Defend Your Enemy. A Qualitative Study on Defending Political Opponents Against Hate Speech Online

Lilian Kojan[1], Hava Melike Osmanbeyoglu[2], Laura Burbach[1], Martina Ziefle[1], and André Calero Valdez[1](✉)

[1] Human-Computer Interaction Center, RWTH Aachen University, Campus Boulevard 57, 52074 Aachen, Germany
{kojan,burbach,ziefle,calero-valdez}@comm.rwth-aachen.de

[2] RWTH Aachen University, Templergraben 44, 52052 Aachen, Germany
melike.osmanbeyoglu@rwth-aachen.de

Abstract. Both hate speech and disinformation negatively influence the internet's potential for public deliberation and lead to polarization between political groups. In this paper, we examine the potential of counter speech to bolster public deliberation and reduce polarization. In two focus groups, we interview participants on what motivates them to engage in counter speech in general as well as counter speech favoring political adversaries. Firstly, we find a sharp distinction between participants who avoid engaging with hate speech and participants who actively engage with hate speech in order to combat it. Thus, the most important predictor for counter speech favoring adversaries is an individual's propensity for counter speech in general. In turn, motivations for counter speech in general are a strong sense of morality, a perception of the internet as an important space for public deliberation, and a sense of responsibility to enforce rules for a fair debate. Many of those participants view their online activitiy as a form of activism. Additionally, individuals engaging in counter speech hope to positively influence not necessarily the hater, but the broader audience.

Keywords: Hate speech · Counter speech · Social media · Political deliberation

1 Deliberation in Digital Media

Since the commercialization of the Internet, the relationship between digital media and political life has grown ever stronger [8]. Apart from election campaigns, one important facet is the Internet's potential to strengthen democratic society by facilitating public deliberation [14,39]. At the same time, several limitations for online public deliberation have emerged. For one, offline power imbalances are often mirrored in the online world through an overrepresentation of groups in power, e.g., well-educated white men [21]. In addition, online groups

M. van Duijn et al. (Eds.): MISDOOM 2020, LNCS 12259, pp. 80–94, 2020.
https://doi.org/10.1007/978-3-030-61841-4_6

tend to be very homogenous meaning that users are seldomly exposed to cross-cutting opinions or differing viewpoints [16]. And when differing opinions do collide, incivility and even hate speech can occur [15,38,40].

1.1 Hate Speech and Misinformation

Although hate speech has been extensively discussed by the public at large as well as studied in academia, finding a universally valid definition is challenging [22]. Legal institutions and social networks alike tend to provide broad definitions that allow for judgement and possible sanctions on a case-by-case basis [22].

From a communication science perspective, Erjavec and Kovačič [20] define hate speech as an expression that is in itself harmful or possibly harm-inciting and targets members of a group determined by characteristics like race or sexual orientation. Similar characteristics can be found in other definitions that consider the purpose and the effects of hate speech. Waldron [44] characterizes speech as hateful when it serves one or both of two functions: Firstly, to dehumanize a target group and diminish its members and secondly, to reinforce a sense of in-group with other like-minded individuals. Similarly, Susan Benesch has coined the term *dangerous speech* which she defines as "[a]ny form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or commit violence against members of another group". [5] Often, the groups targeted are marginalized social groups [1]. But especially in common parlance, hate speech can also describe speech directed at groups like politicians that are arguably powerful [22]. In summary, hate speech both reenforces the boundaries between groups and is harmful to members of the other group, either in itself or in its effects.

Hate speech is inextricably linked to disinformation. For one, online hate often takes the form of over-generalization, exaggeration or even outright deceit about the targeted group. E.g., Awan [3] stresses the use of false stories to exacerbate islamophobic hate. For another, disinformation like fake news often serve the same purpose as hate speech: Polarization, radicalization and othering of the out-group [6].

There is ample evidence for the damaging effects of hate speech, not only on the victim of the hate speech but also on the broader audience. Constant exposition shapes the user's worldview and influences their decision-making [19,27]. Reading hateful and uncivil content increases attitude polarization [7,29]. And by inducing negative emotions, it can also discourage people from engaging in discourse [25,26,29,35]. Thereby, it actively impedes on the Internet's potential for public deliberation.

One possible way to counter hate speech is counter speech. Counter speech can be defined as a dissenting response to hate speech [48]. Although it is sometimes used in a way that also encompasses actions like flagging hateful content, our study focusses on counter speech in the form of content, e.g., comments in answer to the hateful content itself.

1.2 Effectiveness of Counter Speech

Counter speaking is encouraged in many anti-hate speech programs [22]. Furthermore, Chen [14] argues that countering online incivility is necessary to realize the potential of online spaces as a place for political deliberation. In spite of that, only a few studies have actually evaluated the effectiveness of counter speech. Buerger and Wright [11] have reviewed the studies available in November 2019. They differentiate between the effects of counter speech on the hateful speaker and the effects on the wider audience.

