IEAI Technical University of Munich
School of Social Science and Technology
Institute for Ethics in Artificial Intelligence

Research Brief – April 2022

Intervening Against Online Hate Speech: A Case for Automated Counterspeech

by Niklas Felix Cypris, Severin Engelmann, Julia Sasse, Jens Grossklags & Anna Baumert

Online hate speech is a pressing issue that causes great harm to its targets and societal discourse as a whole. Deletion-based approaches commonly used to combat hate speech suffer from practical and ethical issues. In this brief, we discuss the use of automated counterspeech as a supplement to deletion that could address some of those issues. Empirical evidence and psychological theory point toward the potential of automated counterspeech for bystander mobilization against online hate speech and for the general improvement of online discourse. However, the application and the research of automated counterspeech are associated with unique ethical issues. Thus, we propose that automated counterspeech can serve as a valuable supplement to deletion-based approaches to combat online hate speech if it is guided by psychological theory and evaluation, as well as ethical considerations.

In this research brief, we propose automated counterspeech as a supplement to deletion-based approaches that are applied to combat online hate speech. Common deletion-based approaches suffer from ethical as well as practical issues. Deletion often finds itself at odds with the principle of free speech, and current hate speech detection algorithms lack accuracy. Automated counterspeech can serve as a supplement to deletion-based approaches and address some of their shortcomings. Empirical evidence regarding user-generated and (semi-)automated counterspeech (e.g., Miškolci et al., 2020; Munger, 2017), as well as psychological theory (Gambino et al., 2020; Nass & Moon, 2000), point toward the possible effectiveness of automated counterspeech. However, there are also unique ethical challenges that arise from automated counterspeech, namely the definition of transgressions, spillover effects, and the role of deception. We review current research that explores the ethics and feasibility of automated counterspeech and point toward research gaps and future directions.

The Problem: Hate Speech

Online harassment in general and hate speech in particular has become a ubiquitous issue in the virtual space. According to a recent survey (Pew Research Center, 2021), 41% of the participating US adults had been subjected to online harassment of some kind. Often, this online harassment is an expression of prejudice and targets individuals and groups based on their political views, gender, ethnicity, religion or sexual orientation (Pew Research Center, 2021). These kinds of attacks can mostly be subsumed under the term “hate speech.” Hate speech, for our purposes, is “language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used” (Fortuna & Nunes, 2018: p.5).

Research findings suggest that online hate speech may translate into ostracism and actual physical violence

The negative consequences of hate speech are manifold. It can lead to psychological suffering, such as heightened anxiety and depression in targets of racist harassment (Tynes et al., 2008). In a similar vein, an Amnesty International survey (Amnesty Global Insights, 2017) found that many women who were targeted online subsequently experienced stress, anxiety, panic attacks or lowered self-esteem. In South Korea, hateful online harassment has been associated with suicides of celebrities, and subsequent copycat suicide waves in the general population (Nam et al., 2022). Research findings suggest that online hate speech may translate into ostracism and actual physical violence. Frequent exposure to hate speech against specific groups has been associated with increased prejudice towards those groups (Soral et al., 2018). Research from Germany has shown that anti-refugee rhetoric on Facebook can lead to higher crime rates against refugees (Müller & Schwarz, 2021), suggesting that online hate speech may translate into actual physical violence. Furthermore, hate speech can negatively affect online discourse by excluding targeted groups. For example, in the above-mentioned Amnesty International survey, women who had been targeted by harassment reported that they changed their online behavior, up until the point of turning silent and withdrawing from online spaces altogether.

In this research brief, we discuss the application of automated interventions to combat online hate speech and to alleviate its negative impacts on online discourse. First, we cover the practical and ethical challenges of widely used deletion-based approaches. Then, we propose automated counterspeech as a useful supplement to deletion. Regarding its feasibility, we present empirical evidence for the effectiveness of human-generated counterspeech, explore how automated counterspeech can activate the same psychological mechanisms that make human-generated counterspeech effective, and then discuss studies that provide first indications that automated counterspeech can effectively combat hate speech. Finally, we examine the associated ethical challenges of defining transgressions, of spillover effects, and of deception.[1]

How to Counter Hate Speech? Deletion

An established approach to combat hate speech is its deletion, meaning that hate speech is removed by human or automated moderators. Due to the high and rapidly increasing volume of online communication, approaches that involve human moderation become more and more expensive and less tenable in comparison to scalable automated approaches. However, when it comes to automated deletion, there are issues regarding the ethical implications, as well as practical feasibility.

