmic tools due to an avoidance of certain "trigger" words and thus remain online, but also sometimes experiencing content- or (temporary) account removal even enhances offenders' hateful behaviour on either the same or other SMP(s) (Kiesler, Kraut, Resnick, & Kittur, 2012; Laaksonen, Haapoja, Kinnunen, Nelimarkka, & Pöyhtäri, 2020; Siegel, 2019).

Resulting from these shortcomings, focus has been put on a wider array of governance processes and actors. In this regard, the importance of civic engagement in the governance of hate speech on SMPs increasingly became acknowledged. More specifically, this means that there has been growing interest in counter speech among users as a promising new attempt to govern hate speech on SMPs (Benesch, Ruths, Dillon, Saleem, & Wright, 2016; Procter et al., 2019; Siegel, 2019). In this regard, counter speech, defined as "a common crowd-sourced response to (…) hateful content" (Bartlett & Krasodomski-Jones, 2015, p. 5), is believed to be a promising opportunity to tackle the contemporary problem of cyberhate and preferable to previous governance attempts for two reasons. First, unlike previous attempts, user self-governance in the form of counter speech does not build on removal and censorship of communications, thus not imposing restrictions on freedom of expression (Gagliardone et al., 2015). Second, user-initiated counter speech has been suggested to potentially be more effective in reducing the prevalence of hate speech on SMPs than the removal of such speech and thus in protecting human dignity and society (Benesch et al., 2016; Strossen, 2018). In this regard, building on influencing offenders, user self-governance in the form of counter speech can serve to reduce the prevalence of hateful content on SMPs, thus protecting the human dignity of victims and society from the harm such speech may cause, not only in a reactive manner by for instance causing offenders to delete their hateful posts but also in a preventive manner by, at least temporarily, discouraging future transgressions (Sonntag, 2019; Wright, Ruths, Dillon, Saleem, & Benesch, 2017).

However, while the benefits of user-initiated counter speech as compared to previous governance attempts are increasingly acknowledged, it is also clear that its full potential as an appropriate and effective long-needed solution for combatting the problem of an increasing

prevalence of hate speech on SMPs is not yet fully unlocked. More specifically, this means that, while the effectiveness of user self-governance in the form of counter speech for influencing offenders and thus for reducing the prevalence of hate speech and protecting human dignity and society increases with a larger number of counter speakers, research indicates that the number of users who are willing to engage in counter speech oftentimes is limited (Buerger & Wright, 2019; Jubany & Roiha, 2015). Combined, these findings clearly emphasize that, in order to fully exploit the potential of user-initiated counter speech as a promising solution for tackling the contemporary problem of cyberhate, strategies that are effective in increasing counter speech intentions among a large proportion of social media users are needed. Yet, despite this urgent need, research in this area is highly scarce (Leonhard et al., 2018).

In an attempt to contribute to this important, yet under-studied research area, and thus to enhancing the potential of user-initiated counter speech as a promising solution for combatting the rising prevalence of hate speech on SMPs, the present study focussed on a short perspective-taking intervention, namely on providing social media users with a short encouragement to adopt the perspective of hate speech targets, as a strategy for increasing their intentions to counterargue against hateful content. This focus on a brief perspective-taking encouragement was chosen for two main reasons:

First, when considering one of the few studies into the determinants of social media users' counter speech intentions in combination with sociopsychological and neuroscientific theories and research findings, providing social media users with a short perspective-taking encouragement can reasonably be assumed to be an effective strategy for enhancing their counter speech intentions through a presumed invocation of a causal chain of reduced dehumanization of hate speech targets, increased empathy with these victims and facilitated threat perceptions towards hate speech incidents. In this regard, attempting to adopt the perspective of other individuals has been demonstrated to increase feelings of social connectedness towards these people as well as inclusion of these people into the self (Davis, Conklin, Smith, & Luce, 1996; Hutcherson, Seppala, & Gross, 2008; Todd, Bodenhausen, Richeson, & Galinsky, 2011), factors which have been demonstrated to significantly reduce the extent to which individuals are dehumanized (Haslam & Stratemeyer, 2016). This is important because research has not only suggested that dehumanization of hate speech victims, meaning the denial of human attributes to these people, occurs among observers (Fasoli et al., 2016; Haslam, 2006; Murrow & Murrow, 2015), but also that dehumanization reduces observers' empathy towards the individuals in question (Čehajić, Brown, & Gonzáles,

2009; Murrow & Murrow, 2015; Schein & Gray, 2018). Empathy with victims, defined as the capability to be affected by their emotional experience (Čehajić et al., 2009), on the other hand, is a crucial factor influencing observers' threat perceptions, meaning the extent to which observers perceive a particular situation to pose (threat of) harm to the victims impacted by the incident (Krueger et al., 2013; Leonhard et al., 2018; Schein & Gray, 2018). Given these findings and the fact that social media users' threat perceptions towards hate speech incidents are an important determinant of their intentions to counterargue against such speech (Leonhard et al., 2018), an effectiveness of the discussed perspective-taking intervention in terms of enhancing social media users' intentions to counterargue against hateful content on SMPs on grounds of the described causal chain can reasonably be assumed.

Second, if indeed proven effective in increasing social media users' intentions to engage in counter speech, the aforementioned short perspective-taking intervention further excels through its additional practical advantages. As a strategy that neither requires much time nor materials when utilized in practice, it is relatively affordable, can be designed and implemented with relative ease and can thus be distributed by various actors through various channels. Both, high affordability and wide reach are important criteria in terms of the practical utility of intervention strategies (West, Michie, Atkins, Chadwick, Lorencatto, 2019) and allow for a strategy such as the perspective-taking intervention to serve effective in increasing counter speech intentions among a large proportion of social media users.

In order to test these assumptions, an endeavour of clear societal and scientific importance, the present study investigated the following research question:

Is the provision of a short perspective-taking encouragement an effective strategy to increase social media users' counter speech intentions toward hate speech and is potential effectiveness fully attributable to a causal chain of reduced dehumanization, facilitated empathy and increased threat perceptions?

The investigation occurred by means of a quantitative online experimental design. Participants, all of them social media users, were randomly assigned to one of two conditions, meaning that they either received a short perspective-taking instruction (experimental condition) or a control task (control condition). Afterwards, participants in both conditions were presented with the same fictional hate speech post extracted from a previous study and the same questionnaires assessing threat perceptions, dehumanization, empathy and counter speech intentions. The obtained data was analysed by means of a Conditional Process Model with three-way serial mediation performed using the PROCESS macro for IBM SPSS (Hayes, 2018). Utilizing a serial mediation model did not only allow for assessing the effectiveness of

the perspective-taking encouragement in terms of enhancing social media users' counter speech intentions but also for gaining insights into the underlying mechanism by which a potential effectiveness occurs. As such, the results obtained by means of a serial mediation model did not only indicate whether there was an effect of condition on counter speech intentions in favour of the experimental group, that is whether the proposed strategy is effective in enhancing these intentions, but also whether a potential effect was fully mediated by dehumanization, empathy and threat perceptions in serial, meaning whether a potential effectiveness can indeed fully be attributed to the assumed causal chain. The latter is important because having insights regarding why a certain strategy is effective or ineffective offers valuable information for future strategy design attempts by other researchers.

Discussing previous literature, the present chapter explained that research into effective strategies to enhance social media users' counter speech intentions when witnessing hate speech incidents on SMPs is scarce, but that such strategies are urgently needed to fight the prevalence of cyberhate and its detrimental consequences for victims' human dignity as well as, due to adversely affecting societal peace and SMPs ability to foster active participation in democracy, for society as a whole. The remaining part of the present thesis is structured as follows. In the second chapter, the theoretical claims and empirical research findings leading to the conceptualization of the provision of a short perspective-taking encouragement as such an effective strategy to increase social media users' counter speech intentions on SMPs as well as of a causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions as the underlying mechanism for this effectiveness are provided. More specifically, this chapter explains the role of threat perceptions as a determinant for social media users' counter speech intentions as well as the role of empathy in threat perceptions. Furthermore, the role of dehumanization as an empathy-constraining factor and the impact of perspective-taking attempts on dehumanization is elaborated upon. Alongside the written format, a graphical representation of the hypotheses developed based on an integration of the aforementioned literature streams – the so-called conceptual model – is included. During chapter three, the methodological- and analytical choices made to assess the formed hypotheses, thus arriving at an answer to the research question, are elaborated upon in depth. In chapter four, the results of the quantitative analyses are presented and discussed in terms of their implications for the research hypotheses. In chapter five, an answer to the research question is provided and contributions, implications, recommendations, limitations and future research directions are discussed. Chapter six contains the conclusion and in chapter seven a reflection on the thesis project is provided.

Theoretical Background

Threat Perceptions as a Determinant of Social Media Users' Counter Speech Intentions .

Since it has first been hypothesized in the 20th century, the role of observers' threat perceptions in response to a given incident as a determinant of their willingness to intervene in the situation on behalf of the victims has been repeatedly investigated and established in multiple domains (Fischer et al., 2011; Latané & Darley, 1970; Obermaier, Fawzi & Koch, 2016). Most recently, these inquiries also included the empirical issue in question, namely social media users' counter speech intentions when witnessing hate speech incidents on SMPs. More specifically, Leonhard, Rueß, Obermaier and Reinemann (2018) demonstrated users' intentions to counterargue against hateful content on behalf of the targets to depend on their threat perceptions in response to the hate speech incidents. This means that, on average, social media users' counter speech intentions were higher the higher their threat perceptions, meaning the higher the extent to which they perceived the hate speech incident in question as posing threat of harm or actual harm to those victims targeted by it (Leonhard et al., 2018). An explanation for this effect of threat perceptions on social media users' counter speech intentions comes from Latané and Darley's (1970) model on the decision-making process for helping behaviour. According to this model, individuals' intentions to intervene in a critical situation on behalf of the victims affected by it depend on the extent to which four different stages are successfully completed. In the first two stages, individuals need to notice the critical situation in question and interpret it as an emergency, that is as (potentially) harmful to those affected by it. In the third stage, individuals need to consider themselves personally responsible to intervene in the situation in question, meaning that they need to feel as if it is their duty to help those victims. In the fourth stage, individuals need to reflect on and know how to help. Depending on the extent to which the aforementioned stages are successfully completed, individuals are expected to end up with higher or lower intervention intentions. Social media users' threat perceptions in response to a particular hate speech incident, meaning the extent to which they perceive the incident to be (potentially) harmful to its targets, thus are a key factor in determining their counter speech intentions because they directly reflect the second stage of the decision-making process involved in helping behaviour (Leonhard et al., 2018).

The discussed relationship between threat perceptions and counter speech intentions has important implications for the design of strategies that are effective in increasing social media users' intentions to engage in counter speech when witnessing hate speech incidents on SMPs. In this regard, given the key role of threat perceptions as a determinant of users'

counter speech intentions, any strategy that beneficially affects social media users' threat perceptions when witnessing cyberhate on SMPs should, at the same time, also be an effective strategy to increase their intentions to counterargue against such content (Leonhard et al., 2018). Hence, theories and research on the origin of threat perceptions are likely to offer a fruitful starting point for the development of an effective strategy to heighten social media users' counter speech intentions.

The Theory of Dyadic Morality and the Role of Empathy in Threat Perceptions

Drawing on the theory of dyadic morality (TDM) (Schein & Gray, 2018), observers' threat perceptions in response to a given incident are related to the cognitive system. More specifically, according to the TDM, the extent to which a given incident is perceived as (potentially) harmful depends on the extent to which it matches a universally applicable cognitive template of harm. An important component in this cognitive template of harm is the degree of suffering perceived to occur as a result of the incident (Schein & Gray, 2018). More specifically, this means that, the more observers witnessing a certain incident perceive the incident to cause suffering to the persons affected by it, the more the event should match their cognitive template of harm and thus the higher their threat perceptions in response to the incident should be (Schein & Gray, 2018).

While perceptions of the degree of suffering occurring as a result of a given incident are deeply rooted in innate processes of the human mind, other factors also seem to play a role. One of these factors is empathy toward the persons affected by the event (Schein & Gray, 2018). According to the TDM, in order to perceive an event as causing suffering to those impacted by it, observers need to empathize with these victims, that is, they need to be capable to be affected by the victims' emotional experience (Schein & Gray, 2018). The idea is that the more observers empathize with the persons who are affected by a certain event, the more they perceive it as causing suffering to these victims and, as previously discussed, the more the incident matches their cognitive template of harm. A higher match, on the other hand, gives rise to observers experiencing higher threat perceptions in response to the incident (Schein & Gray, 2018).

A recent neuroscientific study offers additional evidence for the theorized effect of observers' empathy with the persons affected by an event on their threat perceptions in response to the incident. In this regard, Krueger et al. (2013) investigated the impact of the exogenous administration of oxytocin, a neuropeptide said to facilitate empathy, on participants' perceptions of the (potential) harmfulness of an offense. Participants who received the neuropeptide consistently perceived the offense as significantly more harmful to

the victims affected by it than participants in the placebo condition, thus further supporting the role of empathy for threat perceptions theorized by the TDM (Krueger et al., 2013).

The Role of Dehumanization as an Empathy-Constraining Factor

While, based on the previous two sections, one could be inclined to conceptualize social media users' empathy with hate speech targets as an important factor for strategies to target in order to increase these users' threat perceptions in response to cyberhate and thus their counter speech intentions, the matter is more complicated than that. In this regard, previous research suggests that empathy in itself can be heavily constrained by dehumanization: The higher observers' dehumanization of the persons that are affected by a given incident, the lower observers' empathy with the victims in question (Čehajić et al., 2009; Haslam & Stratemeyer, 2016; Schein & Gray, 2018). Dehumanization commonly refers to the denial of human attributes to another person (Haslam, 2006). It involves mentally stripping the other person of the characteristics that distinguish humans from other animals or automata, ultimately conceptualizing them as animate beings or inanimate objects (Haslam, 2006). Examples include but are not limited to, completely or partially denying depth or the ability to reason to another person (Haslam, 2006).

A potential explanation for the demonstrated negative effect of dehumanization on empathy originates from neuroscience. In this regard, a recent paper by Murrow and Murrow (2015) proposes that more severely mentally stripping other persons of their human attributes results in lower tendencies to empathize with these people because neural empathetic mechanisms respond best to creatures judged to belong to the same species as the self. Consequently, the more observers dehumanize another person, the worse the response of their neural empathetic mechanisms to this person is assumed to be, ultimately explaining the negative effect of observers' dehumanization on their experienced empathy levels toward the individual in question (Murrow & Murrow, 2015).

Considering this negative effect of dehumanization on empathy as well as the established relationships between empathy, threat perceptions and counter speech in light of the suggestion that the dehumanization of hate speech victims occurs among observers (Fasoli et al., 2016; Murrow & Murrow, 2015), important implications for the design of effective strategies to increase social media users' intentions to counterargue against hate speech incidents on SMPs emerge. More specifically, based on the aforementioned findings, it can be assumed that any strategy that effectively reduces social media users' dehumanization of hate speech targets may be a promising avenue to increase their empathy towards these victims

and thus, by enhancing their threat perceptions, facilitate their counter speech intentions when being exposed to hate speech incidents on SMPs.

The Provision of a Short Perspective-Taking Encouragement as a Potential Strategy

Linking previous research on perspective-taking with research focusing on remedies of dehumanization, encouraging social media users to adopt the perspective of hate speech targets may be such a strategy. Starting with the former aspect, recent research hypothesized that adopting another (group of) person(s)'s perspective may enhance the extent to which one feels socially connected to these people (Hutcherson et al., 2008; Todd et al., 2011). In addition, perspective-taking has also been positively linked to cognitive self-other overlap (Davis et al., 1996). More specifically, Davis, Conklin, Smith and Luce (1996) suggested individuals who received an instruction to adopt another person's perspective to experience a greater proportion of overlap between cognitive representations of the self and of the person whose perspective they were encouraged to adopt, meaning that they perceived the other as more similar to themselves. The provided explanation for this is that actively attempting to adopt another person's perspective causes individuals to ask themselves what they would experience if they were in this person's shoes. Doing so, on the other hand, likely strengthens internal associations between self-relevant information and information about the other person, ultimately resulting in a situation in which the other individual, to a greater extent, is perceived to possess similar traits as oneself (Davis et al., 1996).