The results concerning the effects of counter speech on the hateful speaker are inconclusive [11]. However, there is some evidence that counter speech by users that are perceived as more influential, can curb hate speech at least temporarily (e.g., [36]). Findings on the effects of counter speech on the wider audience are less ambiguous. They all find evidence for something that Buerger and Wright [11] call the "contagion effect", i.e., the presence of hateful comments increases the probability of a user also making a hateful comment. On the opposite hand, civil comments also lead to more civil comments. Moreover, meta-comments urging people to be civil promote further meta-comments about discussion quality [35].

So while further research on the effects of counter speech is desirable, the existing indications for its success prompt us to ask what predicts users engaging in counter speech.

1.3 Predictors for Counter Speech

There already exist several studies examining willingness to intervene against hate speech and incivility in general and even more studies from the field of cyber-bystander research. As bystander intervention in cyber-bullying is similar to counter speech, predictors from a review on cyber-bystanding studies by Lambe et al. [30] are included as well.

The following predictors refer to intervention intention, with intervention ranging from more distanced behavior like using the reporting function (e.g., [47]) to deeply involved behavior like verbally confronting the individual engaging in hate speech (e.g., [18]).

The factors we summarize as *individual factors* concern properties of a would-be counter speaker that make intervention more likely. The predictors found are female gender [30,46], high prosociality and empathy [30], high self-efficacy [30], a negative attitude towards passive bystanding [30], an expectation that defending will help [30], and a high importance of morality, including low moral disengagement, high moral identity scores and individualizing moral foundation [30,46]. Additionally, there are situation-dependent factors like the would-be counter speaker feeling negative affect [14,17,18] and them perceiving social pressure and responsibility to intervene [17,18,47].

Other factors relate to the *properties of the victim* of the hate speech or bullying. Users are more likely to intervene if the victim is an individual person as a victim rather than an abstract social group [37], if the victim is more popular

[30], if they have a friendship or positive relationship with the victim [30] and if they exhibit a low level of prejudice towards the victim's social group [17].

Concerning *situational factors*, users were more likely to intervene if the situation was more deviant [17,18,37], if there was more than one perpetrator [28] and if more steps of the bystander intervention model were met (situation is noticed, fewer number of bystanders, information on how to confront is provided) [30,37].

To summarize, the existing research on hate speech intervention mainly considers the properties of the would-be counter speaker. For this study, we wanted to further the research on hate speech intervention by focussing in on the relationships between the would-be counter speaker and the victim. To be precise, with hate speech as an instrument for social division and polarization, can counter speech bridge the gap between in- and out-group? Therefore, our research question is:

In social media discussions, what are predictors for users to engage in counter speech in support of political adversaries?

2 Method

As laid out in Sect. 1.3, there is some research on predictors for counter speech in general as well as a breadth of studies in the field of cyber-bystander research. To our knowledge, however, there have not been any studies on out-group favoring counter speech. Therefore, an exploratory study design was chosen. Data was gathered in two focus groups. Afterwards, the data was transcribed and analysed to find the most pertinent predictors. The full transcriptions and the full analysis as well as the questionnaire, the slides and the guide used to collect our data can be found in our github repository for this project.[1]

2.1 Focus Groups

We conducted two focus groups, asking participants about their experiences with online hate speech in general and their own reactions to hate speech in particular, i.e., if they engaged in in counter speech at all. Special emphasis was placed on counter speech on behalf of political adversaries, that is, people the participants considered to be their opponents in an online discussion.

Guide and Structure. The focus groups were conducted using a guide which was pre-tested in advance. The sequence was structured into four sections, each concerned with one main topic:

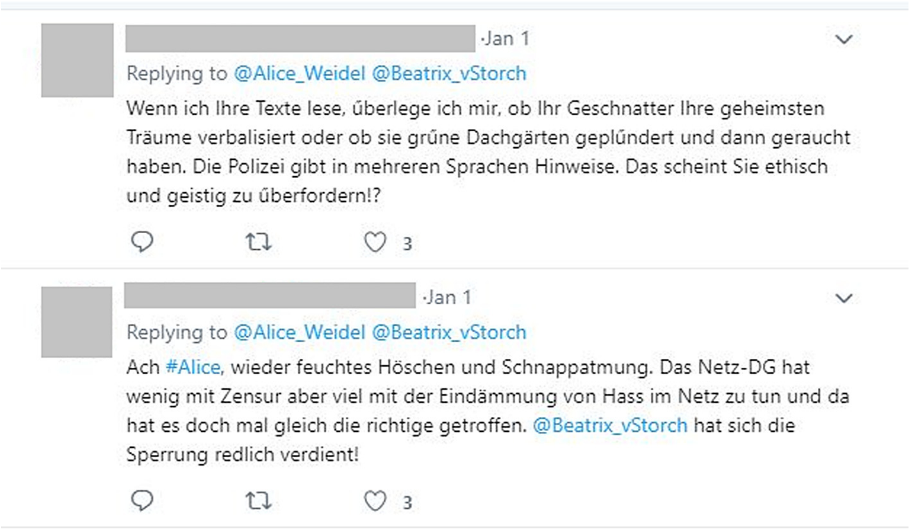
1. Own experiences with hate speech.
2. Engagement in counter speech.
3. Conditions for counter speech for political adversaries.