Ethical Implications of User-Content Deletion

For severely transgressive comments, deletion is the only viable reaction. For example, the German 'Network Enforcement Act' of 2017 requires social media platforms to delete illegal content within a timeframe of up to seven days. However, when it comes to hateful content that does not clearly fall into that category, deletion-based approaches run the risk of going against the principle of free speech. Any kind of censorship will need to be justified, and can only work along clearly defined lines in order to limit its infringements in an open public discourse. When defining clear rules for what counts as hate speech, as well as when applying these rules in concrete decisions, there is a trade-off between wide-reaching hate speech criteria that excessively limit free discourse and more conservative criteria that leave more hate speech on the platform. Besides this normative question of what counts as permissible and impermissible expression, there are challenging meta-ethical questions regarding who should be responsible for defining hate speech and who should be responsible for deleting such hate speech.

Hate speech has become a major ethical challenge in the virtual space, however, there are important normative and meta-ethical concerns regarding the power to delete or otherwise censor people's right to free speech.

For example, should a for-profit company be held responsible for defining and enforcing free-speech by deletion and censorship? In the US, platform operators have been shielded from liability for user-generated content through the Communication Decency Act (CDA). The rationale behind the CDA was to empower platforms to develop their own rules for governing content moderation (Ehrlich, 2002). In response to internal and public pressure, platforms have established structures that resemble non-profit governmental bodies such as Google's Right to be Forgotten Advisory Council and Facebook's Oversight Board. Taken together, it is evident that hate speech has become a major ethical challenge in the virtual space, however, there are important normative and meta-ethical concerns regarding the power to delete or otherwise censor people's right to free speech.

[1] This Brief is based on research from the IEAI project - Personalized AI-Based Interventions Against Online Norm Violations: Behavioral Effects And Ethical Implications

Practical Feasibility of Deletion

Another concern is the feasibility of automated deletion-based approaches, where a big issue is the inaccuracy of hate speech detection. Current algorithms (Badjatiya et al., 2017; Burnap & Williams, 2016; Gitari et al., 2015; Malmasi & Zampieri, 2018; Pitsilis et al., 2018; Vidgen & Yasseri, 2020; Waseem & Hovy, 2016; Watanabe et al., 2018) only achieve up to about 90% precision – that is, 10% of the comments classified as hate speech are actually innocuous speech. Moreover, the algorithms only reach up to around 85% recall – that is, around 15% of the hate speech comments are missed. This, in turn, exacerbates the above-mentioned negative ethical implications of fully-automated deletion-based approaches. On the one hand, substantial amounts of unproblematic communication accidentally get labeled as hate speech and deleted. On the other hand, a substantial amount of hate speech goes undetected. Thus, while deletion-based approaches are very common in the current environment, they also present substantial problems from an ethical as well as a practical perspective.

We define counterspeech as any direct response to a transgression such as openly criticizing the hate comment or expressing solidarity with the target of hate speech.

A Supplement to Deletion: Automated Counterspeech

To address these issues, deletion-based approaches could be supplemented with automated counterspeech. That is, counterspeech by artificial agents such as bots who confront comments that are algorithmically determined to be hate speech. We define counterspeech as any direct response to a transgression such as openly criticizing the hate comment or expressing solidarity with the target of hate speech. For example, artificial agents could post pre-written counterspeech messages to address hate speech against certain groups (e.g., Figure 1). Depending on their sophistication, artificial agents could also generate their own comments (Clever et al., 2022). In the following sections, we discuss the empirical evidence for the positive impact of human-generated counterspeech on bystanders and the general tone of online discourse. Then, we explore how counterspeech by human and artificial actors could leverage similar psychological mechanisms. Finally, we review empirical indications for automated counterspeech as an effective tool to combat online hate speech.

