SCM
Studies in Communication and Media

EDITORIAL

Hate and counter-voices in the Internet: Introduction to the special issue

Deutscher Titel: Hass und Gegenstimmen im Internet: Einleitung in das Sonderheft

Diana Rieger, Josephine B. Schmitt & Lena Frischlich

Diana Rieger (Prof. Dr.), Department of Communication Studies and Media Research Ludwig-Maximilians University, München; Contact: diana.rieger(at)ifkw.lmu.de

Josephine B. Schmitt (Dr.), Department of Communication Studies and Media Research Ludwig-Maximilians University; Contact: josephine.schmitt(at)ifkw.lmu.de

Lena Frischlich (Dr.), Department of Communication, University of Muenster, Münster; Contact: lena.frischlich(at)uni-muenster.de

EDITORIAL

Hate and counter-voices in the Internet: Introduction to the special issue

Deutscher Titel: Hass und Gegenstimmen im Internet: Einleitung in das Sonderheft

Diana Rieger, Josephine B. Schmitt & Lena Frischlich

Abstract: The Internet creates a space in which hate, negativity and the derogation of certain individuals or social groups (e.g., homosexuals, migrants, or women) can be found in various facets—be it in hateful comments of single users under journalistic online articles or below YouTube videos, propagandistic messages of extremists, or populist speech of political parties. At the same time, the Internet can also empower voices against online hate. The articles in this special issue focus on this spectrum, providing new insights into hate and counter-voices in online media.

Following the ARD/ZDF online study, more than 90 percent of Germans are online, more than half of them use the Internet daily (Frees & Koch, 2018). Besides all benefits, online media also offer a space in which *online hate* (i.e., cyber hate, hate speech, and extremism) flourishes. We conceptualize online hate as norm-transgressing communication that is (1) characterized by the derogation and defamation of single individuals (*offensive speech*) as well as members of targeted social groups (*hate speech*), (2) spread by individual users, social bots, as well as social groups or state actors (Marwick & Lewis, 2017), (3) motivated by personal, social, as well as ideological factors (Erjavec & Kovačič, 2012).

In consequence of a recent German law, the *Network Enforcement Act* (NetzDG, 2017), social media platforms were urged to implement procedures that allow users to report illegal content, often consisting of hate speech or even extremist content. Moreover, the platforms are obliged to canvass these user reports immediately and to delete hateful content if applicable. From January to July 2018, users of social media platforms (in this case: Twitter, Facebook and Google) flagged more than 500,000 posts as being inappropriate (Gollatz, Riedl, & Pohlmann, 2018). Most plausible, the number of unreported cases is far higher. While probably not all of the flagged posts are hateful, populist or extremist in nature, the numbers elucidate that the posting and subsequent removal of online hate is not a rare phenomenon.

Taking these numbers into account, it seems plausible that online negativity, hate, disinformation and propaganda can contribute to the radicalization of indi-

viduals, the polarization of societies, or to a radicalization of public discourse. Although the mere contact to online hate does not trigger radicalization, scientific consensus suggests that online media can serve as a catalyst in complex radicalization or polarization processes (Hohnstein & Glaser, 2017; Meleagrou-Hitchens & Kaderbhai, 2017). In consequence, a better understanding of the nature, processes and effects of hate as "dark side" of online-communication and empirical insights into successful ways to counter it, is highly needed.

The articles in this special issue provide such insights into current research addressing the "dark side" of online communication. At the same time, they go beyond focusing on potentially problematic communication by also addressing empirical insights or drawing conclusions for communicative counter-strategies. In the following, we will detail some of the articles' main contributions to the field while providing a very short overview about the phenomenon of online hate and communicative counter-strategies. Finally, we will highlight some directions for future research in this area.