[1] The repo can be found here: (github.com/digitalemuendigkeit/misdoom2020).

4. Motivations for counter speech for political adversaries.

In this context, conditions referred to predictors for counter speech in specific situations (i.e, when will you engage in counter speech) and motivations to general predictors (i.e., why do you engage in counter speech).

Stimuli. During the focus groups, we used screenshots of pertinent online interactions as stimuli, see e.g., Fig. 1. As the research question aimed at counter speech on behalf of one's political adversary, we aimed to select stimuli in a way that different political affiliations were accounted for. Therefore, we chose online interactions and tweets involving Alice Weidel, a member of the German right-wing party AfD, as well as posts aimed at one politician of the Greens, Claudia Roth. Not only can the Greens and the AfD be described as being representative of two ends of the political spectrum [24,31,33]. Also, both the AfD and the Greens, especially Claudia Roth, could be considered highly polarizing in November and December of 2018 when the focus groups were conducted [23,43].



[redacted] [redacted] ·Jan 1

Replying to @Alice_Weidel @Beatrix_vStorch

Wenn ich Ihre Texte lese, überlege ich mir, ob Ihr Geschnatter Ihre geheimsten Träume verbalisiert oder ob sie grüne Dachgärten geplündert und dann geraucht haben. Die Polizei gibt in mehreren Sprachen Hinweise. Das scheint Sie ethisch und geistig zu überfordern!?

♡ 3

[redacted] [redacted] ·Jan 1

Replying to @Alice_Weidel @Beatrix_vStorch

Ach #Alice, wieder feuchtes Höschen und Schnappatmung. Das Netz-DG hat wenig mit Zensur aber viel mit der Eindämmung von Hass im Netz zu tun und da hat es doch mal gleich die richtige getroffen. @Beatrix_vStorch hat sich die Sperrung redlich verdient!

♡ 3

Fig. 1. Stimulus B (insulting replies to one of German politician's Alice Weidel's tweets)

Recording and Transcription. Each focus group was recorded on audio. We then transcribed the recordings using MAXQDA, employing a modified version of GAT 2 as the transcription system [42].

2.2 Participants

The participants were recruited through convenience sampling. Based on preliminary questioning, they were sorted into two homogenous groups, the *moderately active* group and the *very active* group, in order to obtain more detailed results [41]. Potential participants who reported only passive social media use or none at all were excluded. The *moderately active* group (n = 5) included participants who mostly consumed social media but only seldomly posted or commented. Participants who not only used but also commented and posted in social media became part of the *very active group* (n = 6).

Before starting the focus group, each participant was surveyed on demographic details as well as the frequency of their social media use in general, the frequency of them posting and commenting online, and their political left-right self-placement [10]. The results are displayed in table 1.

Table 1. Focus group participants

	Moderately active group (n = 5)	Very active group (n = 6)
Gender	Female: 2, Male: 3	Female: 3, Male: 3
Age	M = 26.6, SD = 4	M = 32.3, SD = 7
Highest Level of Education	Abitur[a]: 1, University Degree: 4	Abitur[a]: 2, University Degree: 4
Occupation	Student: 3, Full-Time Employed: 2	Student: 1, Full-Time Employed: 5
Frequency of Social Media Use[bc]	M = 3.5, SD = 0.5	M = 4.2, SD = 0.4
Frequency of Posting and Commenting Online[b]	M = 2.2, SD = 0.8	M = 5.5, SD = 0.8
Political Left-Right Self-Placement[d]	M = 4, SD = 1.2	M = 3.5, SD = 1.6

[a] General Higher Education Entrance Qualification;
[b] 1 = never, 2 = very rarely, 3 = several times a month, 4 = several times a week, 5 = daily, 6 = several times a day;
[c] averaged over 6 types of platforms (social networking sites, video platforms, blogs, online newspapers, infotainment, social news);
[d] 1 = left, 10 = right

The participants from the *very active* group score somewhat higher on average social media use frequency and much higher on the posting and commenting frequency. Therefore, the classification based on the preliminary questioning was proven valid.