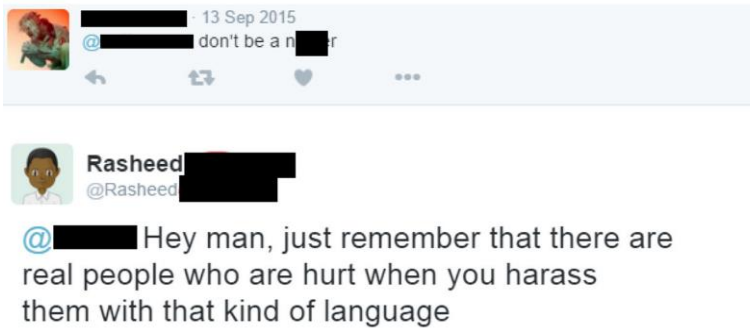

Figure 1
Note. This is part of an image produced by Munger in 2017 displaying an intervention against the racist slur n****r. From "Tweetment effects on the tweeted: Experimentally reducing racist harassment." by K. Munger, 2017, Political Behavior, 39(3), p. 639. Copyright 2018 by American Psychological Association

Human-Generated Counterspeech: Empirical Evidence

Automated counterspeech could exert a positive influence by motivating bystanders to also speak up against hate speech. One indication of this can be found in the positive effects of human-generated counterspeech. While there is mixed evidence regarding the effects of human counterspeech on transgressors (Miškolci et al., 2020; Munger, 2017), positive effects on bystanders are well documented. Users have been found to adjust their comments in line with the rhetoric they encounter in a given online discussion, both with regard to civil and constructive behavior (Berry & Taylor, 2017; Han & Brazeal, 2015; Han, Brazeal & Pennington, 2018; Molina & Jennings, 2018; Seering, Kraut & Dabbish, 2017) as well as uncivil, disruptive online behavior (Cheng, Bernstein, Danescu-Niculescu-

Mizil & Leskovec, 2017; Garland et al., 2020; Gervais & Hillard, 2014; Seering et al., 2017). For example, people were substantially more likely to post comments against Chinese people after having seen negative comments against Chinese people rather than positive ones (Hsueh, Yogeeswaran & Malinen, 2015). Taken together, these results indicate that online conversations tend to strongly shift in line with prior comments.

This emulation effect has also been observed for counterspeech and calls for the moderation of one's rhetoric. For example, Han and colleagues (2018) showed that online users were more likely to speak up in favor of civil discourse online if someone else had already done so, compared to when they only saw hateful comments. Also, if some people criticized group-based hate on Facebook, other users were more likely to speak up in favor of the attacked ethnic group (Miškolci et al., 2020). While there is ample evidence for these effects in studies on social media, the effects have however not been consistently replicated in experimental settings (Leonhard et al., 2018).

Therefore, there is ample evidence that when people interact online, they tend to adjust their own comments to the rhetoric they perceive in a given online space. Moreover, people tend to support counterspeech that they encounter and to be inspired by it themselves.

Counterspeech can also lead to an overall more respectful discourse, thus partially alleviating the negative effects of hate speech. For example, seeing that someone else had already spoken up against Islamophobic hate speech reduced the desire of Muslim participants to reply with hateful comments themselves (Obermaier et al., 2021). There is also tentative evidence that organized counterspeech has a de-escalating effect on the overall discussion climate. A large-scale study investigated communication between Reconquista Germanica (an organized hate group) and Reconquista Internet (an organized counterspeech group) on Twitter (Garland et al., 2020). The amount of counterspeech by Reconquista Internet was associated with decreased aggression as well as the decreased occurrence of hate speech.

Counterspeech can serve as a valuable supplement to deletion-based approaches when it comes to combating hate speech online.