1. The special issue

1.1 Frequency and nature of online hate

Although notions about online media failing to adhere to general norms of *politeness* (Papacharissi, 2004) due to noxious communication such as *trolling* (Buckels, Trapnell, & Paulhus, 2014) or *flaming* (O'Sullivan & Flanagin, 2003) are nothing new, the risk of exposure to *uncivil online discourses* (Coe, Kenski, & Rains, 2014; Muddiman, 2017) seems to be increasing (Kaakinen, Oksanen, & Räsänen, 2018). In Germany, the Joint Competence Center for the Protection of Minors on the Internet, *Jugendschutz.net*, reported that the number of violations against the youth protection law (JuSchG, 2002) increased from around 6,000 registered violations in 2015 (Schindler, Glaser, Herzog, & Özkilic, 2015) to more than 7,000 in 2016 (Schindler, Glaser, Herzog, & Özkilic, 2016). The share of violations due to extremist content more than doubled during that time (ibid., p. 27).

It is thus not surprising that the recent report by the Online Civil Courage Initaitve (OCII, Baldauf et al., 2018) argues that online hate can contribute to societal polarization and extremist radicalization. More than half (67%) of Germans older than 14 years report experiences with "hate comments," and about half of the adolescents in Germany recall pre-experience with extremist online propaganda (Reinemann, Ninierza, Fawzi, Riesmeyer, & Neumann, 2019). Notably, these numbers mainly refer to witnessing such content. At the same time, when asked about extremist online content, in particular adolescents and young adults report uneasy feelings and low certainty in how extremist messages can be detected.

This finding elucidates that one central difficulty in dealing with online hate is the often-veiled nature of online hate. This can be explained by extremists' attempt to cover their messages in a youth-oriented fashion, following a "wolf-in-sheep-clothes"-strategy (Reinemann et al., 2019; Rieger et al., 2013; Schmitt, Ernst, Frischlich, & Rieger, 2017).

The study by Schwarzenegger and Wagner in this issue demonstrates how extreme right actors in four Western democracies strategically abuse satire and humor to disseminate online hate into the mainstream discourse. They further shed light on how users interact with such content. The particularly skilled actors behind online hate are only one reason why an estimation of the overall frequency of online hate is hard to obtain and also make first contacts with online hate low-threshold.

Although the pre-experience with online hate among Finnish Facebook users is as high as in Germany (67%), only 21% of them were attacked personally (Oksanen, Hawdon, Holkeri, Näsi, & Räsänen, 2014). For highly targeted groups such as members of stigmatized groups (Dieckmann, Geschke, & Braune, 2018), politicians, activists or journalists, these numbers can be substantially higher. For instance, Preuß, Tetzlaff, and Zick (2017) found that 42% of German journalists had been personally attacked during 2016.

The study by **Obermaier, Hofbauer, and Reinemann** puts this numbers into context by showing that the clear majority of journalists in Germany (over 70%) is almost never attacked personally but rather observes attacks on colleagues. The study, however, also demonstrates that being targeted by hate speech can overshadow journalists' relationships with their audiences: Journalists who were frequently attacked responded pugnaciously with increased skepticism regarding their audience, more anger and a higher confirmation in their work.

Noteworthy for journalists as well as victims of discrimination, attacks were observed more often offline than online (Dieckmann et al., 2018; Preuß et al., 2017), underlining that hate is not a phenomenon of online media alone but that those spreading hate, calls for violence and extremist propaganda do use online media, too. As terrorism researcher Peter Neumann summarized it recently: "If 22-years-old foreign fighters in Syria post selfies and publish them in social media they do not only act as extremists but mainly like 99 percent of their contemporaries" (Baldauf, Ebner, & Guhl, 2018, p. 5).