2.3 Content Analysis

We conducted a qualitative content analysis as described by Mayring [34] using MAXQDA. After we first developed a categorization, we tested for intercoder reliability by calculating coefficient kappa using the approach of Brennan and Prediger [9] (minimum coding overlap = 60%). Overall, a kappa of 0.26 was calculated. This proposes unsatisfactory reliability which is, however, not out of the ordinary for the first iteration of intercoder reliability examination. [13,32] To resolve the discrepancies between the different coders, we employed the Intercoder Agreement method as described by Campbell et al. [13]. In Fig. 2, an overview of the final categorization is visible.

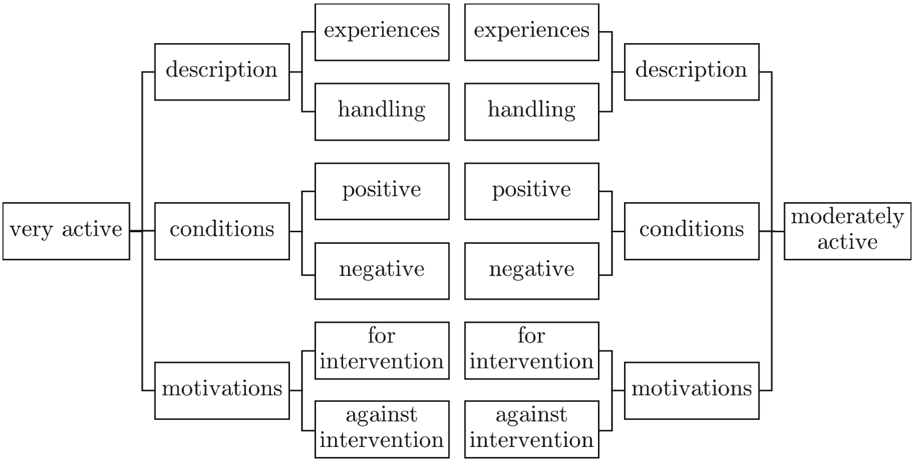


Fig. 2. Overview of the final categorization

3 Results

As Fig. 2 shows, we contrasted the results of both groups for each category, (*description*, *conditions* and *motivations*). As our research question focusses on predictors, only a quick overview will be given for the category *description*.

For the sake of brevity, in the following sections these terms will be used: *Hater:* The perpetrator of the hate speech, *victim(s):* the recipient(s) or subject(s) of the hate speech, and *adversary victim(s):* victim(s) or subject(s) of hate speech that represent a group the participant politically or personally opposes.

3.1 Description: Experiences with Hate Speech and Handling Hate Speech

When describing their experiences with hate speech, both groups mention similar attributes of hate speech (e.g., the online *spaces* where they have most often observed it). However, while four of the six very active participants report to have themselves been victims of hate speech, only one moderately active participant does so as well. There are also notable differences in the way participants of both groups handle hate speech: While participants in the moderately active group tend to look at hateful comments only for entertainment value or avoid looking at comments at all, many participants in the very active group actively seek out hate speech comments in order to fight it.

3.2 Conditions: When to Engage in Counter Speech

For a given situation, participants describe both conditions that make it more likely that they engage in counter speech (*positive conditions*) as well as conditions that make it less likely (*negative conditions*). As mentioned above, the participants in the moderately active group reported to only seldom engage in counter speech at all, much less counter speech favoring *adversary victims*, i.e., politically opposed users that are targetted by hate speech. Therefore, most of the conditions listed are to be understood as conditions for counter speech in general. The only exception is the subcategory *positive conditions* for the very active group where a differentiation between counter speech in general and counter speech favoring *adversary victims* was possible.

Positive Conditions for Counter Speech Favoring Adversaries. When it comes defending people who they are politically opposed to, the following conditions emerged in the very active group: **1) Offenses against a "culture of discussion"**, i.e., the participant feels that the hater breaks the rules for a respectful debate, **2) offenses against the human dignity**, i.e., the participant feels that the hater debases the victim's human dignity, **3) properties of the victim**, e.g., the participants feels sympathy for the victim, and **4) a personal connection to the topic discussed**.

Positive Conditions for Counter Speech in General.

Properties of the Hate Speech. Concerning the properties of the hate speech or the situation where the hate speech occurs, participants of both groups mention they are more likely to step in when they feel that their **1) counter speech is likely to have an impact**, e.g., there are not that many comments overall. Additionally, participants of the very active group name the space as an important factor. They are more likely to step in when there is hate speech **2) outside of hater-dominated spaces**, e.g., not in a dedicated facebook group, or **3) in a more private space**, e.g., in a personal chat group.