Continuing with the latter aspect, research suggests relatively long-lasting remedies of dehumanization to include not only feelings of social connectedness but also enhanced perceptions of similarity (Haslam & Stratemeyer, 2016). More specifically, according to Haslam and Stratemeyer (2016), perceiving another person as more similar to themselves and feeling more socially connected to this person leads to individuals displaying a sustained reduction in dehumanization of the person in question, meaning to them denying human attributes to this person to a lower extent. Given the aforementioned remedies of dehumanization, the discussed beneficial effects of perspective-taking on those remedies and the suggestion that even brief perspective-taking instructions are capable of inducing compliance in terms of perspective-taking attempts and thus of producing these beneficial effects (Davis et al., 1996; Todd et al., 2011), it can be assumed that providing social media users with a short encouragement to adopt the perspective of hate speech targets is effective in reducing their dehumanization of these victims. Furthermore, when considering these findings in light of the relationships discussed in the previous sections of this chapter, it can be assumed that providing social media users with a brief perspective-taking encouragement

will, by means of reducing their dehumanization of hate speech targets and thus through increasing their empathy with these victims and facilitating their threat perceptions, be effective in enhancing their intentions to counterargue against hateful content on SMPs.

Conceptual Model

To sum up, when discussed in connection to each other, the theoretical claims and empirical research findings discussed in the current chapter clearly suggest an effectiveness of the provision of a short perspective-taking encouragement as an intervention strategy to increase social media users' intentions to counterargue against hate speech incidents on SMPs. More specifically, based on linking the discussed theoretical and empirical findings originating from different literature streams, it can be assumed that providing social media users with a brief instruction to adopt the perspective of individuals and social groups that are targeted by hate speech incidents will be effective in increasing their counter speech intentions in response to such incidents. Also, based on literature, it can be assumed that this is the case because attempts to take the perspective of hate speech targets negatively affect social media users' dehumanization of the targets which in turn beneficially affects their empathy towards these victims, ultimately enhancing threat perceptions in response to hate speech incidents. Hence, well-grounded in previous literature, the following answers to the research question of the present study were expected:

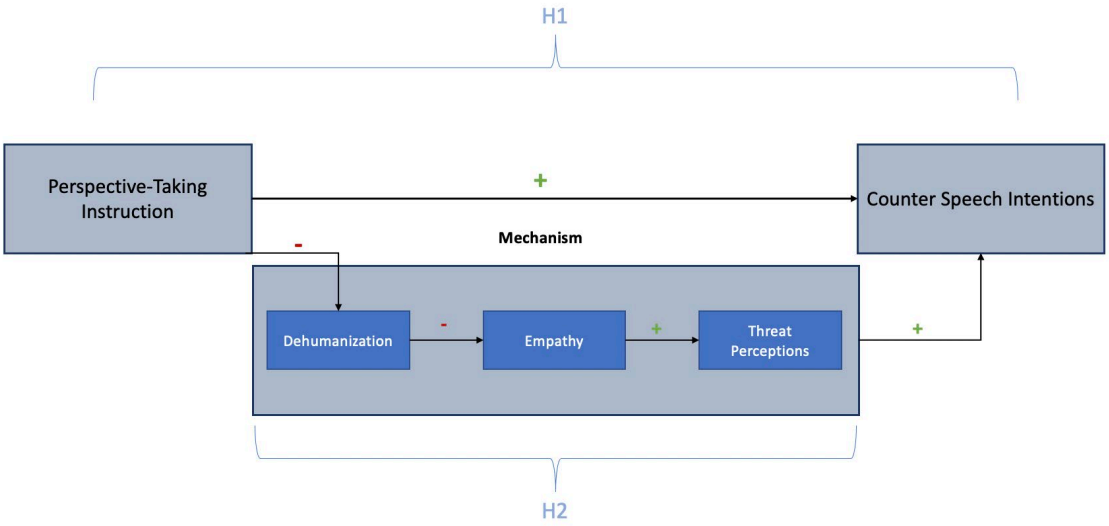
H1: Providing social media users with a short encouragement to adopt the perspective of hate speech targets is an effective strategy to increase their counter speech intentions towards hateful content on SMPs.

H2: This effectiveness of a short perspective-taking encouragement in terms of increasing social media users' counter speech intentions is fully attributable to a causal chain of reduced dehumanization, increased empathy and enhanced threat perceptions.

For a graphical representation of the formed hypotheses, meaning the conceptual model of the present study, please consult Figure 1. For detailed information on how the hypotheses were assessed, please refer to the following chapter of the thesis at hand.

Figure 1

Conceptual Model



Method

Research Design

The research design selected for the present study was a quantitative experimental study conducted online via Qualtrics. Participants, all of them social media users, were randomly assigned to one of two conditions. Participating social media users assigned to the experimental group (*Condition: Perspective-Taking Instruction (Yes)*) received an informational message about the occurrence of hate speech on SMPs and the experimental manipulation, meaning a short perspective-taking encouragement. Participating social media users assigned to the control group (*Condition: Perspective-Taking Instruction (No)*) received the same informational message but instead of the experimental manipulation, they were presented with a control task which matched the perspective-taking encouragement in scope. Afterwards, following the design employed by Leonhard et al. (2018), all participants received the same fictional hate speech post and questionnaires investigating 1) the extent to which they perceive the hate speech post as (potentially) harmful to those victims targeted by it (*Threat Perceptions*) and 2) their intentions to counterargue against the post (*Counter Speech Intentions*). Upon completion, all participants were presented with questionnaires assessing 1) the extent to which they deny human attributes to the targets of the hate speech post (*Dehumanization*) and 2) the degree to which they empathize with these victims (*Empathy*). Last but not least, a manipulation check assessing self-reported perspective-taking attempts was presented. In order to test the formed hypotheses, the collected data was analysed by means of a Conditional Process Model with three-way serial mediation performed using the PROCESS macro for IBM SPSS (Hayes, 2018). More specifically, this analysis allowed to assess the effectiveness of the perspective-taking intervention as well as the mechanism underlying any potential effectiveness. In this regard, it allowed to investigate whether, in line with a hypothesized effectiveness, participating social media users in the experimental group displayed significantly higher counter speech intentions than participants in the control group as well as whether potential significant differences could fully be explained by the proposed causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions. In order to conduct the manipulation check, a Mann-Whitney U Test was selected. Initially, an independent samples t-test was planned to be utilized to conduct the manipulation check but, as a result of the assumption testing, the initial plan had to be adapted. For further details on participants included, the research materials utilized, the exact experimental procedure employed, the methods of data analysis chosen and research ethics, please refer to the respective sections.

Such a quantitative experimental design was chosen as it was the most suitable option to examine the present research question and the associated hypothesized model for multiple reasons. Utilizing a quantitative experimental design minimizes the risk of biases in results and conclusion as a result of confounding variables (Kumar, 2011). Also, it enables the conduction of serial mediation analyses. The ability to conduct a serial mediation analysis was vital for investigating the present research question and associated hypotheses. As such, a serial mediation model is an analysis that is not only capable of providing information about whether there is an effect of condition on counter speech intentions in favour of the experimental group, that is about whether the proposed strategy is indeed effective, but also about whether a potential effect is fully mediated by dehumanization, empathy and threat perceptions in serial. The latter is important because it enables an accurate understanding of whether a potential effectiveness is indeed fully attributable to the assumed causal chain (Hayes, 2018). Such knowledge, on the other hand, can be of great value for future strategy design attempts by other researchers. If it is for instance found that the examined perspective-taking intervention is effective for enhancing participating social media users' counter speech intentions and that this effectiveness is fully attributable to a causal chain of reduced dehumanization of hate speech targets, increased empathy with these victims and facilitated threat perceptions in response to hate speech incidents, future design attempts could focus on finding other strategies targeting dehumanization. Also, if it is found that the examined intervention strategy is ineffective, actively having investigated the presence of the assumed causal chain could provide valuable information about the underlying reasons. As such, it would for instance signal whether the examined strategy is ineffective because it, contrary to the formulated hypotheses, does not manage to reduce dehumanization. Or whether instead it is ineffective because, against assumptions, reduced dehumanization does not lead to increased empathy or that increased empathy does not result in enhanced threat perceptions. Yet another possible reason for ineffectiveness that serial mediation would inform about is if, contrary to Leonhard et al.'s (2018) findings, enhanced threat perceptions do not facilitate participating social media users' counter speech intentions. Besides enabling the conduction of a serial mediation analysis and reducing the risk of biases, the chosen design is useful because quantitative research enables data gathering from a large number of social media users, therefore aiding the generalizability of the findings of the study (Hair, Black, Babin, & Anderson, 2019).

Participants

A total of 132 study participants were recruited via social media platforms including WhatsApp, Facebook, Twitter, Vinted, Reddit and Instagram. Any individual who was of legal age, fluent in the English language and had at least one social media account was eligible to participate in the study at hand. Participation was not reimbursed. Of the initial 132 participants, 30 participants did not agree to the consent form or did not complete the study. There were no extreme outliers. The final sample thus consisted of 102 participants. For information on participant distribution across conditions and demographics, please reference Table 1. In addition, Table 1 provides information on participants' social media behaviour within and across conditions as well as about the percentage of participants recruited via the different SMPs within and across conditions.

Table 1

Participant Demographics and Social Media Behaviour Within and Across Conditions

Condition	Experimental: Perspective-Taking Encouragement (Yes) (n = 48)		Control: Perspective-Taking Encouragement (No) (n =54)		Full Sample (n = 102)	
	n	*%*	*n*	*%*	*n*	*%*
Gender						
Male	22	45.8	23	42.6	45	44.1
Female	19	39.6	24	44.4	43	42.2
Non-Binary	2	4.2	3	5.6	5	4.9
NI	5	10.4	4	7.4	9	8.8
Recruitment Platform						
Reddit	18	37.5	23	42.6	41	40.2
WhatsApp	9	18.7	13	24.1	22	21.5
Facebook	7	14.6	10	18.5	17	16.7
Twitter	8	16.7	3	5.6	11	10.8
Vinted	3	6.3	2	3.7	5	4.9

Instagram	3	6.3	1	1.9	4	3.9
NI	0	0.0	2	3.8	2	2.0
Average Time Spent on Social Media Platforms/ Day						
Less than 1 hour	5	10.4	17	31.5	22	21.6
1-2 hours	23	47.9	20	37.0	43	42.2
2-3 hours	12	25.0	11	20.4	23	22.5
3-4 hours	5	10.4	2	3.7	7	6.9
More than 4 hours	3	6.3	4	7.4	7	6.9

Note. $N = 102$. Participants were on average 31.25 years old ($SD = 12.90$) and, according to an ANOVA, participant age did not differ by condition ($F(1, 101) = .008$, $p = .930$).

Material

All research materials were in English in order to allow for the inclusion of different nationalities in the present study. For an overview, please reference Appendix A. To enhance the reach of the research and consequently the number of participants, the study at hand was conducted online via Qualtrics. For the sake of comprehension, the following section will discuss the research materials separated according to the different components and variables utilized in the study at hand.

Study Conditions

Considerable attention was paid to keep the materials provided to participating social media users in the two different study conditions identical in every regard apart from the experimental manipulation itself. More specifically, this means that the design of the provided materials as well as the content and the wording of the informational message about the occurrence of hate speech on SMPs was identical in the experimental- and the control condition. As such, the only difference between the two conditions lay in the experimental manipulation, meaning in the type of instruction (perspective-taking encouragement vs. control task) participants received. This was done in order to avoid bias in the results and conclusions, meaning in order to ensure that any potential significant differences between the two conditions can indeed be attributed to the effectiveness of a short perspective-taking intervention instead of to confounding factors as for instance differences in wording or design (Coon & Mitterer, 2014; Kumar, 2011). The reason for providing participants with an

informational message about the occurrence of hate speech on SMPs was to provide an adequate context for the subsequent instructions and tasks. More specifically, the content of the message consisted of information about the rising prevalence of hate speech on SMPs, the commonly provided definition of hate speech, the fact that many different individuals and social groups can be affected by it as well as a short example of the forms it can take to ensure better understanding. All the aforementioned included information was extracted from scientific articles (c.f. e.g. Brandwatch & Ditch the Label, 2021; Castaño-Pulgarin et al., 2021; Landesanstalt für Medien NRW, 2020, 2021; Rieger et al., 2018). The brief perspective-taking encouragement provided to the experimental condition was created by adapting the perspective-taking instruction utilized by a previous study to the empirical setting in question, meaning to the context of hate speech incidents and hate speech targets on SMPs (Todd et al., 2011). As such the perspective-taking encouragement asked participating social media users to visualize what hate speech targets may be thinking, feeling and experiencing. Although Todd's, Bodenhausen's, Richeson's and Galinsky's (2011) perspective-taking encouragement intervention was utilized in a different context and for another purpose than reducing victim dehumanization and thus increasing empathy with victims, threat perceptions and helping intentions, a manipulation check incorporated by these authors showed it to be effective in encouraging compliance in terms of perspective-taking attempts among different individuals, thus suggesting its utility for the cause at hand. The content of the control-task provided to the control condition, on the other hand, was selected based on thematic fit with the informational message to avoid suspicion among participants. More specifically, the control-task asked participating social media users to visualize all SMPs on which, in their opinion, hate speech may frequently be found. Furthermore, for previously explained purposes of identicality, the control-task was designed to match the perspective-taking encouragement in scope and wording was kept as similar to the perspective-taking instruction as possible (Coon & Mitterer, 2014; Kumar, 2011). In order to ensure message comprehension among all participants, the language utilized in all materials omitted technical terms as much as possible and was kept relatively simple. The design of the materials was developed and implemented in Microsoft Power Point, was fairly plain and kept in the same colour-scheme as the subsequently presented fictional hate speech post.

Fictional Hate Speech Post

Following the design of Leonhard et al. (2018), a fictional hate speech post was utilized in the study at hand. More specifically, one of the fictional posts created by the aforementioned authors was selected. In line with the authors as well as in order to enhance

the external validity of the research at hand, the fictional hate speech post was designed in the format of a social media webpage. In order to prevent legal claims, a new name and design was developed instead of using the name and design of an existing SMP. The design was developed and implemented by means of Microsoft Power Point.

Threat Perceptions

Participating social media users' threat perceptions in response to the hate speech incident in question, meaning the extent to which they perceived the presented hate speech post as (potentially) harmful to those victims affected by it, was assessed by means of a scale developed by Leonhard et al. (2018). The scale is an English translation of the original German items. This translation was provided by the authors themselves in their article. The scale consists of three closed-end questions, each scored on a 7-point Likert scale (1 = completely disagree, 7 = completely agree). Higher total scores indicated higher threat perceptions. A sample item is "The demonstrated social media post is harmful to the affected people".

Counter Speech Intentions

Participating social media users' intentions to engage in counter speech were assessed by the following item: "I would comment against the demonstrated social media post". Responses were scored on a 7-point Likert scale, ranging from 1 (= completely disagree) to 7 (= completely agree) (Leonhard et al., 2018).

Dehumanization

Following previous research, participating social media users' dehumanization of the targets of the hate speech post, meaning the extent to which they denied human attributes to these victims, was assessed by asking them to rate the extent to which different descriptions apply to the hate speech targets (Lammers & Stapel, 2011). In total 16 items were utilized with examples including "lacking depth" "being sensitive" and "being unmannered" (Lammers & Stapel, 2011). Responses were scored on a 9-point Likert scale, ranging from 1 (= not at all) to 9 (= very much). Higher total scores indicated higher dehumanization of the targets of the hate speech post, creating the need to reverse-code seven positively framed items. One example for such an item is "being sensitive". The original scale by Lammers and Stapel (2011) used two subscales, one for measuring the extent to which participants engaged in the denial of human attributes to another person by mentally stripping this person from the characteristics that distinguish humans from other animals and one for assessing the degree to which participants engaged in the denial of human attributes to another person by stripping this person from the characteristics that separate humans from automata. In the present study,

for the purposes of statistical analysis, more specifically in order to be able to perform the proposed serial mediation model within the chosen framework of Conditional Process Analysis, the subscales were collapsed into a single 16-item scale. In the present study, the reliability of this collapsed scale was very high with a Cronbach's Alpha of approximately .97. This, while indicating exceptional reliability, points at some redundancy among items, which might be alleviated in future studies by reducing the number of superfluous items.