Additionally, relative to overall online content, the share of hateful content in online media is rather small. A large project examining the entire Facebook communication of the Ethiopian diaspora community, a religiously and politically deeply polarized society, found only a small minority of all comments (less than 1%) being hate speech. Similarly, our study on the pre-moderated comments posted in Germany's largest newspaper discussion forum, Spiegel online, shows that only a small share (less than 10%) of user comments entailed direct indicators of hate such as racial slurs or swear words (Boberg, Schatto-Eckrott, Frischlich, & Quandt, 2018). These findings are also supported by a study analyzing hate speech in user comments on YouTube (Ernst et al., 2017). However, even if the amount of hate comments—in relation to the absolute sum of comments—is relatively low, they can have strong effects on people who see them or those who are potentially targeted by them.

1.2 Effects of online hate

There are various studies pointing at potential effects of user comments on perception of the related online content (e.g., Kim, 2015; Lee, 2012; Lee & Jang, 2010; Sikorski & Hänelt, 2016; Weber, 2014). For instance, the valence of comments affects the perceived journalistic quality, the trustworthiness, and persuasiveness of online news articles (Sikorski & Hänelt, 2016). Lee and Jang (2010) demonstrated that people who read comments opposed to news content change their attitudes concerning the news compared to people who read the news article without comments or comments supporting the article's opinion.

Further, online hate has been found to trigger sequences comparable to those caused by other traumatic experiences among members of stigmatized groups (Leets, 2002), diminish social trust among witnesses (Näsi, Räsänen, Hawdon, Holkeri, & Oksanen, 2015), and foster prejudice and aggression amongst those belonging to the perpetrators' social category (Hsueh, Yogeeswaran, & Malinen, 2015; Rösner, Winter, & Krämer, 2016; Soral, Bilewicz, & Winiewski, 2017).

Attempts to repress online hate, such as the German Network Enforcement Law (NetzDG, 2017), have gained substantial critique as potentially diminishing rights of free speech and risking to fuel radicalization itself (George, 2016). In addition, the questions "when does hate begin" and, consequently, "what should be deleted" are difficult to answer. The evaluation of noxious online material can vary substantially between coders (Ross et al., 2016), depending on whether the perspective of the sender, the receiver, or an observer is weighted most (O'Sullivan & Flanagin, 2003). It is therefore not surprising that many actors call for alternative strategies to fight online hate such as the promotion of counter-voices.

1.3 Counter-voices

Counter-voices herein are conceptualized as the large spectrum of communicative responses to online hate. Counter-voices can disseminate messages of tolerance and civility, contributing to civic education in a broader sense. In educational science or clinical context, such messages usually are considered as one *primary prevention* strategy. Counter-voices can also be directed to specific groups or contexts that have been identified as being particularly vulnerable to noxious effects of online hate or that already show an affinity to online hate groups (*secondary prevention*). Finally, counter-voices can respond directly to those spreading hate (*tertiary prevention*). In each of these contexts, other senders (e.g., moderators, other users etc.) can raise their voice against online hate and uncivil communication as the papers in this special issue show.

Addressing the central role of journalistic moderators, **Ziegele, Jost, Bohrmann and Reinbach** show how journalists' own response to online hate can affect subsequent discourses in their comment sections. Their paper makes a strong argument for sociable counter-voices as fruitful and effective strategy against online hate while simultaneously underlying the noxious effects of repressive, authoritarian responses.

Counter-voices are not restricted to moderators or the proprietors of online spaces. Users themselves are a crucial part of most platforms' strategies to fight online hate (for a media report, see Hensel, 2018), yet, the boundary conditions under which users take action against noxious online communication have gained scarce attention so far.

The study by **Leonhard, Rueß, Obermaier, and Reinemann** fills in this void by examining how characteristics of the communication environment, namely the number of other users or "bystanders," of the hate communication itself, and of the individual users, namely their evaluation of the situation and feeling of responsibility, interact in predicting their intentions to raise their voice against online hate.

Noteworthy, raising one's voice alone is not necessarily enough to counter online hate effectively. The simulation model by **Schieb and Preuss** shows that in different communities, different forms of counter-speech are most promising: While in rather neutral contexts, online hate is combated most effectively with harsh rejection, in polarized, hateful environments moderate voices are more effectively in changing the overall tone.