On the other side, participants of the moderately active group mention **4) calls to violence and threats** as well as **5) doxxing**, i.e., finding and disseminating the victim's personal information, as factors making counter speech more likely for them.

Properties of the Victim. Relating to the properties of the victim, both groups mention that they would be more likely to intervene if **1) the victim is a private citizen** or if **2) they know the victim personally**, although that is more important in the moderately active group. Additionally, participants of the very active group would be more likely to engage in counter speech if **3) the victim is an activist**.

Personal Attributes. Participants of both group state that they are more likely to step in if they are **1) well informed about the topic of discussion**. Members of the very active group also name **2) free time and mental energy** as a condition. Some members of the moderately active group, on the other hand, describe **3) feeling frustrated and angry** or **4) having a personal connection** to the topic of discussion as a conductor for counter˜speech.

Negative Conditions for Counter Speech in General. Many of the *negative conditions* mentioned in the group are merely negations of the *positive conditions* already listed and will therefore not be reported again.

Properties of the Hate Speech. Members of both group state that they are less likely to engage in counter speech, when **1) both sides of the discussion engage in hate speech** or when **2) the hate speech is entertaining to them**.

Additionally, participants of the very active group are less willing to intervene when they fear **3) personal risk to themselves**.

Properties of the Victim. Apart from the negatives to the *positive conditions* already mentioned, one member of the very active group reports that they would be less likely to intervene if they suspect **the victim is eager to be seen as a victim by the public**.

3.3 Motivations: Why to Engage in Counter Speech

Just as with the conditions, the *motivations* of the participants could also be categorized into *motivations for intervention* and *motivations against intervention.* In this context, *motivations* relate to the participants' attitudes towards counter speech in general. By contrast, the *conditions* listed above relate to specific situations.

Motivations for Intervention. The *motivations for intervention* can be further categorized into *goals and values* and *personal attributes and experiences.*

Goals and Values. Participants of both groups concede that while they might not be able to dissuade the hater from their destructive behavior, they still **1) hope to positively influence the audience**. Members of the very active group are additionally motivated by the desire to **2) fight disinformation**, **3)**

motivate critical thinking in other users, **4) create a better culture of discussion in online spaces** and **5) engage politically**. Some of the very active participants describe viewing their counter speech activity as a form of activism.

Moderately active participants, on the other hand, worry that *online hate might spark offline violence.*

Personal Attributes and Experiences. When it comes to their personal attributes and experiences that motivate them to engage in counter speech, participants of both groups mention **1) a strong sense of justice** and **2) a sense of responsibility**.

In addition, very active participants name **1) enjoying debating**, **2) their own experiences with bullying and discrimination**, **3) enjoying self-promotion**, as well as **4) being thanked and admired by others**, e.g., by site administrators, as motivators.

Motivations Against Intervention. Members of both groups name one main motivation not to engage in counter speech: They think it is **1) not worth the effort**. On top of that, members of the very active group mainly mention **2) fatigue** with fighting hate speech in general as something that demotivates them from engaging in counter speech.

Among the moderately active participants, a considerable number more motivations are named: **1) A general unwillingness to participate in online communication**, **2) a preference for alternative approaches to hate speech**, e.g. blocking the perpetrator or even reporting them to the police, **3) their perception of the chance to be successful as too small** and **4) their own tendency to avoid reading comments at all**.

4 Discussion and Conclusion

Notably, a vast difference in engagement levels between the participants was found. While some would not engage in online discourse at all and consequently would not engage in counter speech either, others were hyperactive on social media, placing a lot of value on political discourse in online spaces. Participants in the latter group reported a much higher likelihood to engage in counter speech, be it on behalf of opponents or in general.

Overall, the relationship between the counter speaker and the victim which we focussed on in our research question (*In social media discussions, what motivates users to engage in counter speech in support of political adversaries?*) seems to be less important than the willingness to engage in counter speech in general. While we collected and categorized the predictors for counter speech reported by our participants in *conditions*, i.e., situational predictors, and *motivations*, i.e., general predictors, there is only a small number of predictors from the subset *conditions* strictly in answer to our research question:

Three basic motivations for users to engage in counter speech in support of political adversaries can be differentiated: Firstly, the hater violates norms or values that are more important to the counter speaker than political affiliation. Values named here were a "culture of discussion", i.e., an implicit set of rules for a respectful debate, and "human dignity". These conditions tie somewhat into other findings about counter speakers placing high importance on morality [30, 46]. Interestingly, they also mirror the vision of the internet as a place for public deliberation [14]. Productive debates can only happen if all participants follow the rules, no matter which side they are on. Secondly, counter speech is more likely if the participants feels sympathetic towards the victim. This is similar to results from cyber-bullying research [30]. Thirdly, participants are more likely to intervene when they feel a personal connection to the topic of discussion. Both the second and the third motivation are limited in their generalizability. Sympathies are likely to wane the larger the distance on the political spectrum gets. And in many occurrences of hate speech, there will be no connection to a tangible discussion topic.