Empathy

Utilizing a previously established measurement of empathy, participating social media users' empathy with the targets of the hate speech post was assessed by asking them to indicate the extent to which they experienced six different emotion adjectives to these people (Batson & Ahmad, 2001). Examples included "compassionate" and "moved". Responses were scored on a 6-point Likert scale (1 = not at all, 6 = extremely). Higher total scores indicated higher empathy with the victims of the hate speech post.

Manipulation Check

Following the example of Todd et al. (2011), a manipulation check was incorporated in the study at hand. More specifically, in line with the aforementioned authors, participating social media users were asked to report on the extent to which they actually tried to imagine what people that are targeted by hate speech are feeling, thinking and experiencing, (*Perspective-Taking Attempts*). The exact question presented to participants was "To what extent did you try to imagine what people that are targeted by hate speech might be thinking, feeling and experiencing?". Responses were scored on a 7-point scale (1= not at all, 7 = very much so). This was done in order to assess the capability of the experimental manipulation in terms of inducing perspective-taking attempts, meaning to investigate whether, similar to Todd et al.'s (2011) findings, participants who received the brief perspective-taking encouragement indeed complied with the instruction and attempted to adopt the perspective of hate speech targets to a larger extent than participants who received the control-task. In general, including a manipulation check is advisable for the correct interpretation of obtained results (Haslam & McGarty, 2004). Especially in the study at hand its utility is evident. In this regard, the presumed effectiveness of the perspective taking intervention for reducing social media users' dehumanization of hate speech victims and thus, mediated by empathy and threat perceptions, for increasing their counter speech intentions is based on the premise that participants comply with the instruction and indeed attempt to adopt the perspective of hate speech targets. If, in contrast to Todd et al.'s (2011) findings, this should turn out not to be the case, additional care must be taken in interpreting the results. In this regard, potential non-

significant effects may for instance not necessarily indicate an ineffectiveness of the short perspective-taking encouragement at the level of the population in terms of achieving the aforementioned desired outcomes, but rather a failure in the sample to generate compliance in terms of perspective-taking attempts among participants. Conversely, if the manipulation had failed, significant results would not necessarily indicate the effectiveness of the perspective-taking encouragement, but some other, potentially chance factor.

Experimental Procedure

Following recruitment, participants were presented with the study link which forwarded them to Qualtrics. On Qualtrics, they first of all were presented with an informed consent letter. It was explained that the research aim is to examine social media users' perceptions of and reactions to social media communications. To reduce the risk of confounding factors, information about the experimental manipulation, meaning about the fact that, depending on the study condition they are allocated to, they will receive vs. will not receive the short perspective-taking instruction was not yet provided. In addition, it was explained that the study takes approximately 15 minutes, that there will be no compensation for participation, that data will be gathered anonymously and that participants are free to stop their participation at any given moment without facing adverse consequences. Also, they were informed that due to the anonymous nature of the data collection, participants' responses cannot be excluded from the data set and results of the present study anymore once their participation is completed. Furthermore, participants received the researcher's contact information in case they had any remarks or questions and were asked to carefully consider whether they are willing to participate in the study at hand. Afterwards, a consent form was presented.

In case they provided their consent, participants proceeded to the first page of the experimental online study, where they answered three demographic questions focusing on 1) whether they are of legal age, 2) their exact age in years and 3) their gender. Also, they were asked to provide information regarding the SMP on which they encountered the study link as well as about the average daily time that they spend on SMPs. Afterwards, participants were randomly assigned to one of two conditions, namely to either 1) the experimental condition or 2) the control condition. Participants placed in the experimental condition (*Condition: Perspective-Taking Instruction (Yes)*) received the informational message about the occurrence of hate speech on SMPs and the experimental manipulation, meaning the short perspective-taking encouragement, asking them to visualize what targets of hate speech may be feeling, thinking and experiencing. Participants assigned to the control condition

(*Condition: Perspective-Taking Instruction (No)*) received the same informational message but instead of the experimental manipulation, they were presented with the control task which matched the perspective-taking instruction in required time and effort. Then, all participants were presented with the same fictional hate speech post. Following the fictional hate speech post, all participants received the questionnaire assessing 1) the extent to which they perceive the post as posing threat of harm or actual harm to those targeted by it (*Threat Perceptions*) and 2) their intentions to counterargue against the post (*Counter speech Intentions*). Upon completion, all participants were presented with the questionnaire investigating the extent to which they deny human attributes to the targets of the hate speech post (*Dehumanization*) as well as the degree to which they empathize with these victims (*Empathy*). Last but not least, the manipulation check assessing self-reported perspective-taking attempts was presented. Afterwards, participants in all conditions were fully debriefed about the complete nature of the research. In addition, they were again warned that, due to the anonymous nature of the data collection, excluding their responses from the data set and results of the present study is impossible once they completed the study. As such, they were again actively asked for permission to include their data and informed that, if they decline, their responses will be deleted.

Data Analysis

Data Preparation and Descriptive Analyses

Following the data collection, the obtained data was imported to SPSS, where it was prepared for the subsequent statistical analyses. For an overview of all the steps involved in the data preparation process, please consult Table 2. In addition, descriptive analyses were performed. This was done in order to obtain a first overview of participants' average dehumanization-, empathy-, threat perception-, and counter speech intention scores within and across experimental conditions. In addition, the aim was to obtain an overview of the extent to which participants within and across experimental conditions attempted to engage in perspective-taking throughout the study.

Table 2

Overview of the Steps Involved in the Data Preparation Process

Step	Description
Checking for system missing values	Examination of whether data on all variables is available for all participants included in the sample
Reverse-coding of items	Reverse-coding of seven items of the dehumanization scale (in line with Lammers & Stapel, 2011) • Dehu_5: Having self-control • Dehu_6: Having decency • Dehu_7: Being polite • Dehu_8: Being civilized • Dehu_9: Being rational • Dehu_10: Being mature • Dehu_16: Being sensitive
Calculation of total scores	Only for scales consisting of multiple items.

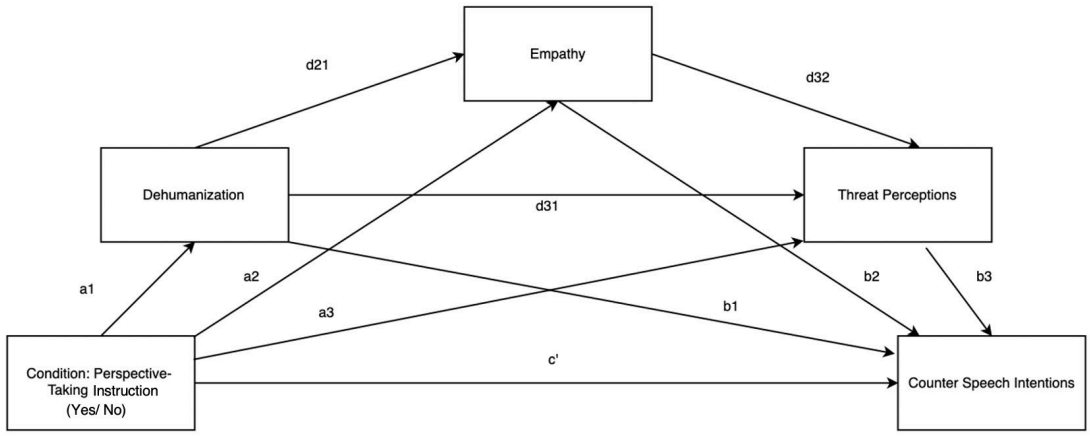
Hypothesis Testing

The hypotheses of the present research were investigated by means of a Conditional Process Model with three-way serial mediation, performed by utilizing the SPSS PROCESS Macro (Hayes, 2018). In the analysis, *Condition: Perspective-Taking Instruction (Yes/ No)*[1] constituted the independent variable, whereas *Counter Speech Intentions* served as the dependent variable. Furthermore, for the analysis, the model number was set to six, meaning to serial mediation with *Dehumanization* serving as the first mediator, *Empathy* constituting the second mediator and *Threat Perceptions* being the third mediator. The analysis in question was conducted with the default setting of 95% confidence intervals and 5000 bootstrap samples. In order to facilitate ease of replication, the random number generator was seeded to a fixed number, namely to 4421. For a graphical representation of all the pathways assessed by the statistical model, please reference Figure 2.

[1]The experimental group (Yes) was coded as "2" in the data-set, whereas the control group (No) was coded as "1".

Figure 2

Graphical representation of the Pathways Assessed by the Three-Way PROCESS-Performed Mediation Model



Of particular interest for the testing of hypothesis H1 is the total effect of *Condition: Perspective-Taking Instructions (Yes/ No)* on *Counter Speech Intentions*. If this total effect was significant with a positive regression coefficient, hypothesis H1 would be accepted. In this regard, a significant total effect with a positive regression coefficient would indicate that counter speech intentions were significantly higher in the experimental condition than in the control condition, thereby indicating that, in line with the first hypothesis, providing social media users with a brief perspective-taking encouragement indeed is an effective strategy for increasing their intentions to counterargue against hateful content. If the total effect, however, was non-significant or had a negative regression coefficient, hypothesis H1 would have to be rejected. In this regard, a non-significant total effect would mean that there are no significant differences in social media users' counter speech intentions according to the condition that they were assigned to, meaning according to whether they received the perspective-taking instruction or the control-task. This would indicate that, in contrast to the first hypothesis, the short perspective-taking intervention was ineffective in increasing social media users' intentions to counterargue against hate speech, depending on the results of the manipulation check, either due to a failure in the sample to induce compliance in terms of perspective-taking attempts among participants or due to an ineffectiveness in general. A significant total effect with a negative regression coefficient would also lead to the rejection of hypothesis H1 because it would signal that, while there were significant differences in counter speech intentions between the two conditions, these differences occurred in favour of the control group instead of the experimental condition.

In addition to the total effect, several other pathways are of interest for the assessment of hypothesis H2. In order to assess this hypothesis it is important to consider the direct effect of *Condition: Perspective-Taking Instruction (Yes/ No)* on *Counter Speech Intentions (c')* as well as all indirect- and individual pathways. If the indirect pathway from *Condition* to *Counter Speech Intentions* via *Dehumanization, Empathy* and *Threat Perceptions* in serial (a1 + d21 + d32 +b3) was significant with a positive regression coefficient, while the direct pathway (c') and all other indirect pathways, meaning

1) *The indirect pathway via dehumanization alone (a1 + b1)*
2) *The indirect pathway via empathy alone (a2 + b2)*
3) *The indirect pathway via threat perceptions alone (a3 +b3)*
4) *The indirect pathway via dehumanization and empathy combined (a1 + d21 +b2)*
5) *The indirect pathway via dehumanization and threat perceptions combined (a1 + d31 + b3)*
6) *The indirect pathway via empathy and threat perceptions (a2 + d32 + b3)*

were non-significant, there indeed would be full serial mediation via the proposed three mediators. This means that such a result would signal that an effectiveness of the proposed perspective-taking intervention, as signalled by a significant total effect with a positive regression coefficient, was indeed fully attributable to a causal chain of dehumanization, empathy and threat perceptions. This alone, however, would not yet be sufficient to accept the second hypothesis. In this regard, for assessing hypothesis H2 it is also necessary to investigate the regression coefficients of the individual effects. Only if

1) *The effect of condition on dehumanization (a1) had a negative regression coefficient*
2) *The effect of dehumanization on empathy (d21) had a negative regression coefficient*
3) *The effect of empathy on threat perceptions (d32) had a positive regression coefficient*
4) *The effect of threat perceptions on counter speech intentions (b3) had a positive regression coefficient*

the causal chain would, as hypothesized, be a causal chain of *reduced* dehumanization, *increased* empathy and *facilitated* threat perceptions. This means that only in this case, any potential beneficial effect of the perspective-taking intervention on counter speech intentions would be due to the fact that the brief perspective-taking encouragement decreased dehumanization which in turn lead to increased empathy which then produced enhanced

threat perceptions. For an overview of the requirements that would need to be met in order to accept the two different hypotheses, please reference Table 3.

Table 3

Overview of the Requirements for Accepting the Hypotheses of the Present Study

Hypothesis	H1: The provision of a short perspective-taking encouragement is an effective strategy for increasing social media users' counter speech intentions	H2: This effectiveness is fully attributable to a causal chain of reduced dehumanization, increased empathy and enhanced threat perceptions
Effect/ Pathway		
Total	Significant; Positive regression coefficient	Significant; Positive regression coefficient
Direct	N.A.	Non-significant
Indirect via all Three Mediators	N.A.	Significant; Positive regression coefficient
Other Indirect	N.A.	Non-significant
Condition on Dehumanization	N.A.	Significant; Negative regression coefficient
Dehumanization on Empathy	N.A.	Significant; Negative regression coefficient
Empathy on Threat Perceptions	N.A.	Significant; Positive regression coefficient
Threat Perceptions on Counter Speech Intentions	N.A.	Significant; Positive regression coefficient

Note: The abbreviation "N.A." stands for non-applicable, meaning that the respective effects are irrelevant to the assessment of the given hypothesis.

Assumption Testing

Before executing the aforementioned statistical analysis, it was examined whether the obtained data meets the four assumptions for PROCESS-performed mediation. An additional fifth assumption, namely statistical independence of errors in estimation, was not explicitly assessed because it could be assumed to be met based on the research design (Field, 2013; Hair et al., 2019). Due to the availability of additional tests and thus higher accuracy of the conclusions, assumption testing for the PROCESS-performed mediation analyses was performed in R rather than in SPSS.

Any serial mediation model that is estimated by PROCESS translates into multiple equations. Applied to the three-way serial mediation model utilized to assess the hypotheses at hand, the statistical model translates into four equations (Hayes, 2018). In the first equation, the first mediator, meaning *Dehumanization*, is estimated from only one predictor, namely from the independent variable, which in this case is *Condition*. In the second equation, the second mediator, in this case *Empathy*, is estimated from two predictors, namely from the independent variable and the first mediator, meaning from *Condition* and *Dehumanization*. In the third equation, the third mediator, in this case *Threat Perceptions*, is predicted from the independent variable, the first- and the second mediator, meaning from *Condition, Dehumanization* and *Empathy.* In the fourth equation, the dependent variable, meaning *Counter Speech Intentions*, is estimated from all four predictors, meaning from *Condition, Dehumanization, Empathy* and *Threat Perceptions* (Hayes, 2018). All four assumptions must be assessed for all equations of the statistical model with the only exception being the first equation (Hair et al., 2019; Hayes, 2018).

For the remaining equations, the first assumption that needs to be assessed is linearity. This assumption states that, in order for PROCESS to enable meaningful interpretations of the results, there cannot be non-linear relationships between any of the predictors and the outcome variables (e.g. for the second equation, between *Condition, Dehumanization* and *Empathy*) (Hair et al., 2019; Hayes, 2018). In addition to examining this assumption graphically via residual plots, the lack-of-fit test was conducted for each predictor of the mediation model to enhance the accuracy of the reached conclusions. For the lack-of-fit test, insignificant p-values ($p \geq .05$) suggest the linearity assumption to be met, whereas significant p-values suggest it to be violated (Fox & Weisberg, 2011). The second assumption is no multicollinearity, meaning that the different predictors should not have high correlations with each other. This assumption was investigated by assessing both, the values of the Variance Inflation Factors (VIFs) and the Tolerance. Whereas, the VIF scores should not be higher than

10 for the assumption to be fulfilled, the Tolerance values should not be smaller than .10 as both, VIFs higher than 10 and Tolerance values smaller than .10 can be indicative of problems with multicollinearity (Hair et al., 2019). The third assumption is normality of the residuals. This assumption refers to the idea that the residuals need to be normally distributed in the sample (Field, 2013). It was assessed by means of the Shapiro-Wilk test. For the Shapiro-Wilk test, insignificant p-values ($p \geq .05$) suggest the normality assumption to be met, whereas significant p-values suggest it to be violated (Field, 2013). Last but not least, the fourth assumption is homoscedasticity of the residuals. It means that the variance of the residuals should be constant over the range of values of the predictors (Hair et al., 2019). This assumption was assessed by inspecting the residual plots, where dots which appear to be relatively randomly and evenly spread indicate homoscedasticity (Field, 2013).