Overall, the study by Ziegele et al., and the work by Schieb and Preuss underline prior research's notion about potentially unwanted side effects of counter-voices (Legault, Gutsell, & Inzlicht, 2011)—particularly when counter-voices do not hold up to civic, participatory standards themselves but act in an authoritative manner or mock about those falling for extremist ideologies (compatible to our own findings, Frischlich, Rieger, Morten, & Bente, 2017).

2. Directions for future research

Although the articles in this special issue contribute to our understanding of online hate and counter voices, they can present only a part of the picture. We will describe some other aspects of online hate and counter-voices in the following aiming to provide starting points for future research beyond the valuable questions raised within the single papers.

2.1 Who spreads online hate?

Overall, little attention is paid to the role of individual dispositions for the voluntarily reception or conscious distribution of online hate. As such, the question of "who spreads online hate" is surely a fruitful venue for future research. The imbalance of studying "who raises their voice against hate" versus "who spreads hate" might be partially due to larger empirical challenges in approaching those spreading online hate. Besides human agents, algorithms and algorithm-based agents (i.e., social bots) can also play a role in the spreading of online hate (Frischlich, Boberg, & Quandt, 2017). Further, algorithms also determine or guide selective exposure to certain contents in online media. Through topical linkage (e.g., reliance on keywords), for example in the case of recommendation algorithms, counter-voices and online hate can be linked in undesired manners (Schmitt, Rieger, Rutkowski, & Ernst, 2018). Initial studies show that including

an actor-centered perspective is crucial for understanding the multiple motives (personal, ideological, economical) underlying the spread of online hate (Buckels et al., 2014; Erjavec & Kovačič, 2012; Jablonska & Kozak, 2017; Marwick & Lewis, 2017).

2.2 In which channel?

A substantial part of communication research on online hate and counter-voices focuses on large platforms such as YouTube, Twitter, Facebook, or the comment sections of journalistic online media. This focus is—to some extent—justified by the reach and hence potential impact of these channels. However, particularly strategic online hate (e.g., by right-wing hate groups) often emerges from more fringe communities or networks such as gap, 4chan, or reddit (Zannettou et al., 2018), respectively is created in specific threads on these communities or in closed groups on larger platforms (Morin & Flynn, 2014; Musgrave, 2017). Even more difficult to study: Online hate and extremism are increasingly spread via end-to-end encrypted instant messengers such as telegram or WhatsApp (Bloom, Tiflati, & Horgan, 2017; Yayla & Speckhard, 2017). Although studying such counter-public or semi-private spheres poses unique challenges for researchers, ranging from theoretically sound sampling over ethical questions up to personal risks when entering these contexts, more systematical research in this area is necessary to understand how online hate is born and breed before entering mainstream online spaces.

In a related vein, online hate might substantially benefit from "approval" by established authorities such as politicians or journalistic media (see e.g. Neumann & Baugut, 2016 on reciprocal media effects on extremists). Increasing numbers of hate crimes in the US (Levin & Grisham, 2017) as well as the UK (Corcoran & Smith, 2016) seem to underline the noxious potential of hateful governments on intergroup relations. Studying the effects of online hate by governmental actors beyond considering it as part of the populist communication logic (Engesser, Fawzi, & Larsson, 2017; Fawzi et al., 2017), thus seems to be a fruitful venue for further research.

2.3 Towards which audience?

The aforementioned success by right-wing populists and their often harsh and hateful rhetoric might partially explain why such forms of hate gained the lion's share of attention over the course of the last months (Bennett & Livingston, 2018; Marwick, Blackwell, & Lo, 2016). It is further complemented by an ongoing focus on extremism proclaimed to be inspired by Islam (Chatfield, Reddick, & Brajawidagda, 2015; Farkas, Schou, & Neumayer, 2018; Winter, 2018). However, online hate and extremism are not bound to certain ends of the political spectrum or certain religions as the spread of hate by for instance nationalist Buddhist monks against the Rohingya has again shown (Gowen & Bearak, 2017).