The other predictors we found refer to counter speech in general and therefore do not strictly answer the research question. However, as posited above, we did not observe the expected divide between people engaging in counter speech only for friends or members of their in-group and people engaging in counter speech for everyone—including adversaries. Rather, the divide was between people engaging in counter speech for everyone, regardless of political or group affiliation, and people not generally engaging in counter speech. As such, we feel that the *motivations* of the very active group questioned also partly answer the question of what motivates counter speakers.

The most important *motivations* we found were deep-seated moral convictions and a feeling of responsibility to uphold those convictions. This does not only match the findings by on the importance of morality by Wilhelm and Joeckel [46] and Lambe et al. [30]. The acceptance of responsibility also matches the bystander model of intervention often used to describe bystander behavior in cyber-bullying incidents (e.g., [37]). Moreover, the participants felt that online discourse is an important part of political participation [14]. Many of the active counter speakers we talked to saw their actions as a form of activism. One of their major goals was not to change the behavior of the people engaging in hate speech, but to positively influence the broader audience. This matches what Buerger and Wright [11] call the contagion effect.

In conclusion, when looking at what motivates a person to regularly engage in counter speech, their relationship to the victim appears to be secondary. Of greater importance seems to be what part morality plays in that person's self-image and how willing they are to accept and defend online spaces as a place for public deliberation.

Finally, some limitations have to be noted: Firstly, although we tried to emphasize the relationship aspect (i.e., counter speech *in favor of adversaries*) in our research design, stressing this emphasis during the focus groups proved challenging. Rather, participants tended to talk about their experiences with counter speech *in general*. This holds especially true for the participants of the moderately active group, many of whom never had engaged in counter speech at all. Therefore, our results are not suitable to evaluate whether there are differences between predictors for counter speech *in favor of adversaries* and counter speech *in general*. Secondly, our sample was comparatively young, highly educated and politically left-aligned. It is entirely possible that other predictors not mentioned here are important with counter speakers who are, e.g., more politically right-leaning. In any case, the predictors identified in this study should be further tested in a quantitative study. Thirdly, the predictors identified in this study as well as most other studies listed in Sect. 1.3 are self-reported. Conducting an experiment could shed light on whether or not these translate to actual defending behavior.

Acknowledgements. We would like to thank Maximilian Geulen and Merten Wothge for their help in developing the research design as well as collecting the data. This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia. We would further like to thank the authors of the packages we have used. We used the following packages to create this document: `knitr` [49], `tidyverse` [45], `rmdformats` [4], `kableExtra` [50], `rmdtemplates` [12], `citr` [2].

References

1. Álvarez-Benjumea, A., Winter, F.: Normative change and culture of hate: an experiment in online environments. Eur. Sociol. Rev. **34**(3), 223–237 (2018). https://doi.org/10.1093/esr/jcy005. ISSN: 0266–7215, 1468–2672
2. Aust, F.: citr: RStudio add-in to insert markdown citations. R package version 0.3.2. (2019). https://CRAN.R-project.org/package=citr
3. Awan, I.: Islamophobia on social media: a qualitative analysis of the Facebook'S walls of hate (2016). https://doi.org/10.5281/ZENODO.58517
4. Barnier, J.: rmdformats: HTML output formats and templates for 'rmarkdown' documents. R package version 0.3.6. (2019). https://CRAN.R-project.org/package=rmdformats
5. Benesch, S., et al.: Dangerous speech: a practical guide (2020). https://dangerousspeech.org/guide/
6. Bennett, W.L., Livingston, S.: The disinformation order: disruptive communication and the decline of democratic institutions. Eur. J. Commun. **33**(2), 122–139 (2018). https://doi.org/10.1177/0267323118760317. ISSN: 0267-3231, 1460-3705
7. Borah, P.: Does it matter where you read the news story? interaction of incivility and news frames in the political blogosphere. Commun. Res. **41**(6), 809–827 (2014). https://doi.org/10.1177/0093650212449353. ISSN: 0093–6502, 1552–3810
8. Boulianne, S.: Twenty years of digital media effects on civic and political participation. Commun. Res. (2018). ISSN: 0093–6502, 1552–3810. https://doi.org/10.1177/0093650218808186
9. Brennan, R.L., Prediger, D.J.: Coefficient kappa: some uses, misuses, and alternatives. Educ. Psychol. Meas. **41**(3), 687–699 (1981)
10. Breyer, B.: Left-right self-placement (allbus) (2015)
11. Buerger, C., Wright, L.: Counterspeech: a literature review (2019)