In summary, counterspeech can serve as an effective supplementary tool to combat hate speech and avoids many of the problems connected to deletion. First, counterspeech leaves the original content classified as hate unchanged. Therefore, it does not infringe on free speech as heavily as deletion does. Moreover, although automated counterspeech encounters the same issue of insufficient hate speech detection, due to its positive effects on bystanders and the discursive tone in general, it could further exert a positive impact by shaping a general anti-hate atmosphere. Hence, it can even help to alleviate the negative impact of hate speech that goes unnoticed by hate speech-detection algorithms. Thus, counterspeech can serve as a valuable supplement to deletion-based approaches when it comes to combating hate speech online.[2]

Psychological Mechanisms of User-Generated and Automated Counterspeech

Two major psychological mechanisms are relevant when it comes to the positive effects of human-generated counterspeech on bystanders and it is likely that these mechanisms are similarly affected by automated counterspeech. Comments can, for one, bring a specific mode and tonality of commenting to the forefront and thus make it more accessible to bystanders. Automated counterspeech could function as a behavioral

[2] For a more comprehensive discussion of the effects of human-generated counterspeech and the psychological mechanisms associated with it, see the book chapter by Sasse et al., (2022), on which these passages are based.

prompt for bystanders in the same way. Secondly, counterspeech by commenters to whom bystanders feel some kind of social connection could affect further comments by informing bystanders about social norms. Here, the initial commenters would serve as exemplars – people who represent a social group as a whole and through their behavior, shape perceptions of group norms (Klein et al., 2007; Zillmann, 2002).

Humans have been found to display ingroup bias for artificial actors – that is, they preferred computational agents that are members of their team over ones that are not, just as they do with humans.

Based on the Computers Are Social Agents framework (CASA) (Nass & Moon, 2000), computational agents can trigger and shape social norms in very similar ways to human agents (see also Gambino et al. (2020)). The central tenets of CASA are that humans apply the same social scripts and heuristics to artificial agents that they apply to human counterparts if the artificial agents communicate a minimum of social cues and if they can be perceived as agentic instead of just executing commands (Nass & Moon, 2000). For example, humans have been found to display ingroup bias for artificial actors – that is, they preferred computational agents that are members of their team over ones that are not, just as they do with humans (Nass et al., 1996). In a similar vein, people displayed group conformity with ingroup computer agents if these agents were perceived as being similar to real humans (Xu & Lombard, 2017).

Automated Counterspeech: Empirical Indications

There is ample evidence for the effectiveness of user-generated counterspeech. In addition, there are theoretical indications that artificial agents may be equally effective. Moreover, empirical evidence has emerged that further supports this assumption.

Simulation studies show that social bots can have substantial discursive impacts even when they make up a very small proportion of an overall discursive network (Ross et al., 2019). Further, one study demonstrated the effectiveness of counterspeech by a bot addressing Twitter users who used the racial slur “n****r” to insult others (Munger, 2017). The users were contacted by one of four accounts created by the author and whose profile picture was either White (i.e., the addressee’s in-group) or a Person of Color (i.e., the addressee’s out-group) who had either few or many followers. Regardless of the condition, the accounts would address the person who used the racial slur and post: “Hey man, just remember that there are real people who are hurt when you harass them with that kind of language.” Transgressors were less likely to use the slur subsequently, but only after counterspeech by a popular ingroup member (i.e., White with many followers), compared to a no-intervention control condition.

Single-statement type of counterspeech that was used in this study could easily be automated and used in a multiplicity of interactions. Furthermore, while the addressees were still partially selected manually in that study, automated detection and confrontation in a comparable framework would also be implementable in a relatively straightforward fashion. However, further research is needed to apply such an automated framework and investigate the outcomes of automated counterspeech and to contrast them with human-generated counterspeech.

Ethics of Automated Counterspeech

Even if automated counterspeech proves to be effective under certain conditions – as the reviewed theoretical and empirical literature suggests – such automated counterspeech produces unique ethical challenges.

One key ethical concern stems from the fundamental question of when automated counterspeech is justified. That is, what types of transgressions qualify for counterspeech by bot intervention? For example, in Germany, speech

acts that explicitly threaten the democratic constitutional state are prohibited by law and thus need to be deleted by platform operators. For such unlawful speech acts, counterspeech can hardly be justified because the speech act in question remains published on the platform. Therefore, automated counterspeech may justifiably operate within the scope of highly undesirable online behaviors that are not, strictly speaking, illegal. To reiterate, such definitional work produces normative and meta-ethical challenges. That is, how can we define what speech acts are permissible or impermissible? And who should determine this scope? If automated counterspeech turns out to be an effective "silencer" of online discussions, authoritarian regimes could exploit this technique to shift, re-frame, or subdue user discourse that they deem "undesirable". Importantly, automated counterspeech may be subtle to the extent that it cannot be identified and contested at any time.