Previous studies additionally argue that for extreme communication (e.g., populist communication, online hate, extremist messages) to have an effect, the audi-

ence' context plays an important role: Uncertainty or the feeling being threatened can increase the vulnerability (Frischlich, Rieger, Hein, & Bente, 2015; Rieger, Frischlich, & Bente, 2017). Extending research into these areas would help to understand the general mechanisms underlying online hate as communicative phenomenon beyond current use cases.

This becomes even more important when considering the network structure of the internet in which algorithms also decide what users find and stumble upon (Schmitt et al., 2018). Even attempts to fight hate online, for example through the production and distribution of counter-voices can evoke new online hate (Ernst et al., 2017). Studies on the effects of online hate predominantly address the perspective of majority members although hate speech is most often directed to specific minority groups. The effects of hate speech or extremist propaganda, for instance, on members of religious minorities are more seldom. Including their perspectives, however, is crucial to evaluate the impact of online hate and counter-voices as research with minority members convincingly demonstrates (Appel, 2012; Harrell, Hall, & Taliaferro, 2003; Morten, Frischlich, Rieger, & Bente, 2017; Neumann, Arendt, & Baugut, 2017; Rieger, Frischlich, & Bente, 2013).

3. Conclusion

This overview aimed at framing the topic of this special issue and the content presented in the distinct contributions. While the papers in this special add to the empirical evidence in the field of hate and counter-voices in the internet, they also stimulate new perspectives on hate speech, radicalization and prevention strategies in online media.

On a more abstract level, the contributions in this special issue also show the methodological diversity of approaches in communication science to detect, analyze or test the effect of online hate and prevention. They address the phenomenon of online hate and (communicative) counter means with qualitative (Schwarzenegger & Wagner) and quantitative content analyses (Ziegele, Jost, Bohrmann, & Reinbach), a survey (Obermaier, Hofbauer & Reinemann), experiments (Leonhard, Rueß, Obermaier & Reinemann) as well as a computational simulation (Schieb & Preuss), making a strong argument for the value of diverse perspectives in understanding current communication phenomena. We hope that this special issue will contribute to scholarly as well as application-oriented discussions regarding what is arguably one of the mostly debated aspects of online communication.

References

Appel, M. (2012). Anti-immigrant propaganda by radical right parties and the intellectual performance of adolescents. *Political Psychology*, *33*(4), 483–493. https://doi.org/10.1111/j.1467-9221.2012.00902.x

Baldauf, J., Ebner, J., & Guhl, J. (Eds.) (2018). *Hassrede und Radikalisierung im Netz—der OCCI Forschungsbericht* [Hate speech and radicalization in the Internet—the OCCI research report] London, UK: Institute for Strategic Dialogue.

Benesch, S. (2012). *Dangerous speech: A proposal to prevent group violence*. Retrieved from https://tinyurl.com/y8nqmpxj.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bloom, M., Tiflati, H., & Horgan, J. (2017). Navigating ISIS's preferred platform: Telegram. *Terrorism and Political Violence, 6553*, 1–13. https://doi.org/10.1080/09546553.2017.1339695

Boberg, S., Schatto-Eckrott, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, *6*(4), 58–69. https://doi.org/10.17645/mac.v6i4.1493

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences, 67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Chatfield, A. T., Reddick, C. G., & Brajawidagda, U. (2015). Tweeting propaganda, radicalization and recruitment: Islamic State supporters multi-sided Twitter networks. In J. Zhang & Y. Kim (Eds.), *Proceedings of the 16th Annual International Conference on Digital Government Research: Digital Government and Wicked Problems: Climate Change, Urbanization, and Inequality* (pp. 239–249). New York, NY: USA: ACM Press.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Corcoran, H., & Smith, K. (2016). Hate Crime, England and Wales, 2015/16. London, UK: Home Office. Retrieved from https://tinyurl.com/y9km7fqo.