12. Valdez, A.C.: rmdtemplates: rmdtemplates – an opinionated collection of rmarkdown templates. R package version 0.3.3.0001 (2019). https://github.com/statisticsforsocialscience/rmd_templates
13. Campbell, J.L., et al.: Coding in-depth semistructured interviews: problems of unitization and intercoder reliability and agreement. Sociol. Methods Res. **42**(3), 294–320 (2013). https://doi.org/10.1177/0049124113500475. ISSN: 0049–1241, 1552-8294
14. Chen, G.M.: Online Incivility and Public Debate. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56273-5
15. Coe, K., Kenski, K., Rains, S.A.: Online and uncivil? patterns and determinants of incivility in newspaper website comments. J. Commun. **64**(4), 658–679 (2014). https://doi.org/10.1111/jcom.12104. ISSN: 00219916
16. Colleoni, E., Rozza, A., Arvidsson, A.: Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data: political homophily on twitter. J. Commun. **64**(2), 317–332 (2014). https://doi.org/10.1111/jcom.12084. ISSN: 00219916
17. Dickter, C.L.: Confronting hate: heterosexuals' responses to anti-gay comments. J. Homosex. **59**(8), 1113–1130 (2012). https://doi.org/10.1080/00918369.2012.712817. ISSN: 0091–8369, 1540–3602
18. Dickter, C.L., Newton, V.A.: To confront or not to confront: non-targets' evaluations of and responses to racist comments: responses to racist comments. J. Appl. Soc. Psychol. **43**, E262–E275 (2013). https://doi.org/10.1111/jasp.12022. ISSN: 00219029
19. Erjavec, K.: Readers of online news comments: why do they read hate speech comments? Ann. Series historia et sociologia **24**(3), 451–462 (2014)
20. Erjavec, K., Kovačič, M.P.: You don't understand, this is a new war!' analysis of hate speech in news web sites' comments. Mass Commun. Soc. **15**(6), 899–920 (2012). https://doi.org/10.1080/15205436.2011.619679. ISSN: 1520-5436, 1532-7825
21. Feezell, J.T.: Predicting online political participation: the importance of selection bias and selective exposure in the online setting. Polit. Res. Q. **69**(3), 495–509 (2016). https://doi.org/10.1177/1065912916652503. ISSN: 1065–9129, 1938-274X
22. Gagliardone, I., et al.: Countering online hate speech. Technical report, UNESCO (2015)
23. Gensing, P.: Grünen-Politikerin Roth: Im Visier des Hasses. de (2018). http://www.tagesschau.dehttp://faktenfinder.tagesschau.de/inland/kampagnen-roth-101.html
24. Hambauer, V., Mays, A.: Wer wählt die AfD? – Ein Vergleich der Sozialstruktur, politischen Einstellungen und Einstellungen zu Flüchtlingen zwischen AfD-WählerInnen und der WählerInnen der anderen Parteien de. Z. Vgl. Polit. Wiss **12**(1), 133–154 (2018). https://doi.org/10.1007/s12286-017-0369-2. ISSN: 1865-2646, 1865-2654
25. Hwang, H., Kim, Y., Huh, C.U.: Seeing is believing: effects of uncivil online debate on political polarization and expectations of deliberation. J. Broadcast. Electron. Media **58**(4), 621–633 (2014). https://doi.org/10.1080/08838151.2014.966365. ISSN: 0883-8151, 1550-6878
26. Hwang, H., et al.: Does civility matter in the Blogo– sphere? examining the interaction effects of incivility and disagreement on citizen attitudes. In: Annual Convention of the International Communication Association Montreal, Canada (2008)