Results of Assumption Testing and Implications for Planned Statistical Analysis

Starting with the first assumption, the residual plots suggest an absence of non-linear relationships between any of the predictors and the outcome variables in any of the equations of the statistical models, thus confirming the assumption of linearity. This interpretation was confirmed by the requested lack-of-fit tests. In this regard, the lack-of-fit tests showed insignificant p-values ($p \geq .05$) for all predictors in all equations of the three-way serial mediation model. For the residual plots as well as a detailed overview of the results of the lack-of-fit tests, please consult Appendix B. Continuing with the second assumption, all VIF scores were far beneath 10 and all Tolerance values were far above .10, suggesting the assumption of no multicollinearity to be met. For an overview of the exact values, please reference Appendix B. Continuing with the third assumption, significant p-values on the Shapiro-Wilk test indicate normality to be violated for the third as well as the fourth equation of the model. For the results of the Shapiro-Wilk test, please consult Appendix B. Last but not least, regarding the fourth assumption, in the residual plots for the second and third equation of the three-way serial mediation model, the dots appear to be relatively randomly and evenly spread, suggesting homoscedasticity. In the residual plot for the fourth equation this, however, does not seem to be the case, suggesting a partial violation of the homoscedasticity assumption. For the residual plots for all equations, please reference Appendix B.

Due to discussed deviations from normality of the residuals as well as heteroscedasticity of the residuals for some of the equations of the statistical model, additional measures were taken in setting up the mediation analysis in order to still ensure the validity of the results. More specifically, robust bootstrap confidence intervals for the regression coefficients of each regression equation defining the model were requested. This was done in

order to be able to examine (non)significance of effects not only based on p-values as these type of significance statistics could be biased by violations of the normality assumption, but also based on bootstrap confidence intervals which are relatively robust against such violation (Hayes, 2018). Moreover, to ensure the validity of the results despite the violation of the homoscedasticity assumption, the analysis was performed with the HC3 estimator (Hayes & Cai, 2007). For the complete syntax of the three-way serial mediation analysis utilized to assess the hypotheses of the study at hand, please refer Appendix C.

Manipulation-Check

Assumption Testing

Initially, the manipulation check was planned to be conducted by means of an independent samples t-test. In order to assess the feasibility of this analysis plan, it was examined whether the data meets two crucial assumptions underlying this parametric statistical test. These assumptions are normality and homogeneity of variances. Applied to independent samples t-tests, the assumption of normality refers to the idea that the dependent variable should be approximately normally distributed for each group of the independent variable (Field, 2013). In order to assess this assumption, a Shapiro-Wilk test, utilizing *Condition* as the independent- and *Perspective-Taking Attempts* as the dependent variable, was conducted. Insignificant p-values ($p \geq .05$) indicate the normality assumption to be met, whereas significant p-values suggest it to be violated (Field, 2013). The assumption of homogeneity of variances, on the other hand, means that the variance of the dependent variable must be equal over all groups of the independent variable (Field, 2013). In order to assess this assumption a Levene's test was conducted. Again*, Condition: Perspective-Taking Instruction (Yes/ No)* served as the independent variable whereas *Perspective-Taking Attempts* was utilized as the dependent variable. Insignificant p-values ($p \geq .05$) suggest the homogeneity assumption to be met whereas insignificant p-values suggest it to be violated (Field, 2013). Similarly to the assumption testing conducted for the PROCESS-performed serial mediation analysis, the assumption of independence was not explicitly assessed because it could be assumed to be met based on the research design (Field, 2013; Hair et al., 2019).

Results of Assumption Testing and Implications for Planned Statistical Analysis

Starting with the first assumption, significant p-values on the Shapiro-Wilk test indicate normality to be violated for both levels of the independent variable, meaning that in both, the experimental and the control group, participants' average perspective-taking attempts scores appear to not be normally distributed ($W(48) = .913$, $p = .002$ and $W(54) = .916$, $p = .001$, respectively). Continuing with the second assumption, an insignificant p-value

on the Levene's test indicates the assumptions of homogeneity of variances to be fulfilled ($F(1,100) = 1.542$, $p = .217$). As a consequence of the detected deviations from normality, the initially planned analysis of an independent samples t-test was not suitable. Instead, an independent sample Mann-Whitney U test was chosen as non-parametric tests do not require the normality assumption to be fulfilled (Field, 2013). Within the Mann-Whitney U test, *Condition: Perspective-Taking Instruction (Yes/ No*) was utilized as the independent variable while *Perspective-Taking Attempts* served as the dependent variable.

Research Ethics

Several measurements to ensure the research's adherence to research ethics were taken (American Psychological Association, 2017). First, consensual rather than coerced participation in the study was ensured by 1) not offering reimbursements for partaking, 2) explicitly stating that participation is voluntarily and can be stopped at any given point in time without the risk of adverse consequences and 3) actively asking for instead of assuming consent at the outset of the study. Second, informed consent was safeguarded to the highest extent possible without interfering with the experimental manipulation. More specifically this means that, at the outset of the study, participants were informed that the research aim is to examine social media users' perceptions of and reactions to social media communications. To reduce the risk of confounding factors, information about the experimental manipulation, that is about the fact that, depending on their condition, they would vs. would not receive a short perspective-taking intervention, was not yet provided to participants. To the extent that this deception was necessary, a full debriefing was provided after completion, that is on the last page, of the study to conform to ethical guidelines. Third, confidentiality was ensured by 1) collecting as little demographic data as possible and 2) gathering data anonymously by selecting the "anonymize responses" option in Qualtrics, an option that ensured that no personally identifiable information such as for instance the IP address was collected. Due to the aforementioned anonymous nature of the data collection, offering participants the opportunity to have their responses deleted from the data set and the results of the present study at a later point in time, meaning after their completion of the study, was impossible. This is because it would have been impossible to know which responses belong to which participant. In order to ensure that the study still conforms to ethical standards, participating social media users were thus explicitly informed about this issue twice. The first time they were informed about the fact that, due to the anonymous nature of the data collection, it is impossible to have their responses excluded from the data set and results of the present study after completion was in the informed consent letter, before being presented with the consent

form. The second time they were provided with this information was after having been presented with the debriefing of the study at hand, meaning after they have been made aware of the true nature of the study. As such, after having been presented with the debriefing, participants were again actively asked for permission to have their responses included in the data set and results of the present study and informed that if they agreed it would be impossible to delete their responses at a later point in time. Furthermore, they were ensured that there would be no adverse consequences if they declined and that, in this case, their responses would be deleted immediately.

Results

Descriptive Analyses

For a written overview of participants' average dehumanization-, empathy-, threat perception-, and counter speech intention scores within and across experimental conditions as well as their average reported level of perspective-taking attempts, please consult Table 4.

Table 4

Overview of Participants' Average Dehumanization-, Empathy-, Threat Perception-, and Counter Speech Intention Scores as Well as Their Average Reported Levels of Perspective-Taking Attempts Within and Across Conditions.

Condition	Experimental – Perspective-Taking Encouragement: Yes (n = 48)		Control – Perspective-Taking Encouragement: No (n = 54)		Full Sample (n = 102)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Users' Perceptions and Intentions						
Dehumanization	64.250	32.204	85.241	32.250	75.363	33.752
Empathy	23.854	8.257	18.074	7.883	20.794	8.529
Threat Perceptions	14.313	4.785	11.982	5.688	13.078	5.386
Counter Speech Intentions	3.830	2.137	2.910	1.794	3.340	2.007
Manipulation Check						
Perspective-Taking Attempts	5.020	1.604	3.610	1.764	4.270	1.825

Hypothesis Testing

The following paragraphs contain the results of the statistical analysis and their implications for the hypotheses of the present research. For a graphical representation of the results, including the unstandardized regression coefficients for the significant pathways, please reference Figure 3. For a detailed tabular overview of the acceptance criteria presented in the previous chapter and the respective decisions for both research hypotheses, please consult Table 5.

Hypothesis H1:

Condition: Perspective-Taking Instruction (Yes/ No) had a significant positive total effect on *Counter Speech Intentions* (*B* = .926, *t*(100) = 2.330, *p* = .022, 95% CI [.138, 1.714]) with participating social media users who received the perspective-taking instruction, on average, demonstrating higher counter speech intentions than participants who were presented with the control task (*M* = 3.830 and *M* = 2.910, respectively). This means that the short perspective-taking encouragement utilized in the present study indeed was an effective strategy in enhancing social media users' counter speech intentions towards hateful content on SMPs. Thus, the first hypothesis of the present study was confirmed.

Hypothesis H2:

Furthermore, the aforementioned significant positive total effect was fully mediated by 1) *Dehumanization*, 2) *Empathy* and 3) *Threat Perceptions* in serial. In this regard, the indirect pathway of the effect of *Condition: Perspective-Taking Instruction (Yes/ No*) on *Counter Speech Intentions* via *Dehumanization, Empathy* and *Threat Perceptions* was significant (*B* = .087, 95% CI [.012, .220]). This pathway fully accounted for the overall impact of *Condition: Perspective-Taking Instruction (Yes/ No)* on *Counter Speech Intentions* with the direct effect being non-significant (*B* = .150, *t*(100) = .384, *p* = .702, 95% CI [-.556, .930]) and all remaining indirect pathways being non-significant as well. The indirect pathways of the effect of *Condition: Perspective-Taking Instruction (Yes/ No*) on *Counter Speech Intentions* via 1) *Dehumanization* alone (*B* = .217, 95% CI [-.154, .579]) *Empathy* alone (*B* = .073, 95% CI [-.144, .253]) and 3) *Threat Perceptions* alone (*B* = .232, 95% CI [-.137, .680]) were non-significant. Similarly, the indirect pathways of the effect of *Condition Perspective-Taking Instruction (Yes/ No)* on *Counter Speech Intentions* via 1) only *Dehumanization* and *Empathy* (*B* = .058, 95% CI [-.071, .307]), 2) only *Dehumanization* and *Threat Perceptions* (*B* = -.001, 95% CI [-.191, .159]) and 3) only *Empathy* and *Threat Perceptions* (*B* = .111, 95% CI [-.005, .255]) were non-significant.

Examining the nature of this full serial mediation via 1) *Dehumanization*, 2) *Empathy* and 3) *Threat Perceptions* in detail by considering the individual effects of the significant indirect pathway reveals that receiving encouragement to adopt the perspective of hate speech targets caused participating social media users in the experimental condition to display significantly lower dehumanization of the people that were targeted by the fictional hate speech post than participating social media users in the control condition (*B* = -20.991, *t*(100) = -3.251, *p* = .002, 95% CI [-33.607, -8.690]). This reduced denial of human attributes to these victims of the fictional hate speech post, in turn, significantly increased the extent to

which social media users empathized with those targets (B = -.121, t(100) = -4.433, p < .001, 95% CI [-.173, -.070]). Facilitated empathy with the hate speech victims, on the other hand, gave rise to a significant increase in social media users threat perceptions, meaning that it prompted them to perceive the hate speech incident in question as posing significantly higher (threat of) harm to the people targeted by the hate speech (B = .186, t(100) = 2.683, p = .009, 95% CI [.040, .312]). Those enhanced threat perceptions then ultimately gave rise to the aforementioned elevation of counter speech intentions in the experimental- as compared to the control condition (B = .184, t(100) = 5.899, p < .001, 95% CI [.123, .241]).

Taken together, the aforementioned results regarding the existence and nature of the full serial mediation via *1) Dehumanization, 2) Empathy and 3) Threat Perceptions* suggest that the observed effectiveness of the utilized brief perspective-taking encouragement in terms of social media users' counter speech intentions was indeed fully attributable to a causal chain of reduced dehumanization of hate speech targets, increased empathy with those victims and enhanced threat perceptions in response to hate speech. Therefore, the second hypothesis of the present study was confirmed as well.

Figure 3

Graphical Representation of the Results Including the Unstandardized Regression Coefficients for the Significant Pathways

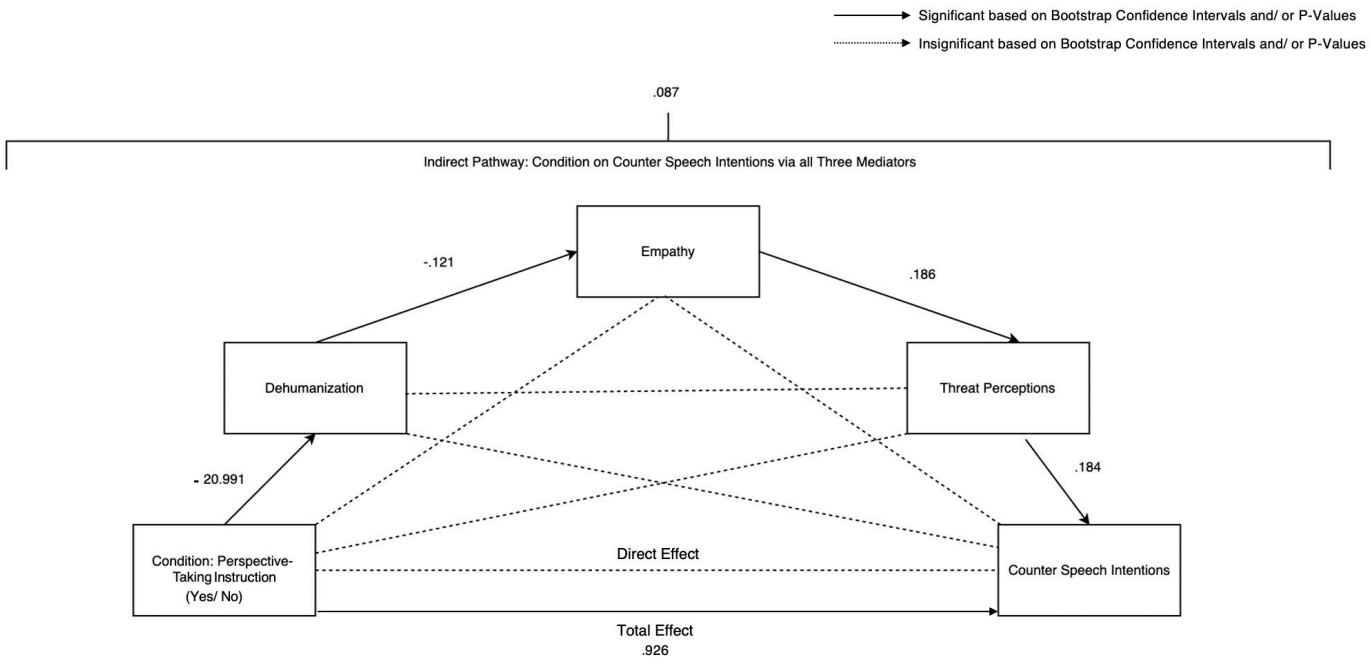


Table 5

Decision for Acceptance or Rejection of the Hypotheses of the Present Study

Hypothesis	H1: The provision of a short perspective-taking encouragement is an effective strategy for increasing social media users' counter speech intentions		H2: This effectiveness is fully attributable to a causal chain of reduced dehumanization, increased empathy and enhanced threat perceptions	
	Requirements for Acceptance	Results	Requirements for Acceptance	Results
Effect/ Pathway				
Total	Significant; Positive regression coefficient	✓	Significant; Positive regression coefficient	✓
Direct	N.A.	N.A.	Non-significant	✓
Indirect via all Three Mediators	N.A.	N.A.	Significant; Positive regression coefficient	✓
Other Indirect	N.A.	N.A.	Non-significant	✓
Condition on Dehumanization	N.A.	N.A.	Significant; Negative regression coefficient	✓
Dehumanization on Empathy	N.A.	N.A.	Significant; Negative regression coefficient	✓
Empathy on Threat Perceptions	N.A.	N.A.	Significant; Positive regression coefficient	✓
Threat Perceptions on Counter Speech Intentions	N.A.	N.A.	Significant; Positive regression coefficient	✓
Decision		✓		✓

Note: The abbreviation "N.A." stands for non-applicable, meaning that the respective effects are irrelevant to the assessment of the given hypothesis.