Dieckmann, J., Geschke, D., & Braune, I. (2018). *Diskriminierung und ihre Auswirkungen für Betroffene und die Gesellschaft* [Discrimination and its effects for victims and society]. Jena, Germany. Retrieved from https://www.idz-jena.de/fileadmin/user_upload/PDFS_WsD2/Diskriminierung_und_ihre_Auswirkungen.pdf

Engesser, S., Fawzi, N., & Larsson, A. O. (2017). Populist online communication: introduction to the special issue. *Information, Communication & Society, 20*(9), 1279–1292. https://doi.org/10.1080/1369118X.2017.1328525

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society, 15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Bente, G., & Roth, H. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization, 10*, 1–49. http://journals.sfu.ca/jd/index.php/jd/article/view/91

Farkas, J., Schou, J., & Neumayer, C. (2018). Cloaked facebook pages: Exploring fake Islamist propaganda in social media. *New Media & Society*, *20*(5), 1850–1867. https://doi.org/10.1177/1461444817707759

Fawzi, N., Obermeier, M., Reinemann, C., Alberg, T., Esser, F., Reinemann, C., … de Vreese, C. H. (2017). Germany: Is the Populism Laggard Catching Up? In Toril Aalberg, F. Esser, C. Reinemann, J. Stromback, & C. H. de Vreese (Eds.), *Populist communication in Europe* (pp. 111–126). New York, NY: USA: Routledge.

Frees, B., & Koch, W. (2018). ARD / ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation [ARD/ZDF online-study 2018: Increase in mediated Internet usage and communication]. *Media Perspektiven, 9*, 398–413.

Frischlich, L., Boberg, S., & Quandt, T. (2017). Unmenschlicher Hass: Die Rolle von Empfehlungsalgorithmen und Social Bots für die Verbreitung von Cyberhate [Inhuman hate: The role of recommendation algorithms and social bots for the dissemination of cyberhate]. In K. Kaspar, L. Gräßer, & A. Riffi (Eds.), *Online Hate Speech—Perspektiven auf eine Form des Hasses* (pp. 71–80.). München: kopaed.

Frischlich, L., Rieger, D., Hein, M., & Bente, G. (2015). Dying the right-way? Interest in and perceived persuasiveness of parochial extremist propaganda increases after mortality salience. *Frontiers in Psychology: Evolutionary Psychology and Neuroscience, 6*(1222). https://doi.org/10.3389/fpsyg.2015.01222

Frischlich, L., Rieger, D., Morten, A., & Bente, G. (2017). Wirkung [Effect]. In L. Frischlich, D. Rieger, A. Morten, & G. Bente (Eds.), *Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand* (in cooperation with the research unit Terrorism/Extremism (FTE) of the federal crime police; pp. 81–140). Wiesbaden: Germany: Griebsch und Rochol Druck GmbH, Hamm.

George, C. (2016). Regulating "hate spin": The limits of law in managing religious incitement and offense. *International Journal of Communication, 10,* 2955–2972.

Gollatz, K., Riedl, M. J., & Pohlmann, J. (2018*). Removals of online hate speech in numbers*. HIIG Science Blog. https://doi.org/10.5281/zenodo.1342325

Gowen, A., & Bearak, M. (December 8, 2017). Fake news on Facebook fans the flames of hate against the Rohinga in Burmar. Washington Post.

Harrell, J. P., Hall, S., & Taliaferro, J. (2003). Physiological responses to racism and discrimination: An assessment of the evidence. *American Journal of Public Health, 93*(2), 243–248. https://doi.org/10.2105/AJPH.93.2.243

Hensel, A. (August 21, 2018). *Facebook now gives users who flag fake news a credibility score*. Venturebeat.com.