If automated counterspeech turns out to be an effective "silencer" of online discussions, authoritarian regimes could exploit this technique to shift, re-frame, or subdue user discourse that they deem "undesirable".

Other ethical challenges of automated counterspeech result from spillover effects. Counterspeech interventions appeal to injunctive norms: they showcase to members of a given group what kind of norm transgressions are permissible and impermissible by demonstrating what happens to those that engage in such transgressions (Alvarez-Benjumea & Winter, 2018). In comparison to deletion or censorship, this requires the transgressive content to remain on the platform, at least for a certain time period. Even if counterspeech leads to a broad understanding among users that such content is morally reprehensible, "the damage may be done" if the visible transgressive content spills over and motivates even few bystanders to share similarly toxic content in the future. Moreover, on social media platforms, content that leads to outrage often draws more attention and results in more engagement than other content. User feedback on content such as likes, comments, or shares determines whether the platform's underlying recommendation and news feed algorithms amplify the content on the platform. The ethical benefits gained from pointing out transgressions to members of a social media group may be less significant in light of possible spillover effects that may be enhanced by the platform's recommender system.

While previous field studies have not performed informed consent and individual debriefing, there may be justifiable reasons to do so in order to protect the studies' validity and the privacy of research subjects

Finally, research that explores the feasibility of automated counterspeech against hate speech on social media platforms creates its own set of unique ethical challenges. Studying counterspeech interventions on social media platforms is an example of field research. Such interventions aim to directly influence online hate speech posts, where they are published, and who they interact with. This, however, comes at a cost. First, research studies cannot ask the study's subjects (i.e., transgressors or bystanders) whether they consent to participate in the research. However, informed consent to participate in a research study is a hallmark of ethical research practice. It seems unlikely that future transgressors would want to participate in a study as moral transgressors in the first place. Moreover, providing sufficient information on the study's goals would very likely influence their future behaviors and thereby decrease the validity of the study. Second, previous studies (e.g., Munger, 2017) have not conducted debriefing of individual subjects after the research study was over. Individual debriefing on Twitter would only be possible if the transgressors "followed" the bot

accounts created by the researchers. Researchers could publicly inform subjects that they had participated in the study. However, publicly declaring that a specific social media user posted hate speech would likely result in more harm than the absence of individual debriefing. While previous field studies have not performed informed consent and individual debriefing, there may be justifiable reasons to do so in order to protect the studies' validity and the privacy of research subjects. Overall, automated counterspeech requires some degree of deception in order to bring about its intended effect: less hateful interactions on social media. The role of the initiator, coordinator and implementer of the automated counterspeech, be it a platform operator, a governmental body or a research team, will matter in the ethical evaluation and justification of this approach.

Conclusion

Online hate speech is a pressing issue that causes a plethora of negative consequences in virtual spaces. Common deletion-based approaches to combat hate speech suffer from ethical as well as practical issues. Deletion often runs contrary to the principle of free speech and has the potential to curtail open discourse if implemented too liberally. Moreover, current hate speech detection algorithms are far from perfectly accurate which further exacerbates ethical issues.

Automated counterspeech could serve as a supplement to deletion-based approaches that addresses some of their shortcomings.

Automated counterspeech could serve as a supplement to deletion-based approaches that addresses some of their shortcomings. User-generated counterspeech has been shown to positively impact online discourse by motivating bystanders to also reject hate speech and to improve the discursive tone on platforms. Thus, it presents itself as an effective and less intrusive intervention against hate speech since it does not delete any user-generated content while nonetheless reducing the negative impact of hate speech.

In addition, through its positive effects on bystanders and the general discourse, it can also alleviate the negative impact of hate speech that it does not confront directly. Psychological theory indicates that automated counterspeech can work in a similar way to user-generated counterspeech.