27. Jubany, O.: Backgrounds, experiences and responses to online hate speech: an ethnographic multi-sited analysis. In: 2nd Annual International Conference on Social Science and Contemporary Humanity Development (SSCHD 2016). Atlantis Press(2016). ISBN: 978-94-6252-227-5. https://doi.org/10.2991/sschd-16.2016.143
28. Kazerooni, F., et al.: Cyberbullying bystander intervention: the number of offenders and retweeting predict likelihood of helping a cyberbullying victim. J. Comput. Mediated Commun. **23**(3), 146–162 (2018). https://doi.org/10.1093/jcmc/zmy005. ISSN: 1083-6101
29. Kim, Y., Kim, Y.: Incivility on facebook and political polarization: the mediating role of seeking further comments and negative emotion. Comput. Hum. Behav. **99**, 219–227 (2019). https://doi.org/10.1016/j.chb.2019.05.022. ISSN: 07475632
30. Lambe, L.J., et al.: Standing up to bullying: a social ecological review of peer defending in offline and online contexts. Aggression Violent Behav. **45**, 51–74 (2019)
31. Lees, C.: The 'alternative for Germany': the rise of right-wing populism at the heart of Europe. Politics **38**(3), 295–310 (2010). https://doi.org/10.1177/0263395718777718. ISSN: 0263-3957, 1467-9256
32. MacPhail, C., et al.: Process guidelines for establishing intercoder reliability in qualitative studies. Qual. Res. **16**(2), 198–212 (2016). https://doi.org/10.1177/1468794115577012. ISSN: 1468-7941, 1741-3109
33. Mader, M., Schoen, H.: The European refugee crisis, party competition, and voters' responses in Germany. West Eur. Polit. **42**(1), 67–90 (2019). https://doi.org/10.1080/01402382.2018.1490484. ISSN: 0140-2382, 1743-9655
34. Mayring, P.: Qualitative content analysis: theoretical foundation, basic procedures and software solution. Klagenfurt (2014)
35. Molina, R.G., Jennings, F.J.: The role of civility and metacommunication in facebook discussions. Commun. Stud. **69**(1), 42–66 (2018). https://doi.org/10.1080/10510974.2017.1397038. ISSN: 1051-0974, 1745-1035
36. Munger, K.: Tweetment effects on the tweeted: experimentally reducing racist harassment. Polit. Behav. **39**(3), 629–649 (2017). https://doi.org/10.1007/s11109-016-9373-5. ISSN 0190-9320, 1573-6687
37. Naab, T.K., Kalch, A., Meitz, T.G.K.: Flagging uncivil user comments: effects of intervention information, type of victim, and response comments on bystander behavior. New Media Soc. **20**(2), 777–795 (2018). https://doi.org/10.1177/1461444816670923. ISSN: 1461-4448, 1461-7315
38. Papacharissi, Z.: Democracy online: civility, politeness, and the democratic potential of online political discussion groups. New Media Soc. **6**(2), 259–283 (2004). https://doi.org/10.1177/1461444804041444. ISSN: 1461-4448, 1461-7315
39. Papacharissi, Z.: The virtual sphere: the internet as a public sphere. New Media Soc. **4**(1), 9–27 (2002). https://doi.org/10.1177/14614440222226244. ISSN: 1461-4448, 1461-7315
40. Rains, S.A., et al.: Incivility and political identity on the internet: intergroup factors as predictors of incivility in discussions of news online: incivility and political identity online. J. Comput. Mediated Commun. **22**(4), 163–178 (2017). https://doi.org/10.1111/jcc4.12191. ISSN: 10836101
41. Schulz, M., Mack, B., Renn, O. (eds.): Fokusgruppen in Der Empirischen Sozialwissenschaft: Von Der Konzeption Bis Zur Auswertung. Springer VS, Wiesbaden (2012). https://doi.org/10.1007/978-3-531-19397-7. ISBN: 978-3-531-19396-0

42. Selting, M., et al.: A system for transcribing talk-in-interaction: GAT 2 translated and adapted for english by elizabeth couper-kuhlen and dagmar barth-weingarten. Gesprächsforschung - Online-Zeitschriftzur verbalen Interaktion **12**, 1–51 (2011). ISSN: 1617-1837
43. Volksverhetzung: Hunderte Anzeigen Gegen AfD-Fraktionsvize von Storch — ZEIT ONLINE (2018). https://www.zeit.de/politik/2018-01/volksverhetzungbeatrix-von-storch-strafanzeigen-silvester
44. Waldron, J.: The Harm in Hate Speech. Harvard University Press, USA (2012). ISBN: 978-0-674-06589-5
45. Wickham, H.: tidyverse: easily install and load the 'tidyvers'. R package version 1.3.0. (2019) . https://CRAN.R-project.org/package=tidyverse
46. Wilhelm, C., Joeckel, S.: Gendered morality and backlash effects in online discussions: an experimental study on how users respond to hate speech comments against women and sexual minorities. Sex Roles **80**(7), 381–392 (2018). https://doi.org/10.1007/s11199-018-0941-5. ISSN: 0360-0025, 1573-2762
47. Wong, R.Y.M., Cheung, C.M.K., Xiao, B.: Combating online abuse: what drives people to use online reporting functions on social networking sites. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 415–424. IEEE (2016). ISBN: 978-0-7695-5670-3. https://doi.org/10.1109/HICSS.2016.58
48. Wright, L., et al.: Vectors for counterspeech on twitter. In: Association for Computational Linguistics, pp. 57–62 (2017). https://doi.org/10.18653/v1/W17-3009
49. Xie, Y.: knitr: a general-purpose package for dynamic report generation in R. R package version 1.26 (2019). https://CRAN.Rproject.org/package=knitr
50. Zhu, H.: kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0 (2019). https://CRAN.R-project.org/package=kableExtra