Hohnstein, S., & Glaser, M. (2017). Wie tragen digitale Medien zu politisch-weltanschaulichem Extremismus im Jugendalter bei und was kann pädagogische Arbeit dagegen tun? Ein Überblick über Forschungsstand, präventive und intervenierende Praxis im Themenfeld [How do digital media contribute to political-ideological extremism during adolescence and what can educational science do against it? An overview about researcj, preventive and interventive practice in the area]. In S. Hohnstein & M. Herding (Eds.), *Digitale Medien und politisch-weltanschaulicher Extremismus im Jugendalter. Erkenntnisse aus Wissenschaft und Praxis* (pp. 243–281). Halle a.d. Saale, Germany: Deutsches Jugendinstitut.

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

Jablonska, M. R., & Kozak, B. (2017). Digital natives towards sponsored online hate speech. *PRZEDSIĘBIORCZOŚĆ I ZARZĄDZANIE, 18*(4), 261–273.

Kaakinen, M., Oksanen, A., & Räsänen, P. (2018). Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior, 78*, 90–97. https://doi.org/10.1016/J.CHB.2017.09.022

Kim, Y. (2015). Exploring the effects of source credibility and others' comments on online news evaluation. *Electronic News*, *9*(3), 160–176. https://doi.org/10.1177/1931243115593318

Lee, E.-J. (2012). That's not the way it is: How user-generated comments on the news affect perceived media bias. *Journal of Computer-Mediated Communication, 18*(1), 32–45. https://doi.org/10.1111/j.1083-6101.2012.01597.x

Lee, E.-J., & Jang, Y. J. (2010). What do others' reactions to news on internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. *Communication Research, 37*(6), 825–846. https://doi.org/10.1177/0093650210376189

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

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science, 22*(12), 1472–1477. https://doi.org/10.1177/0956797611427918

Levin, B., & Grisham, K. E. (2017). *Final U.S. status report: Hate crime analysis & forecast 2016/2017*. San Bernadino, CA, USA: Center for the Study of Hate and Extremism.

Marwick, A., Blackwell, L., & Lo, K. (2016). *Best practices for conducting risky research and protecting yourself from online harassment*. Data & Society. New York, NY: USA: Data & Society Institute.

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society. New York, NY: USA: Data & Society Institute.

Meleagrou-Hitchens, A., & Kaderbhai, N. (2017). *Perspectives on online radicalization, Literature review 2006–2016*. London, UK: Vox Pol. Retrieved from https://tinyurl.com/yb57o9fk

Morin, D. T., & Flynn, M. A. (2014). We are the Tea Party! The use of Facebook as an online political forum for the construction and maintenance of in-group identification during the "GOTV" weekend. *Communication Quarterly, 62*(1), 115–133. https://doi.org/10.1080/01463373.2013.861500

Morten, A., Frischlich, L., Rieger, D., & Bente, G. (2017). Wirksamkeit [Efficacy]. In L. Frischlich, D. Rieger, A. Morten, & G. Bente (Eds.), *Videos gegen Extremismus? Counter-Narrative auf dem Prüfstand* (in cooperation with the research unit Terrorism/Extremism (FTE) of the federal crime police; pp. 161–224). Wiesbaden, Germany: Griebsch und Rochol Druck GmbH.

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication, 11,* 3182–3202. https://ijoc.org/index.php/ijoc/article/view/6137

Musgrave, S. (August 9, 2017). *The secret Twitter rooms of Trump nation*. Politica.

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. Information Technology & People, 28(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

Netzwerkdurchsetzungsgesetz (NetzDG) (2017). *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken* [Networkenforcement act].

Neumann, K., Arendt, F., & Baugut, P. (2017). *News and Islamist radicalization processes: Investigating muslims' perceptions of negative news coverage of Islam*. Paper presented at 67th Annual Conference of the International Communication Association, San Diego, USA.