Counter comments by artificial agents can be expected to leverage the same psychological mechanisms as human commenters. While there is initial empirical evidence that points in this direction, further research is needed to draw comprehensive conclusions.

From an ethical perspective, the development of automated counterspeech by bots needs to address three major questions: Which comments justify automated interventions? What are possible spillover effects? Which issues arise from research about automated counterspeech on social media (e.g., ethical concerns)?

In summary, we argue that supplementing deletion-based approaches with automated counterspeech may be a promising approach to combat online hate speech, if guided by psychological theory and evaluation as well as ethical considerations.

References

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

Amnesty Global Insights. (2017). Unsocial media: The real toll of online abuse against women. Amnesty International. Retrieved from: https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 759–760. https://doi.org/10.1145/3041021.3054223

Berry, G., & Taylor, S. J. (2017). Discussion quality diffuses in the digital public square. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 1371–1380. https://doi.org/10.1145/3038912.3052666

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 11. https://doi.org/10.1140/epjds/s13688-016-0072-6

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1217–1230. https://doi.org/10.1145/2998181.2998213

Clever, L., Klapproth, J., Frischlich, L. (2022). Automatisierte (Gegen-)Rede? Social Bots als digitales Sprachrohr ihrer Nutzer*innen. In: Ernst, J., Trompeta, M., Roth, HJ. (eds) *Gegenrede digital. Interkulturelle Studien*. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-36540-0_2

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 51(4), 1–30. https://doi.org/10.1145/3232676

Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, *1*, 71–86. https://doi.org/10.30658/hmc.1.5

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. *ArXiv:2006.01974 [Cs]*. http://arxiv.org/abs/2006.01974

Gervais, S. J., & Hillard, A. L. (2014). Confronting sexism as persuasion: Effects of a confrontation's recipient, source, message, and context. *Journal of Social Issues*, *70*(4), 653–667. https://doi.org/10.1111/josi.12084

Gitari, N. D., Zhang, Z., Damien, H., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

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

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

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave Your Comment Below": Can Biased Online Comments Influence Our Own Prejudicial Attitudes and Behaviors?: Online Comments on Prejudice Expression. Human Communication Research, 41(4), 557–576. https://doi.org/10.1111/hcre.12059

Kang, N., Kuo, T., & Grossklags, J. (2022) Closing Pandora's Box on Naver: Toward Ending Cyber Harassment. *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM)*, forthcoming.

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

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

Malmasi, S., & Zampieri, M. (2018). Challenges in Discriminating Profanity from Hate Speech. *ArXiv:1803.05495 [Cs]*. http://arxiv.org/abs/1803.05495

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review*, *38*(2), 128–146. https://doi.org/10.1177/0894439318791786

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

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

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

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? International Journal of Human-Computer Studies, 45(6), 669–678. https://doi.org/10.1006/ijhc.1996.0073

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

Netzwerkdurchsetzungsgesetz v. 1.9.2017 (BGBl. I S. 3352)

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

Pew Research Center (2021). The State of Online Harassment. Pew Research Center. Retrieved from: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48(12), 4730–4742. https://doi.org/10.1007/s10489-018-1242-y

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. European Journal of Information Systems, 28(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Seering, J., Kraut, R., & Dabbish, L. (2017). Shaping pro and anti-social behavior on twitch through moderation and example-setting. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17, 111–125. https://doi.org/10.1145/2998181.2998277

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

Sasse, J., Cypris, N., & Baumert, A. (under review), Online Moral Courage. In J. C. Cohrs, N. Knab, & G. Sommer (Eds.), Handbuch der Friedenspsychologie

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online Racial Discrimination and Psychological Adjustment Among Adolescents. Journal of Adolescent Health, 43(6), 565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. Journal of Information Technology & Politics, 17(1), 66–78. https://doi.org/10.1080/19331681.2019.1702607

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL Student Research Workshop, 88–93. https://doi.org/10.18653/v1/N16-2013

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access, 6, 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394

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

Zillmann, D. (2002). Media effects: Advances in theory and research (J. Bryant, Ed.; 2nd ed). L. Elbaum Associates.