Manipulation Check

Reported levels of perspective-taking attempts were higher in the experimental condition than in the control condition ($M = 5.020$ and $M = 3.610$, respectively). A Mann-Whitney U test indicated this difference to be statistically significant ($U = 1878.500$, $p < .001$). More specifically, these results indicate that participating social media users who received the perspective-taking encouragement indeed attempted to adopt the perspective of hate speech targets to a significantly higher degree than participants in the control condition, thus suggesting a capability of the examined intervention strategy to generate compliance among its receivers within the sample.

Discussion

The study at hand investigated whether providing social media users with an encouragement to adopt the perspective of hate speech targets is an effective strategy to enhance their counter speech intentions when confronted with hateful content. Furthermore, it examined whether potential effectiveness is fully attributable to a causal chain of reduced dehumanization of hate speech victims, increased empathy with these targets and facilitated threat perceptions. The results of the utilized experimental online investigation revealed that participating social media users in the experimental condition, meaning users who were presented with the encouragement to visualize what hate speech targets may be feeling, thinking and experiencing, displayed significantly higher counter speech intentions when confronted with a fictional hate speech post than participating social media users who received a control instruction. In addition, the results revealed that the aforementioned significant differences between the two groups were fully mediated by reduced dehumanization, enhanced empathy and facilitated threat perceptions in serial. As established, this means that participating social media users who received the perspective-taking encouragement exhibited higher counter speech intentions than participating social media users in the control condition because receiving encouragement to adopt the perspective of hate speech targets reduced the extent to which participating social media users in the experimental condition denied human attributes to the victims of the fictional hate speech post which then prompted them to display facilitated empathy with these victims, ultimately enhancing the extent to which they perceived the hate speech incident in question as (potentially) harmful to the victims affected by it. Hence, confirming both, the first and the second research hypothesis, the answer to the research question of the present study is:

The provision of a brief perspective-taking encouragement is an effective strategy to increase social media users' counter speech intentions toward hate speech. This effectiveness is fully attributable to a causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions.

Having adopted a micro-level focus by approaching a gap in the hate speech governance literature through a sociopsychological lens, the present research offers important theoretical contributions and implications for both fields. In the following sections, these contributions and implications will be discussed separately, starting with hate speech governance.

Theoretical Contributions and Implications for Hate Speech Governance

As elaborated upon in the introduction of the present paper, combatting the rising prevalence of hate speech on SMPs by means of governance is an increasingly important but also complex contemporary societal issue (Schwoon et al., 2021). For effective and appropriate governance of hateful content on SMPs, the importance of civic engagement in the form of user-initiated counter speech is increasingly acknowledged. In this regard, user self-governance in the form of counter speech is often described as preferable to other governance attempts due to 1) not infringing on freedom of expression and 2) a potentially higher effectiveness in reducing the prevalence of hateful content online and thus in protecting the human dignity of targets and society as a whole from the, often severe, harm such speech may cause (Benesch et al., 2016; Gagliardone et al., 2015; Strossen, 2018). However, while the benefits of user-initiated counter speech as compared to previous governance attempts for combatting the contemporary problem of cyberhate are increasingly acknowledged, it is also clear that its full potential is not yet unlocked with its effectiveness in influencing offenders to change their behaviour and thus in reducing the prevalence of online hate speech increasing with a larger number of counter speakers and the number of users willing to engage in counter speech oftentimes being limited (Buerger & Wright, 2019; Jubany & Roiha, 2015). While this clearly emphasizes the need for strategies that are effective in enhancing counter speech intentions among a large proportion of social media users in order to be able to fully exploit the potential of user-initiated counter speech as a promising solution for tackling the contemporary problem of cyberhate, up to now research in this area is highly scarce (Leonhard et al., 2018).

By not only empirically investigating and, ultimately proving, the effectiveness of a brief perspective-taking encouragement as a strategy to increase social media users' counter speech intentions but also assessing and, ultimately illustrating, the underlying mechanism of this effectiveness by means of serial mediation, the present study took two steps forward in closing the aforementioned research gap and thus in contributing to enhancing the potential of user self-governance in the form of counter speech as a promising solution for addressing the problem of a rising prevalence of hate speech on SMPs. First and quite obviously, it provided a strategy which can immediately be utilized in practice in order to enhance social media users' intentions to engage in counter speech when witnessing hate speech on SMPs. Second, by demonstrating that the effectiveness of the examined strategy was fully attributable to a causal chain of reduced dehumanization, enhanced empathy and facilitated threat perceptions, the present study offers a starting point for future research. In this regard, by expanding the

scientific understanding of the dynamics of counter speech intentions among users, the present study offers future research aiming to contribute to tackling the problem of a rising prevalence of hate speech on SMPs by means of enhancing the potential of user-initiated counter speech heretofore, a novel, not previously discovered factor to target. Having gained the knowledge that the short perspective-taking encouragement was effective because it reduced social media users' dehumanization of hate speech targets which then, mediated by facilitated empathy and threat perceptions, sparked increased counter speech intentions, future research aiming at increasing social media users' counter speech intentions, thus enhancing the potential of user-initiated counter speech as a promising solution for combatting the contemporary problem of cyberhate, could look into additional remedies of such dehumanization to inform their own strategy design attempts.

Theoretical Contributions and Implications for Social Psychology

Continuing with social psychology, the present study contributes to existing research in three main ways which will be elaborated upon in the section at hand.

To begin with the first contribution, the findings of the present research are in line with existing sociopsychological theories and research findings, thus offering (further) support for these theoretical claims and empirical findings as well as for their applicability in the context of counter speech and online hate speech. In this regard, the present finding that participating social media users' threat perceptions toward the hate speech incident in question, meaning the extent to which they perceived the hate speech post as (potentially) harmful to the victims affected by it, determined their counter speech intentions is in line with Latané and Darley's (1970) model of the decision-making process for helping behaviour. In this regard, the model describes threat perceptions as one of the four stages that eventually shape the degree of helping intentions which observers of a critical situation experience towards the person affected by the incident. Although this model has been investigated and validated in a variety of different empirical contexts, its application in the domain of social media users' counter speech intentions toward hateful content is still scarce, with the exception of Leonhard et al.'s (2018) study which served as a starting point for the present research. With replicating Leonhard et al.'s (2018) findings that the extent to which social media users perceive a given hate speech incident as (potentially) harmful to the victims affected by it determines their intentions to counterargue against the incident in question, the present research thus further emphasizes the model's relevance for the empirical context at hand, thus potentially opening the door for developing further strategies to enhance social

media users' counter speech intentions based on the same or other stages of the theorized decision-making process.

Furthermore, the fact that, in the present study, the extent to which participating social media users experienced empathy with the targets of the hate speech post determined their threat perceptions is in line with an existing sociopsychological theory, namely the TDM (Schein & Gray, 2018). In this regard, the TDM theorizes that higher empathy with the people who are affected by a given incident increases the extent to which individuals perceive the incident as (potentially) harmful to the people affected by it because empathy is one of the factors determining the extent to which the incident matches these individuals' cognitive template of harm (Schein & Gray, 2018). With, apart from Krueger et al.'s (2013) neuroscientific study, investigations of these claims still being scarce or, as in the case of online hate speech, absent, the present study thus offers (further) empirical support for the theory's assumptions and their applicability to the context of online hate speech and counter speech.

Also, the present finding that participating social media users' empathy with the targets of the hate speech post increased with a reduction of their dehumanization of these hate speech victims is in line with and thus provides further support for the theorized but yet scarcely researched role of dehumanization as an empathy-constraining factor (Čehajić et al., 2009; Haslam & Stratemeyer, 2016; Murrow & Murrow, 2015; Schein & Gray, 2018).

The second, and perhaps biggest, theoretical contribution of the present research in the sociopsychological domain can be found in its integration of all of the aforementioned sociopsychological theories and research findings into one conceptual model. In this regard, although previous scholars have concerned themselves with the interrelationships between the variables forming the core of the present research, namely 1) dehumanization, 2) empathy, 3) threat perceptions and 4) intervention- or counter speech intentions, they generally limited themselves to examining or theorizing the relationships between only a subset of these variables. For instance, among others, Latané and Darley (1970) and Leonhard et al. (2018) assert that threat perceptions play a role in helping- or counter speech intentions, while Krueger et al. (2013) and Schein and Gray (2018) suggest a positive relationship between threat perceptions and empathy. Finally, among others, Čehajić, Brown and Gonzáles (2009) and Schein and Gray (2018) argue for the existence of a negative relationship between dehumanization and empathy. One of the biggest theoretical contributions of the present research is that it took these individual, disjoint theories and findings, proposed a single causal chain linking all these four variables together and tested the proposed causal chain,

using a single serial mediation model, ultimately verifying not only the relationships found or theorized by the above mentioned scholars but also the more complex single causal chain proposed in the present thesis. This is of crucial importance because, while individual results from different papers connecting the four theoretical constructs, namely 1) dehumanization, 2) empathy, 3) threat perceptions and 4) intervention- or counter speech intentions, may be conceptually indicative of the existence of an underlying causal chain linking all of them, the existence of such a chain would be impossible to confirm or reject based only on such individual results due to divergences in sample sizes, methodological choices and measurement scales. Thus, by providing empirical proof for the existence of a single causal chain linking all four variables together via a single statistical model, the present research made a crucial contribution to the study of these constructs and related theories.

To end with the third contribution, a novel finding emerged in the present research, namely that encouraging social media users to adopt the perspective of hate speech targets reduced their dehumanization of hate speech victims. This alleviating influence of the perspective-taking encouragement on social media users' dehumanization of the hate speech targets was expected based on previous literature. In this regard, previous studies did not only demonstrate the employed perspective-taking encouragement to induce compliance in terms of perspective-taking attempts, a finding replicated by the manipulation check included in the present study, but also perspective-taking attempts to directly target relatively long-lasting remedies of dehumanization such as increased feelings of social connectedness and perceptions of similarity towards the person whose perspective was adopted (Davis et al.,1996; Haslam & Stratemeyer, 2016; Hutcherson et al., 2008; Todd et al., 2011). Despite being expected based on the aforementioned research findings and despite perspective-taking encouragements previously having been theorized to be effective in reducing dehumanization in different empirical contexts, the present research is the first to provide empirical evidence hereto. This is a valuable insight because it is well known that dehumanization has a variety of negative consequences beyond reducing observers' empathy towards individuals that are affected by a negative situation. In this regard, other consequences of dehumanization include, among others, aggressive behaviour towards the person who is dehumanized (Haslam, 2006). Furthermore, to provide a recent example, dehumanization has even been linked to risk perceptions and conspiracy beliefs regarding COVID-19 with individuals who dehumanized people of an Asian descent to a higher extent being less likely to perceive the virus as a health risk for others or themselves, being less likely to believe that they could contract it or assuming it to be a biochemical weapon (Markowitz et al., 2021). Although it

yet remains to be seen whether the established alleviating influence of the brief perspective-taking encouragement generalizes beyond the current empirical context, the present research potentially offers some starting points for researchers to investigate possible solutions to the aforementioned problems as well as related issues in which dehumanization plays a focal role. If the alleviating influence of perspective-taking encouragements on dehumanization generalizes beyond the current empirical context and beyond the specific encouragement utilized in the present study, it is possible that the aforementioned and related issues, including aggressive and potentially even hateful behaviour on SMPs could be tackled by encouraging people to engage in perspective-taking. However, future research is needed to draw conclusions in this regard.

Practical Implications

The rising prevalence of hate speech on SMPs is a serious contemporary problem with potential consequences including not only extensive psychological damage and physical harm for the individuals and social groups targeted by it but also harm for society as a whole by interfering with the platforms' ability to foster each individual's active participation in democracy and threatening societal peace (c.f. e.g. Citron & Norton, 2011; Edwards et al., 2021; Nemes, 2002; Ring, 2013). With the need of combatting the rising prevalence of hate speech on SMPs thus being clear, user-initiated counter speech possessing several advantages over prevailing governance attempts in this regard, its potential as a promising solution hereto increasing with a larger number of counter speakers and the number of users willing to engage in counter speech oftentimes being limited (Benesch et al., 2016; Buerger & Wright, 2019; Gagliardone et al., 2015; Jubany & Roiha, 2015; Strossen, 2018), the strategy investigated by the present study should be implemented in practice. In implementing this affordable and effective strategy to enhance social media users' counter speech intentions, attention should be paid to reach as many social media users as possible for maximum impact regarding a reduction of cyberhate. In the following section, concrete implementation recommendations for different actors, different channels and different subgroups of the target audience will be provided.

One concrete recommendation to implement the examined strategy is to incorporate the encouragement to adopt the perspective of hate speech targets by visualizing what targets of hate speech may be feeling, thinking and experiencing into existing awareness-raising campaigns on hate speech. In this regard, awareness-raising campaigns on hate speech are already regularly executed by different actors. Examples include, among many others, the Hate Hurts Wales campaign and the #EngageResponsibly initiative. Starting with the former,

the Hate Hurts Wales campaign is a Welsh governmental initiative attempting to raise awareness for both, hate crimes and offline as well as online hate speech. Besides providing victim resources, it involves efforts to raise public awareness and understanding of the issue of hate and its consequences for targets and wider society by including real-life experience reports (Bridgend Association of Voluntary Organisations, n.d.). Continuing with the latter, the #EngageResponsibly initiative is a multi-stakeholder effort piloting in the US, involving, among others, the Association of National Advertisers (ANA), the Global Alliance for Responsible Media (GARM), Non-Governmental Organizations (NGOs), multiple big brands and small and medium-sized businesses as well as social media companies. It involves educational efforts including the provision of information on the issue of online hate speech, its consequences in terms of real-world violence as well as suggestions on how to take action (Markets Insider, 2021; PRNewswire, 2021). Considering previous literature and the results of the present study, adding the investigated perspective-taking encouragement to these and related campaigns is highly recommended to tackle the problem of cyberhate on SMPs and thus for reducing or preventing its detrimental consequences for hate speech targets and society as a whole. In this regard, based on the present results, adding this brief encouragement is expected to make a major difference in terms of increasing social media users' intentions to take action against online hate speech incidents by means of counterarguing and thus in increasing the potential of counter speech as a promising solution for reducing hate speech on SMPs but, due to its limited length and material requirements, is unlikely to increase the costs or workload associated with these campaigns in any significant way.

Another concrete suggestion to implement the examined strategy which specifically targets children and adolescents would be to incorporate the examined strategy into existing pedagogical workshops on online safety in schools or to launch new workshops specifically focussing on hate speech on SMPs including the investigated perspective-taking encouragement. Similarly, to reach young adults, information sessions on online hate speech on SMPs incorporating the examined strategy could also be launched in universities and vocational colleges.

In order to also reach older individuals of the target group and those having restricted access to higher education, another, relatively inexpensive, option to implement the examined strategy may be to distribute posters and flyers, starring the encouragement that was provided to the experimental group of the present study. Similarly, to maximize the strategy's reach

and thus its practical utility, another inexpensive option would be for SMPs to distribute the provided instruction on their platforms.

Limitations and Future Research

Although careful attention has been paid to constructing and conducting the present study in the best possible way, the derived conclusions, implications and recommendations should be read with certain limitations kept in mind.

Starting with the methodology of the present study, four main limitations need to be acknowledged. First, the present study assessed participating social media users' intentions to counterargue against a fictional hate speech incident in a laboratory setting rather than against an existent hate speech post in a field setting. The main rationale underlying this choice was controllability. Albeit well justified and in line with methodology utilized by previous research (Leonhard et al., 2018), the chosen approach carries certain limitations. In this regard, previous research suggests that there can be small discrepancies between individuals' beliefs regarding how they would act in certain situations and their actual behaviour when confronted with these situations in reality (Kang, Rangel, Camus, & Camerer, 2011). Hence, it is possible that the provided perspective-taking encouragement aiming at enhancing social media users' counter speech intentions in response to hateful content vindicates less effective in a field setting than what would initially be expected from the results of the present study. Considering the fact that, in the present study, effectiveness was very strong, as indicated by significance statistics which were far from the cut-off point, it is, however, highly likely that the examined strategy is still effective in enhancing counter speech intentions in a field setting, albeit perhaps to a lower extent. Employing qualitative observational research to validate the results of the present study in a field setting may be an opportunity to arrive at decisive conclusions in this regard.