Neumann, K., & Baugut, P. (2016). *Neonazis im Scheinwerferlicht der Medien-Eine Analyse reziproker Medieneffekte innerhalb der Neonazi-Szene in Deutschland*. Wiesbaden: Germany: Springer, VS.

O'Sullivan, P. B., & Flanagin, A. J. (2003). Reconceptualizing "flaming" and other problematic messages. *New Media & Society*, *5*(2), 1461–4448. https://doi.org/10.1177/1461444803005001908

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., & Räsänen, P. (2014). Exposure to online hate among young social media users. In N. M. Warehime (Ed.), *Soul of society: A focus on the lives of children & youth*. Oklahoma City: Emerald Books.

Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Preuß, M., Tetzlaff, F., & Zick, A. (2017). *"Publizieren wird zur Mutprobe": Studie zur Wahrnehmung von und Erfahrung mit Angriffen unter Journalist_innen* ["Publishing becomes a dare": A study on the perception and experience with attacks among journalists]. Bielefeld, Germany.

Reinemann, C., Ninierza, A., Fawzi, N., Riesmeyer, C., & Neumann, K. (2019). *Jugend—Medien—Extremismus* [Youth – Media - Extremism]. Wiesbaden: Springer VS.

Rieger, D., Frischlich, L., & Bente, G. (2013). *Propaganda 2.0: Psychological effects of right-wing and Islamic extremist internet videos*. Cologne, Germany: Wolters Kluwer Deutschland GmbH.

Rieger, D., Frischlich, L., & Bente, G. (2017). Propaganda in an insecure, unstructured world: How psychological uncertainty and authoritarian attitudes shape the evaluation of right-wing extremist internet propaganda. *Journal for Deradicalization*, *10*, 203–229.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461–470. https://doi.org/10.1016/j.chb.2016.01.022

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In Beißwenger, M., & Wojatzek, M. (Eds.), *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, September 2016*. Bochum: Sprachwissenschaftliches Institut.

Schindler, F., Glaser, S., Herzog, H., & Özkilic, M. in collaboration with the team of jugendschutz.net (2015). *Protection of minors on the Internet –Annual Report 2015*. Mainz, Germany: Jugendschutz.net.

Schmitt, J. B., Ernst, J., Frischlich, L., & Rieger, D. (2017). Rechtsextreme und islamistische Propaganda im Internet: Methoden, Auswirkungen und Präventionsmöglichkeiten [Right-wing extremist and Islamist extremist propaganda in the internet: Methods, effects and possibilities for prevention]. In R. Altenhof, S. Bunk, & M. Piepenschneider (Eds.), *Politischer Extremismus im Vergleich* (pp. 171-210). Münster, Germany: LIT Verlag.

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube recommendation algorithms. *Journal of Communication*, *68*(4), 780–808. https://doi.org/10.1093/joc/jqy029

Sikorski, C. von, & Hänelt, M. (2016). Scandal 2.0 How valenced reader comments affect recipients' perception of scandalized individuals and the journalistic quality of online

news. *Journalism & Mass Communication Quarterly*, *93*(3), 551–571. https://doi.org/10.1177/1077699016628822

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

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, *60*(3), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x

Weber, P. (2014). Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New Media & Society*, *16*(6), 941–957. https://doi.org/10.1177/1461444813495165

Winter, C. (2018). Apocalypse, later: a longitudinal study of the Islamic State brand. *Critical Studies in Media Communication*, *35*(1), 103–121. https://doi.org/10.1080/15295036.2017.1393094

Yayla, A. S., & Speckhard, A. (2017). *Telegram: the mighty application that ISIS loves*. ICSVE Brief Reports. International Center for the Study of Violent Extremism. Retrieved from https://www.voxpol.eu/telegram-mighty-application-isis-loves/

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. *arXiv preprint arXiv:1805.12512*. https://arxiv.org/pdf/1805.12512.pdf