The second limitation also relates to the hate speech post utilized in the study at hand. In this regard, the hate speech post selected to assess the effectiveness of a perspective-taking encouragement in terms of enhancing social media users' counter speech intentions as well as the underlying mechanism possessed very specific stimulus features. As such, following the example of Leonhard et al. (2018), only one type of hate speech post, namely a racial hate speech post targeting ethnic minorities, more specifically refugees, was utilized. Although in line with previous research and well justified given the limited scope of the master thesis trajectory, the aforementioned choice comes with certain limitations regarding the generalizability of the findings and conclusions of the study at hand. The present study provides evidence for the effectiveness of the utilized perspective-taking encouragement on

social media users' counter speech intentions as well as insights into the underlying causal chain in the case of racial cyberhate. However, the extent to which these findings regarding the effectiveness of the examined intervention strategy and the underlying mechanism of this effectiveness can be generalized to other forms of hate speech on SMPs such as for instance gendered cyberhate is yet unclear. While generalizability was attempted to be ensured to the greatest extent possible in the chosen research design by utilizing a perspective-taking instruction generic to all hate speech targets instead of an instruction specifically focussing on adopting the perspective of refugees and ethnic minorities, potential differences in findings according to the various forms of hate speech could not explicitly be assessed. Thus, although generalizability may be relatively safely assumed, from the present study, no certain conclusions regarding the effectiveness of the utilized perspective-taking encouragement and the underlying mechanism of this effectiveness can be drawn beyond racial hate speech posts. As a consequence, in order to draw definitive conclusions regarding the effectiveness of the utilized perspective-taking encouragement on social media users' counter speech intentions in response to other types of cyberhate, future research is needed. In this regard, it may be worthwhile to replicate the present study with a different type of hate speech post such as for instance a hate speech post targeting women or the LGBTQIA+ community. Perhaps even a comparative design could be employed.

A third limitation relates to the generalizability of the effectiveness of the brief perspective-taking encouragement utilized in the present study to the effectiveness of other short perspective-taking instructions. In this regard, it is unclear whether and to what extent the results can be generalized to other brief perspective taking instructions that possess different characteristics such as for instance different wording or emotional charge. More specifically, this means that, from the present findings, it is impossible to say whether providing social media users with a different short perspective taking encouragement will be (equally) effective in reducing their dehumanization of hate speech targets and thus, by enhancing their empathy with these victims and facilitating their threat perceptions, (equally) effective in increasing their counter speech intentions. While it is likely that perspective-taking instructions which are known to induce compliance in terms of perspective-taking attempts are (equally) effective in terms of producing the desired outcomes, future research is needed to reach clear conclusions. In order to address this limitation, larger-scale studies are needed. More specifically, future studies could replicate the present study while exposing participating social media users to a range of different brief perspective-taking instructions

and then investigate whether there are differences in effectiveness according to the type of encouragement utilized.

Fourth, due to not employing a longitudinal design, no clear information about the long-term effectiveness of the utilized brief perspective-taking encouragement on social media users' counter speech intentions can be derived from the present study. Previous literature suggests that attempting to adopt another person's perspective enhances perceptions and feelings that are relatively long-lasting remedies of dehumanization, namely perceived similarity and feelings of social connectedness (Davis et al., 1996; Haslam & Stratemeyer, 2016; Hutcherson et al., 2008; Todd et al., 2011). Consequently and based on the insight of the present study that the effectiveness of the utilized perspective-taking encouragement was fully attributable to a causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions, long-term effectiveness of the examined strategy is likely. However, without explicitly assessing longevity by means of a longitudinal design, no definitive conclusions can be drawn in this regard. Hence, in order to arrive at definitive conclusions regarding the long-term effectiveness of the administered short perspective-taking intervention on social media users' counter speech intentions, future studies should consider utilizing a longitudinal research design.

A last aspect that deserves mentioning and needs to be taken into account is an error on the side of the researcher. In this regard, when setting up the study in Qualtrics, one of the utilized measurement scales, namely the scale to assess empathy was unintentionally changed. In this regard, the original scoring of this scale occurs on a 7-point scale (1 = not at all, 7 = extremely) (Batson & Ahmad, 2001). Unfortunately, due to the fact that only the extreme ends of the scale are labelled, when setting up the study in Qualtrics, one scale point was overlooked. As a consequence, in the present study, participants' responses on the empathy measurement items were scored on a 6-point- rather than on a 7-point scale. While this change is unlikely to have affected the results of the study in any way, future researchers potentially aiming to replicate the present study should be aware of this incident.

Further Suggestions for Future Research

Besides addressing the aforementioned limitations of the present study wherever possible, there are other aspects that future research could take into account.

Firstly, as mentioned in the first theoretical implications section of the present chapter, future studies aiming at contributing to the research avenue of developing potential intervention strategies to increase social media users' intentions to counterargue against hate speech on SMPs could take the present study as their starting point. Considering the insight

that the effectiveness of the examined perspective-taking intervention was fully attributable to a causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions, future researchers could look into additional opportunities to decrease social media users' dehumanization of hate speech victims to inform their own strategy design attempts.

Secondly and lastly, future strategy design attempts could focus on the other stages of Latané and Darley's (1970) model of the decision-making process for helping behaviour. In this regard, according to this model, observers' threat perceptions, meaning the extent to which they perceive a critical situation as (potentially) harmful to those affected by it, is only one of the factors influencing their intentions to intervene in the situation on behalf of those victims. The extent to which the other stages of the model are successfully completed also has an influence on the extent to which observers intend to intervene in a critical situation on behalf of the victims affected by the situation (Latané & Darley, 1970). The other stages of the model include, among others, assuming personal responsibility for intervening in the incident in question and reflecting on as well as knowing how to do so. Especially the third stage of the decision-making process for helping behaviour, meaning feeling personally responsible to help the victims affected by a given incident, may be of particular interest for future studies. In this regard, in addition to being investigated and established in multiple domains, the role of feelings of personal responsibility for intervention intentions has recently also been proven to apply to the case of social media users' intentions to counterargue against hateful content on SMPs (Leonhard et al., 2018). Consequently, future studies attempting to contribute to the development of potential intervention strategies to increase social media users' counter speech intentions could explore potential ways to increase the extent to which these users feel responsible to intervene in hate speech incidents on behalf of the hate speech targets.

Conclusion

To conclude, in an experimental manner, the study at hand investigated not only the effectiveness of the provision of a brief perspective-taking encouragement as a strategy for enhancing social media users' intentions to counterargue against hateful content on SMPs but also the underlying mechanism of potential effectiveness. Findings suggest that the provided short perspective-taking instruction is indeed effective in increasing users' counter speech intentions and that this effectiveness is fully attributable to a causal chain of reduced dehumanization, increased empathy and facilitated threat perceptions. More specifically, the present findings indicate that the brief perspective-taking encouragement is effective in enhancing social media users' intentions to counterargue against hate speech because it reduces the extent to which they dehumanize the people that are targeted by a given hate speech post which in turn enhances their empathy with these victims. Enhanced empathy, on the other hand, causes social media users to perceive the hate speech incident as (potentially) harmful to the hate speech targets to a greater extent. Facilitated threat perceptions then ultimately give rise to increased counter speech intentions. These findings carry crucial practical and theoretical implications discussed extensively throughout this thesis.

Reflection on the Thesis Project

Conducting the research project at hand within the master thesis trajectory of the "Organisational Design & Development" program of Radboud University was a challenging but highly educational experience. Starting with the former aspect, the main challenge was to plan, conduct and, above all, write up a graduation project at a faculty at which I only had a single year of experience, namely at the Nijmegen School of Management. It took a lot of dedication and hard work to reflect on practices and structural regulations that I, over the years, have internalized from my previous study to the extent that these internalized practices and regulations became consciously available and thus possible to adapt. To name an example, in the beginning, it was difficult for me to write the introduction and the theory chapter separately from each other because I did not have any prior experiences with making such a split. Furthermore, the aforementioned challenge became aggravated by personal traits and circumstances. Starting with the personal traits, the main difficulties were my often perfectionistic attitudes paired with my insecurities. In this regard, I regularly questioned whether what I am doing, for instance regarding the introduction-theory separation, is in line with what is expected of me in this master program. Paired with my ambition to deliver a good final product, these insecurities caused me to experience a lot of stress. Continuing with the personal circumstances, the fact that the last year was not an easy year for me with not only several private problems but also recurring illnesses, further contributed to the aforementioned experience of stress. Learning to deal with this stress, reflecting on my insecurities and perfectionism, adapting to a new faculty and pushing through regardless of all these aspects, on the other hand, provided me with valuable experiences for my professional future, irrespective of whether this future will be in academia or not. In this regard, it 1) caused me to look up and adopt novel stress- and time-management tactics (such as the pomodoro technique), 2) taught me the importance and acceptability of asking for professional feedback and emotional support, 3) made me realize the importance of reflecting on internal biases, 4) caused me to reflect on the adverse effects of my often perfectionistic attitudes and insecurities, and 5) upon completion provided me with novel confidence in my professional abilities. All these experiences are likely come in handy in my occupational future. In this regard, for the next major project, I would among other things, attempt to keep track of my stress-levels from the beginning and, should it be necessary, ask for feedback and support on time in order to avoid stress levels and illnesses to accumulate. Besides benefitting my well-being, I believe that doing so would also be helpful to avoid mistakes such as the research error regarding the empathy-scale made in the study at hand.

References

Alkiviadou, N. (2019). Hate speech on social media networks: Towards a regulatory framework? *Information & Communications Technology Law, 28*(1), 19-35. doi:10.1080/13600834.2018.1494417

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct [PDF file]. Retrieved from https://www.apa.org/ethics/code/ethics-code-2017.pdf

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology, 24*(3), 233-239. doi:10.1080/13600869.2010.522323

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-speech. Examining content that challenges extremism online [PDF file]. Retrieved from http://www.demos.co.uk/wp-content/up-loads/2015/10/Counter-speech.pdf

Batson, C. D., & Ahmad, N. (2001). Empathy-induced altruism in a prisoner's dilemma II: What if the target of empathy has defected? *European Journal of Social Psychology, 31*(1), 25-36. doi:10.1002/ejsp.26

Benesch, S., Ruths, D., Dillon, K., Saleem, H., & Wright, L. (2016). *Counterspeech on Twitter: A field study*. Retrieved from https://de.scribd.com/document/327586365/Counterspeech-on-Twitter-A-Field-Study

Brandwatch, & Ditch the Label. (2021). *Uncovered: Online hate speech in the covid era*. Retrieved from https://www.ditchthelabel.org/wp-content/uploads/2021/11/Uncovered_Online-_Hate_Speech_DTLxBW_V2-1.pdf

Bridgend Association of Voluntary Organisations. (n.d). *Hate hurts Wales: Let's stand up to hate crime together #hatehurtswales*. Retrieved from https://www.bavo.org.uk/hate-hurts-wales-lets-stand-up-to-hate-crime-together-hatehurtswales/

Buerger, C., & Wright, L. (2019). Counterspeech: A literature review. *SSRN*. Retrieved from https://deliverypdf.ssrn.com/delivery.php?ID=628003127096098108092074113093066024029024036061084064120099097030084000101072029104060110030124042026061123122116116086016125126034071019079109117109093115085087022071044031100080085027100069021095104125125015089123088091065012028083118127075111015106&EXT=pdf&INDEX=TRUE

Castaño-Pulgarin, S. A., Suárez-Betancur, N., Tilano Vega, L. M., & Herrera López, H. M. (2021). Internet, social media and online hate speech: Systematic review. *Aggression and Violent Behavior, 58*. doi:10.1016/j.avb.2021.101608

Čehajić, S., Brown, R., & Gonzáles, R. (2009). What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. *Group Processes & Intergroup Relations, 12*(6), 715-729. doi:10.1177/1368430209347727

Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review, 91*, 1435-1484. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/bulr91&collection=journals&id=1443&startid=&endid=1492#

Clark, M., Fine, M. B., & Scheuer, C. (2017). Relationship quality in higher education marketing: The role of social media engagement. *Journal of Marketing for Higher Education, 27*(1), 40-58. doi:10.1080/08841241.2016.1269036

Coon, D., & Mitterer, J. O. (2014). *Psychology: A journey*. Belmont, CA: Cengage Learning.

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology, 70*(4), 713-726. doi:10.1037/0022-3514.70.4.713

De Smedt, T., De Pauw, G., & Van Ostaeyen, P. (2018). *Automatic detection of online jihadist hate speech*. (CLiPS Technical Report No. 7). Antwerp, Belgium: University of Antwerp, Computational Linguistics & Psycholinguistics Research Center.

Edwards, A., Webb, H., Housley, W., Beneito-Montagut, R., Procter, R., & Jirotka, M. (2021). Forecasting the governance of harmful social media communications: Findings from the digital wildfire policy Delphi. *Policing and Society, 31*(1), 1-19. doi:10.1080/10439463.2020.1839073

Fasoli, F., Paladino, M. P., Carnaghi, A., Jetten, J., Bastian, B., & Bain, P. G. (2016). Not "just words": Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology, 46*(2), 237-248. doi:10.1002/ejsp.2148

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics*. Thousand Oaks, CA: SAGE Publications.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., … Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517-537. doi:10.1037/a0023304

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression*. Thousand Oaks, CA: SAGE Publications.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Paris, France: UNESCO.

Gimmler, A. (2001). Deliberative democracy, the public sphere and the internet. *Philosophy & Social Criticism, 27*(4), 21–39. doi:10.1177/019145370102700402

Grygiel, J., & Brown, N. (2019). Are social media companies motivated to be good corporate citizens? Examination of the connection between corporate social responsibility and social media safety. *Telecommunications Policy, 43*, 445-460. doi:10.1016/j.telpol.2018.12.003

Guo, L., & Johnson, B. G. (2020). Third-person effect and hate speech censorship on Facebook. *Social Media + Society, 6*(2). doi:10.1177/2056306120923003

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*. Andover, United Kingdom: Cengage.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252-264. doi:10.1207/s15327957pspr1003_4

Haslam, N., & Stratemeyer, M. (2016). Recent research on dehumanization. *Current Opinion in Psychology, 11*, 25-29. doi:10.1016/j.copsyc.2016.03.009

Haslam, S. A., & McGarty, C. (2004). Experimental design and causality in social psychological research. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The SAGE handbook of methods in social psychology* (pp. 237-264). Thousand Oaks, CA: SAGE publications.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY: The Guilford Press.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*(4), 709-722. doi:10.3758/bf03192961

Hutcherson, C. A., Seppala, E., & Gross, J. J. (2008). Loving-kindness mediation increases social connectedness. *Emotion, 8*(5), 720-724. doi:10.1037/a0013237

Jane, E. A. (2017). Gendered cyberhate, victim-blaming, and why the internet is more like driving a car on a road than being naked in the snow. In E. Martellozzo, & E. A. Jane (Eds.), *Cybercrime and its victims* (pp. 61-78). doi:10.4324/9781315637198-4

Jubany, O., & Roiha, M. (2015). *Backgrounds, experiences and responses to online hate speech: A comparative cross-country analysis*. Retrieved from https://sosracismo.eu/wp-content/uploads/2016/07/Backgrounds-Experiences-and-Responses-to-Online-Hate-Speech.pdf

Kang, M. J., Rangel, A., Camus, M., & Camerer, C. F. (2011). Hypothetical and real choice differentially activate common valuation areas. *The Journal of Neuroscience, 31*(2), 461-468. doi:10.1523/JNEUROSCI.1583-10.2011

Kiesler, S., Kraut, R. E., Resnick, P., & Kittur, A. (2012). Regulating behavior in online communities. In R. E. Kraut & P. Resnick (Eds.), *Building successful online communities: Evidence-based social design*. Cambridge, MA: Massachusetts Institute of Technology.

Klausen, J., Tschaen Barbieri, E., Reichlin-Melnick, A., & Zelin, A. Y. (2012). The YouTube jihadists: A social network analysis of Al-Muhajiroun's propaganda campaign. *Terrorism Research Initiative, 6*(1), 36-53. Retrieved from https://www.jstor.org/stable/26298554?seq=1

Krueger, F., Parasuraman, R., Moody, L., Twieg, P., de Visser, E., McCabe, K., … Lee, M. R. (2013). Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. *Social Cognitive and Affective Neuroscience, 8*(5), 494-498. doi:10.1093/scan/nss026

Kumar, R. (2011). *Research methodology. A step-by-step guide for beginners*. Retrieved from http://www.sociology.kpi.ua/wp-content/uploads/2014/06/Ranjit_Kumar-Research_Methodology_A_Step-by-Step_G.pdf

Laaksonen, S., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020). The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data, 3*. doi:10.3389/fdata.2020.00003

Lammers, J., & Stapel, D. A. (2011). Power increases dehumanization. *Group Processes & Intergroup Relations, 14*(1), 113-126. doi:10.1177/1368430210370042

Langvardt, K. (2018). Regulating online content moderation. *Georgetown Law Journal, 106*(5), 1353-1388. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/glj106&collection=journals&id=1367&startid=&endid=1402

Landesanstalt für Medien. (2020). *Ergebnisbericht: Forsa-befragung zu hate speech 2020*. Retrieved from https://www.medienanstalt-nrw.de/fileadmin/user_upload/Neue Website_0120/Themen/Hass/forsa_LFMNRW_Hassrede2020_Ergebnisbericht.pdf

Landesanstalt für Medien. (2021). *Ergebnisbericht: Forsa-befragung zu hate speech 2021*. Retrieved from https://www.medienanstalt-nrw.de/fileadmin/user_upload/Neue Website_0120/Themen/Hass/forsa_LFMNRW_Hassrede2021_Ergebnisbericht.pdf

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

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. *Studies in Communication and Media*, *7*(4), 555-579. doi:10.5771/2192-4007-2018-4-555

Markowitz, D. M., Shoots-Reinhard, B., Peters, E., Silverstein, M. C., Goodwin, R., & Bjällkebring, P. (2021). Dehumanization during the COVID-19 pandemic. *Frontiers in Psychology, 12*. doi:10.3389/fpsyg.2021.634543

Murrow, G. B., & Murrow, R. (2015). A hypothetical neurological association between dehumanization and human rights abuses. *Journal of Law and the Biosciences, 2*(2), 336-364. doi:10.1093/jlb/lsv015

Nemes, I. (2002). Regulating hate speech in cyberspace: Issues of desirability and efficacy. *Information & Communications Technology Law, 11*(3), 193-220. doi:10.1080 /1360083022000031902

Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society, 18*(8), 1491-1507. doi:10.1177/1461444814563519

PR Newswire. (2021, September 20). Advertising industry doubles down on commitments to combat online hate speech. *Markets Insider*. Retrieved from https://markets.business insider.com/news/stocks/advertising-industry-doubles-down-on-commitments-to-com bat-online-hate-speech-1030810145

Procter, R., Webb, H., Jirotka, M., Burnap, P., Housley, W., Edwards, A., & Williams, M. (2019). A study of cyber hate on Twitter with implications for social media governance strategies. *Proceedings of the 2019 Truth and Trust Online Conference*. doi:10.48550/arXiv.1908.11732

Puppis, M. (2010). Media governance: A new concept for the analysis of media policy and regulation. *Communication, Culture & Critique, 3*(2), 134-149. doi:10.1111/j.1753-9137.2010.01063.x

Rieger, D., Schmitt, J. B., & Frischlich, L. (2018). Hate and counter-voices in the internet: Introduction to the special issue. *Studies in Communication and Media, 7*(4), 459-472. doi:10.5771/2192-4007-2018-4-459

Ring, C. E. (2013). *Hate speech in social media: An exploration of the problem and its proposed solutions* (Doctoral dissertation). Retrieved from https://scholar.colorado. edu/concern/graduate_thesis_or_dissertations/12579s395

Sageman, M. (2004). *Understanding terror networks*. Philadelphia, PA: University of Pennsylvania Press.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgement by redefining harm. *Personality and Social Psychology Review, 22*(1), 32-70. doi:10.1177/1088868317698288

Schwoon, B., Schembera, S., & Scherer, A. G. (2021). *A multi-level process perspective on issue management in the context of grand challenges: The case of online hate speech governance at traditional media organizations*. Paper presented at the 37th EGOS Colloquium 2021, Amsterdam, the Netherlands.

Siegel, A. A. (2019). Online hate speech. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform*. Retrieved from https://alexandra-siegel.com/wp-content/uploads/2019/08/Siegel_Online_Hate_Speech_v2.pdf

Sonntag, J. (2019). Social norms and xenophobia: Evidence from Facebook [PDF file]. Retrieved from https://jfasonntag.github.io/docs/xenophobia.pdf

Statista Research Development. (2022a). *Daily time spent on social networking by internet users worldwide from 2012 to 2020*. Retrieved from https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista Research Development. (2022b). *Number of global social networks users 2017-2025*. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Strossen, N. (2018). *Hate: Why we should resist it with free speech, not censorship*. New York, NY: Oxford University Press.

Tahmasbi, F., Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., & Zannettou, S. (2021). „Go eat a bat, chang!“: On the emergence of sinophobic behavior on web communities in the face of COVID-19. *Proceedings of the Web Conference 2021*. doi:10.1145/3442381.3450024

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology, 100*(6), 1027-1042. doi:10.1037/a0022308

Tsesis, A. (2002). *Destructive Messages: How hate speech paves the way for harmful social movements*. New York, NY: New York University Press.

Uyheng, J., & Carley, K. M. (2021). Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Special Issue on Epidemics Dynamics & Control on Networks. Applied Network Science, 6*(1). doi:10.1007/s41109-021-00362-x

Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., . . . Burnap, P. (2015). Digital wildfires: Hyper-connectivity, havoc and a global ethos to govern social media. *ACM SIGCAS Computers and Society, 45*(3), 193-201. doi:10.1145/2874239.2874267

West, R., Michie, S., Atkins, L., Chadwick, P., & Lorencatto, F. (2019). *Achieving behaviour change: A guide for local government and partners*. London, England: Public Health England.

Wilson, R. A., & Land, M. K. (2021). Hate speech on social media: Content moderation in context. *Connecticut Law Review, 52*(3), 1029-1076. Retrieved from https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/conlr52&id=1056&men_tab=srchresults

Wright, L., Ruths, D., Dillon, K. P., Saleem, H. M., & Benesch, S. (2017). Vectors for counterspeech on twitter. *Proceedings of the First Workshop on Abusive Language Online*. doi:10.18653/v1/w17-3009

Appendix A: Research Materials

Informed Consent

Information Letter
Concerning the Study: Social Media Users' Perceptions of and Reactions to Social Media Communications

Introduction

Welcome and thank you for your interest in the research at hand. Are you of legal age, meaning at least 18 years old, and have at least one social media account? If yes, you are invited to participate in the research study at hand which is conducted within the master thesis trajectory of the "Organisational Design & Development" program at Radboud University Nijmegen (Netherlands). If you decide to participate, I will ask you to indicate your consent on the following page. Before you decide whether or not you want to take part in the study, I will, however, give you information about the study. Please take your time to read the following information letter carefully. If the provided information is not clear or if you would like to receive more information, please do not hesitate to contact me. My e-mail address is: sarah.simon@ru.nl.

Purpose of the Research Study and Procedure:

The purpose of this research is to examine social media users' perceptions of and reactions to social media communications. The study procedure involves being presented with examples of social media communications and filling out an online survey. In this research, there are no right or wrong answers. As a consequence, you are asked to fill out the questions of the online survey as honest and intuitively as possible. Participating in this research will take approximately 15 minutes. Participation will not be compensated for.

Anonymity of the Research Data:

In the study at hand, data will be collected completely anonymous. This means that no information will be collected that can identify who you are. Also, no information you share in this study can be traced electronically back to you or the computer that you used. As a consequence of the fact that your responses cannot be traced back to you, I cannot inform you

about your personal results. I can, however, inform you about the results of the study as a whole. If you wish to be informed about the results of the study as a whole, please do not hesitate to contact me under the e-mail address indicated at the top of this letter.

Voluntary Participation

Your participation in this research is voluntary. This means that you can withdraw your participation and consent at any given time ***before*** completing the study, without giving a reason and without facing adverse consequences. Because, in this study, research data is collected completely anonymously, responses cannot be traced back to specific individuals. This means that ***once you completed this study, it is impossible for me to know which research data belongs to you. As a consequence, once you completed this study, your research data cannot be excluded from the data set and results of this research anymore***. As a consequence, in addition to providing you with the opportunity to give or decline your consent at the beginning of this study, permission to use your research data will actively be asked for again at the end of the study at hand. If you provide permission to use your research data when being asked again at the end of this study, you have completed the study and your data cannot be excluded from the data set and research results anymore. If you decline permission to use your research data when being asked again at the end of this study, your responses will be deleted immediately and thus your research data will not be included in the data set and results of this research.

Usage of the Anonymous Data

If you give permission to use your research data when being asked again at the end of the study, the anonymous research data collected during this study will be used by me as part of my master thesis and a presentation. Me, the first reviewer of my master thesis, the second reviewer of my master thesis, the examination board and the secretary's office may have access to the anonymous data. In addition, the anonymous research data may be available to other scientists for a period of at least 10 years. All research data are safely stored following the Radboud University guidelines.

More Information, Questions, Remarks or Complaints?

If you have any questions or remarks, would like additional information about the study at hand or should you have any complaints regarding this research study, please do not hesitate to contact me using the e-mail address provided at the top of this letter.

Consent Form

Now, I ask you to take sufficient time to carefully consider whether you want to participate in the study at hand. You are of course free to decide that you do not want to participate in this study. In that case I thank you for time. If you decide that you want to participate in this study, I will ask you to indicate your consent on the following page. By doing so, you indicate, among others, that you are sufficiently informed about the study and that you voluntarily want to participate in the study.

Kind regards,
Sarah-Luisa Simon

- o I have read the information letter displayed above.

Consent Form

Consent Form for Study Participation

Hereby I confirm that

- I have been sufficiently informed about the present study and have read and fully understood the provided information letter.
- I have the opportunity to ask questions about the study as well as to provide remarks.
- I have been given sufficient time to carefully consider whether I would like to participate in the study at hand.
- I voluntarily take part in the study.

I understood that

- I can end my participation in the present study at any given point in time without having to provide reasons for my decision and without experiencing any disadvantages or adverse consequences as a result.
- all information is collected anonymously and therefore cannot be traced back to me.
- as a result of the anonymous data collection, I cannot be informed about my personal results.
- as a result of the anonymous data collection, my data cannot be excluded from the data set and results of this research anymore once I completed the study.

- o I agree.
- o I disagree.

Demographic Data

1) Are you of legal age?
    - Yes.
    - No.

2) How old are you (in years)?
    - _____

3) What is your gender?
    - Female.
    - Male.
    - Non-binary.
    - I do not want to say.

Recruitment Platform and Social Media Behaviour

1) On which social media site did you encounter the link to this study (e.g. Reddit, Twitter, WhatsApp, etc.)?
    - _____

2) Roughly how many hours per day do you spend on social media sites? Choose one option. (*Item extracted from Clark, Fine, & Scheuer, 2017*)
    - Less than 1 hour.
    - 1-2 hours.
    - 2-3 hours.
    - 3-4 hours.
    - More than 4 hours.

Study Conditions

1) Experimental: Condition: Perspective-Taking Encouragement (Yes)

Please read the displayed information and instructions carefully. To ensure adherence to the instructions, the "next" button will appear only after 60 seconds have passed.

The prevalence of hateful speech on social media platforms increases. Hateful speech means content which defames or derogates individuals or social groups for possessing certain characteristics or holding certain beliefs. It can target many different individuals and social groups and take various forms such as for instance shaming or insults.

Please take a minute to clearly and vividly visualize what people that are targeted by hateful speech may be thinking, feeling and experiencing.

2) Control: Condition: Perspective-Taking Encouragement (No)

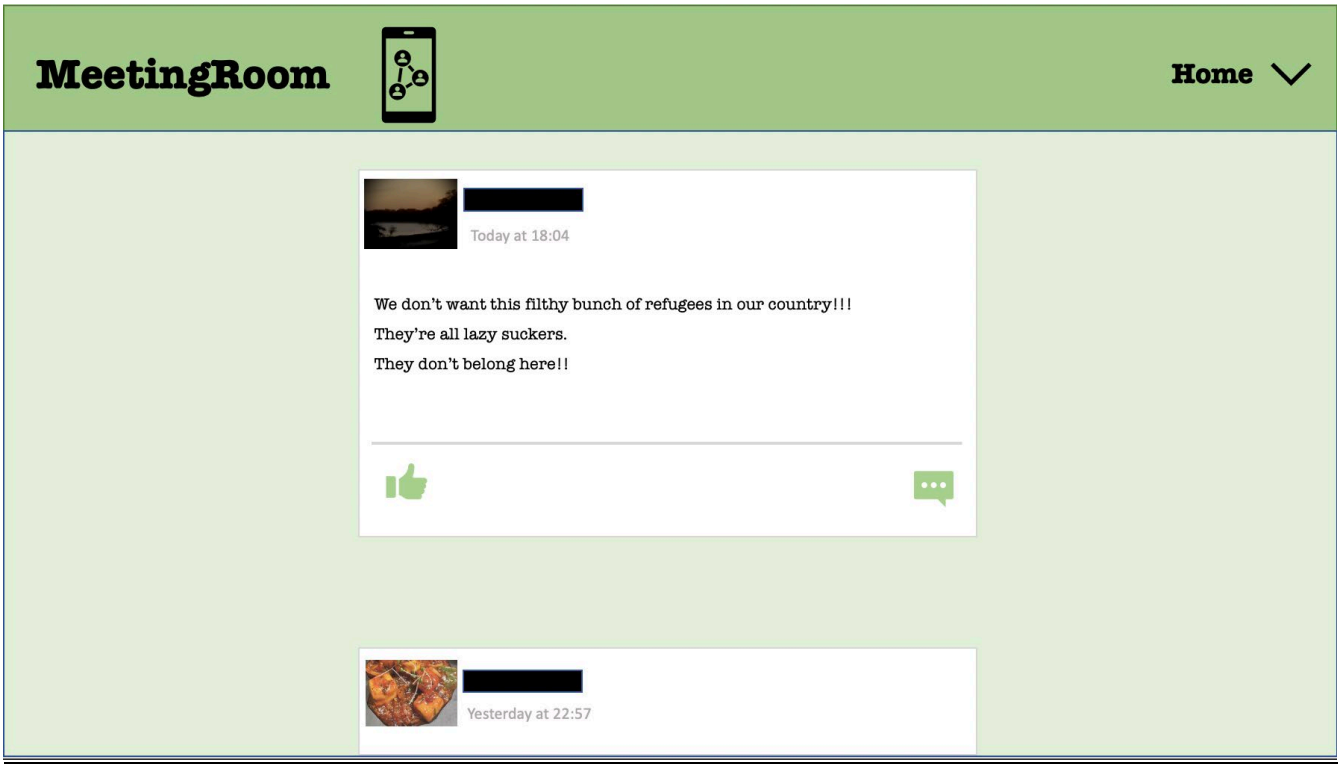
Please read the displayed information and instructions carefully. To ensure adherence to the instructions, the "next" button will appear only after 60 seconds have passed.

The prevalence of hateful speech on social media platforms increases. Hateful speech means content which defames or derogates individuals or social groups for possessing certain characteristics or holding certain beliefs. It can target many different individuals and social groups and take various forms such as for instance shaming or insults.

Please take a minute to clearly and vividly visualize all social media platforms on which, in your opinion, hateful speech may frequently be found.

Fictional Hate Speech Post

Please read the demonstrated social media post carefully as you will receive questions about it. To ensure careful reading, the "next" button will only appear after 30 seconds have passed.



Threat Perceptions (1-3) and Counter Speech Intentions (4)

The following items concern the social media post demonstrated on the previous page of this study. For each of these statements, you are asked to indicate the degree to which the statement reflects your own opinions and/ or intentions. There are no right or wrong answers. Please answer as honest and intuitive as possible.

1. The demonstrated social media post is harmful to the affected people.
    - I completely disagree.
    - I disagree.
    - I somewhat disagree.
    - I neither agree nor disagree.
    - I somewhat agree.
    - I agree.
    - I strongly agree.

2. The demonstrated social media post it threatening.
    - I completely disagree.
    - I disagree.
    - I somewhat disagree.
    - I neither agree nor disagree.
    - I somewhat agree.
    - I agree.
    - I strongly agree.

3. The demonstrated social media post has the potential to incite to violence.
    - I completely disagree.
    - I disagree.
    - I somewhat disagree.
    - I neither agree nor disagree.
    - I somewhat agree.
    - I agree.
    - I strongly agree.

4. I would comment against the demonstrated social media post.
    - Highly unlikely.
    - Unlikely.
    - Somewhat unlikely.
    - Neither likely nor unlikely.
    - Somewhat likely.
    - Likely.
    - Highly likely.

Dehumanization (1st Matrix) and Empathy (2nd Matrix)

The following items concern the people that were mentioned by the social media post that was displayed at the beginning of this study. For each of these statements, you are asked to indicate the degree to which the statement reflects your own opinions. There are no right or wrong answers. Please answer as honest and intuitive as possible.

Please rate the extent to which the following descriptions apply to the people that were mentioned by the demonstrated social media post. There are no right or wrong answers. Please answer as honest and intuitive as possible.

	Not at all.								Very much.
Lacking self-control	○	○	○	○	○	○	○	○	○
Being childish	○	○	○	○	○	○	○	○	○
Being irrational	○	○	○	○	○	○	○	○	○
Being unmannered	○	○	○	○	○	○	○	○	○
Having self-control (R)	○	○	○	○	○	○	○	○	○
Having decency (R)	○	○	○	○	○	○	○	○	○
Being polite (R)	○	○	○	○	○	○	○	○	○
Being civilized (R)	○	○	○	○	○	○	○	○	○
Being rational (R)	○	○	○	○	○	○	○	○	○
Being mature (R)	○	○	○	○	○	○	○	○	○
Being cold	○	○	○	○	○	○	○	○	○
Lacking responsiveness	○	○	○	○	○	○	○	○	○
Being passive	○	○	○	○	○	○	○	○	○
Being superficial	○	○	○	○	○	○	○	○	○
Lacking depth	○	○	○	○	○	○	○	○	○
Being sensitive (R)	○	○	○	○	○	○	○	○	○

Please indicate for the following list of emotion adjectives, the degree to which you experienced each toward the persons mentioned in the demonstrated social media post. There are no right or wrong answers. Please answer as honest and intuitive as possible.

	Not at all.					Extremely
Sympathetic	○	○	○	○	○	○
Compassionate	○	○	○	○	○	○
Warm	○	○	○	○	○	○
Softhearted	○	○	○	○	○	○
Tender	○	○	○	○	○	○
Moved	○	○	○	○	○	○

Manipulation Check

To what extent did you try to imagine what people that are targeted by hate speech might be thinking, feeling and experiencing?

Not at all.						Very much so.
○	○	○	○	○	○	○

Debriefing

Debriefing
Concerning the Study: Social Media Users' Perceptions of and Reactions to Social Media Communications

Thank you for participating in this research study. For this study, it was crucial to withhold some information from you. Now that your participation is completed, I will describe the withheld information to you. Afterwards, I will provide you with the opportunity to decide whether you would still like to have your data included in this study or want to withdraw your consent instead.

What you should know about this study:
In the informed consent letter you were informed that the aim of this research study is to examine social media users' perceptions of and reactions to social media communications. While this is true, some information was withheld from participants.

In this regard, the first aim of this research study was to investigate those factors that determine whether social media users' intent to counterargue (i.e. comment) against hateful social media content (also referred to as hate speech or cyberhate). Based on previous research, it was hypothesized that social media users' who attribute a higher extent of humanness to people that are targeted by cyberhate, meaning users' who judge hate speech targets to possess a greater extent of attributes that set them apart from animals and inanimate objects, will have higher intentions to counterargue against hate speech posts than social media users' who attribute a lower extent of humanness to these people. Furthermore, in line with previous research, it was hypothesized that this effect of humanness attributions on counter speech intentions occurs due to differences in empathy and threat perceptions. The idea was that people who attribute a higher extent of humanness to people that are targeted by cyberhate, feel more empathy for these people which then causes them to perceive hateful social media content as more threatening and harmful to these hate speech targets. Perceiving hateful social media content as more threatening and harmful to hate speech targets, in turn, was assumed to result in higher intentions to defend these hate speech targets by means of counter speech.

A second aim of this study was to investigate whether encouraging social media users to

adopt the perspective of people that are targeted by hate speech enhances the extent to which these users attribute humanness to hate speech targets and thus, via the mechanisms described above, their intentions to counterargue against cyberhate. Based on previous research, perspective-taking was assumed to be an effective strategy in this regard. This means that, based on previous research, it was hypothesized that social media users who were induced to adopt the perspective of cyberhate targets will judge people who are targeted by hateful social media content to possess a greater extent of attributes that set them apart from animals and inanimate objects (i.e. humanness) and thus, via the described mechanisms, will have higher intentions to defend these people by means of counter speech.

Fulfilling these research aims is important because of the increasing prevalence of hateful content on social media platforms, the harm hate speech poses to both, its targets and society as a whole and the fact that user-initiated counter speech has emerged as one of the most promising opportunities to regulate such hateful content on social media platforms.

In order to fulfill these research aims, the following experimental procedure was utilized:

1. Study participants were randomly assigned to either the experimental- or the control group.
2. Participants in the experimental group received an overview and example of hate speech on social media platforms and were asked to visualize what hate speech targets may be feeling, thinking and experiencing (perspective-taking induction). Participants in the control group received the same overview and example but were asked to visualize the social media platforms on which, in their opinion, hateful content may frequently be found (unrelated control-task).
3. Following a previous study, all participants received the same fictional hate speech post that targets refugees.
4. All participants filled out the same questionnaires asking about
    1. their perceptions of the threat and harm posed by the demonstrated social media post.
    2. their intentions to comment against this social media post.
    3. the extent to which they attribute humanness to the targets of the demonstrated social media post (i.e. to refugees).
    4. the extent to which they feel empathy for the targets of the demonstrated social media post (i.e. for refugees).

5. All participants filled out a question about the extent to which they tried to adopt the perspective of hate speech targets in general when being presented with the informational message about the prevalence of hateful content on social media platforms at the beginning of the study. This question is used as a manipulation check - to evaluate whether participants in the experimental group indeed engaged in perspective-taking to a higher degree than participants in the control group.
6. In the coming weeks, statistical analyses will be applied to evaluate whether the formulated hypotheses can be confirmed or have to be rejected.

Why information was withheld:

Withholding information about the exact nature of the study and the experimental manipulation was necessary to reduce response biases that may occur when participants want to help the experimenter to confirm his/ her hypotheses. Avoiding these response biases is crucial to ensure that the obtained results reflect reality as accurately as possible. Accurate reflection is, among others, important to avoid that resources are invested in implementing strategies which, in reality, are ineffective/ only marginally effective.

In case that you know other potential participants, I kindly ask you to keep the methodology used in this experimental study confidential by not informing them about the information provided in this debriefing before they have completed the study. As explained in the aforementioned paragraph, It is important for the purposes of my research that future participants are naïve to my research questions, hypotheses and exact research design.

If you have questions or remarks:

If you have any questions or remarks regarding this study, please feel free to contact me at the following e-mail address: sarah.simon@ru.nl.

Right to withdraw data:

You now have the opportunity to withdraw your permission to use the data you provided prior to this debriefing without facing any adverse consequences. If you withdraw your permission now, your data will not be included in the data set and results of this research and will be deleted immediately. Having your data excluded from the data set and results of this research at a later point in time unfortunately is not possible. This is because, in this study, all responses have been collected completely anonymously and thus cannot be traced back to

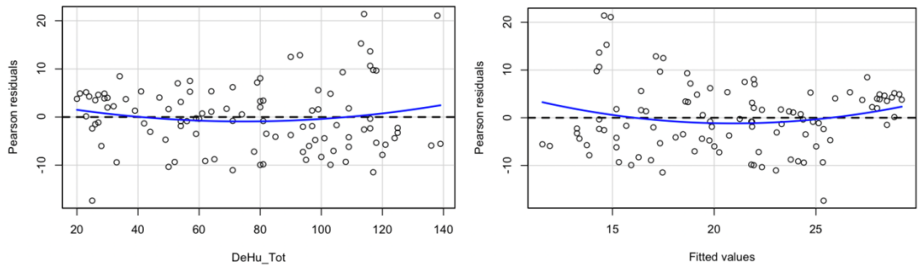
specific individuals. As a consequence, I now ask you to take sufficient time to carefully consider whether you do or do not give permission to have your data included in the study. Please indicate your decision below:

Appendix B: Assumption Testing Results

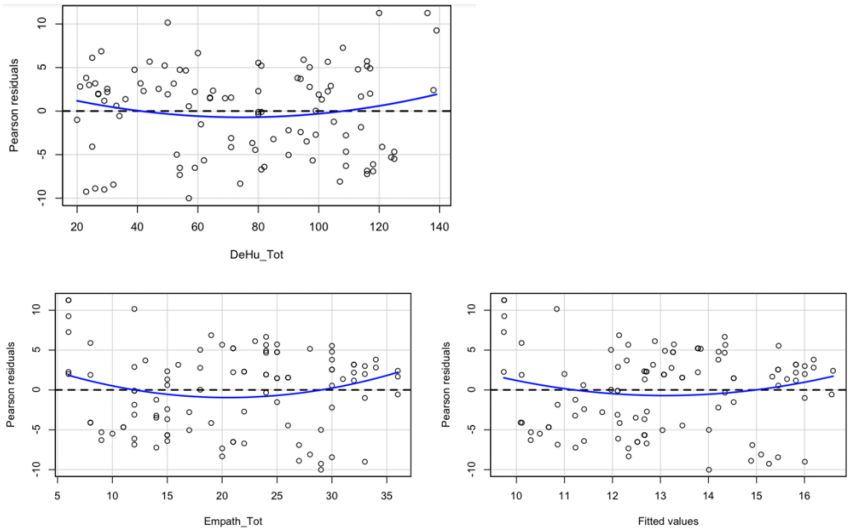
Tables and Plots for PROCESS-Performed Three-Way Mediation Model

Linearity

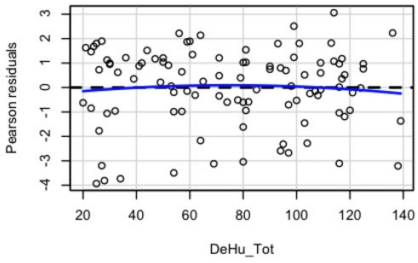
Residual Plots Equation 2

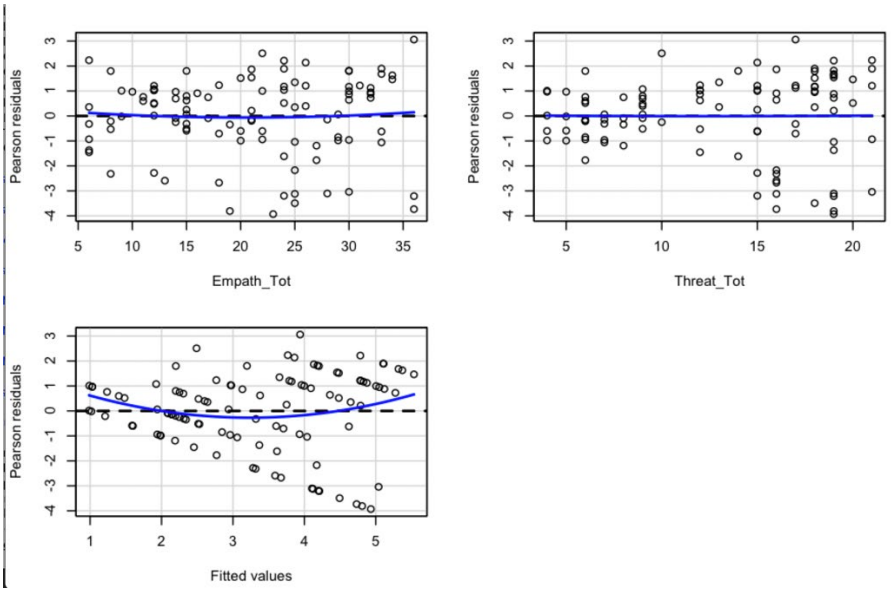


Residual Plots Equation 3



Residual Plots Equation 4





Lack-of-Fit Tests

	Lack-of-Fit Tests		
	Statistic (= *t*)	Sig. (= *p*)	Significant?
Equation 2 (Empathy ~ Condition + Dehumanization)			
Condition	N.A.	N.A.	N.A.
Dehumanization	1.162	.248	No
Equation 3 (Threat Perceptions ~ Condition + Dehumanization + Empathy)			
Condition	N.A.	N.A.	N.A.
Dehumanization	1.270	.207	No
Empathy	1.828	.071	No
Equation 4 (Counter Speech Intentions ~ Condition + Dehumanization + Empathy + Threat Perceptions)			
Condition	N.A.	N.A.	N.A.
Dehumanization	-.052	.607	No
Empathy	.388	.699	No
Threat Perceptions	.048	.962	No

Multicollinearity

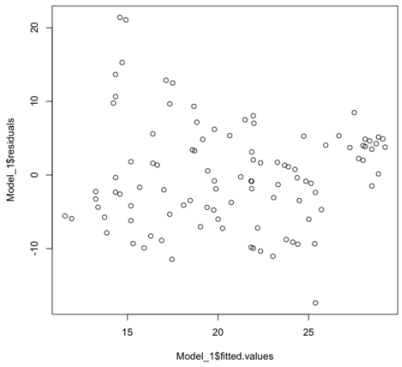
	VIF	Tolerance
Equation 2 (Empathy ~ Condition + Dehumanization)		
Condition	1.108	.903
Dehumanization	1.108	.903
Equation 3 (Threat Perceptions ~ Condition + Dehumanization + Empathy)		
Condition	1.161	.861
Dehumanization	1.448	.691
Empathy	1.478	.677
Equation 4 (Counter Speech Intentions ~ Condition + Dehumanization + Empathy + Threat Perceptions)		
Condition	1.177	.850
Dehumanization	1.448	.691
Empathy	1.577	.634
Threat Perceptions	1.141	.876

Normality of the Residuals

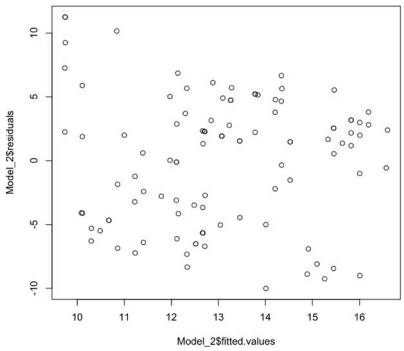
	Shapiro-Wilk			
	Statistic (= *W*)	df	Sig. (= *p*)	Significant?
Equation 2 (Empathy ~ Condition + Dehumanization)	.980	102	.133	No
Equation 3 (Threat Perceptions ~ Condition + Dehumanization + Empathy)	.968	102	.014	Yes
Equation 4 (Counter Speech Intentions ~ Condition + Dehumanization + Empathy + Threat Perceptions)	.943	102	< .001	Yes

Homoscedasticity of the Residuals

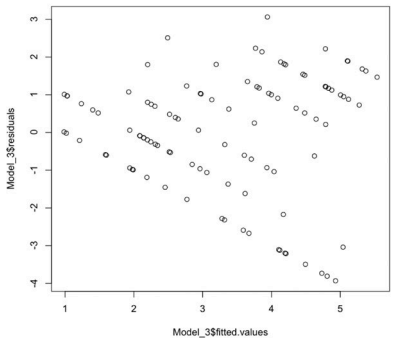
Residual Plot Equation 2



Residual Plot Equation 3



Residual Plot Equation 4



Appendix C: Syntax for PROCESS-Performed Three-Way Mediation Model

process y=Cspeech / x=Condt/ m=DeHu_Tot Emph_Tot Thr_Tot/ model=6/ effsize=1/ hc=3/ total=1/ seed=4421/ modelbt=1