Article

new media & society
2024, Vol. 26(5) 2614–2632

DOI: 10.1177/14614448221091185
journals.sagepub.com/home/nms
SAGE

How social media users perceive different forms of online hate speech: A qualitative multi-method study

Ursula Kristin Schmid
LMU Munich, Germany

Anna Sophie Kümpel
TU Dresden, Germany

Diana Rieger
LMU Munich, Germany

Abstract

Although many social media users have reported encountering hate speech, differences in the perception between different users remain unclear. Using a qualitative multi-method approach, we investigated how personal characteristics, the presentation form, and content-related characteristics influence social media users' perceptions of hate speech, which we differentiated as first-level (i.e. recognizing hate speech) and second-level perceptions (i.e. attitude toward it). To that end, we first observed 23 German-speaking social media users as they scrolled through a fictitious social media feed featuring hate speech. Next, we conducted remote self-confrontation interviews to discuss the content and semi-structured interviews involving interactive tasks. Although it became apparent that perceptions are highly individual, some overarching tendencies emerged. The results suggest that the perception of and indignation toward hate speech decreases as social media use increases. Moreover, direct and prosecutable hate speech is perceived as being particularly negative, especially in visual presentation form.

Corresponding author:
Ursula Kristin Schmid, Department of Media and Communication, LMU Munich, Oettingenstr. 67, 80538 Munich, Germany.
Email: ursula.schmid@ifkw.lmu.de

Keywords
Hate speech, incivility, perceptions, qualitative research, social media

The conversational tone online has become noticeably more aggressive in recent years as social media platforms have emerged as a space in which hostility and hate speech flourish (e.g. Mondal, 2017). As representative surveys have shown, many Internet users around the globe have reported being exposed to hate speech. Whereas self-reported exposure has ranged from approximately 28% among New Zealanders at least 18 years of age (Pacheco and Melhuish, 2018) to 64% among 13- to 17-year-old Americans (Common Sense, 2018), in Germany, the number has continuously risen and even reached 78% in 2021 (Steppat, 2021). Young users especially report frequent exposure to the form of incivil online communication called *hate speech*, defined as the expression of "hatred or degrading attitudes toward *a collective*" (Hawdon et al., 2017: 254). Although the results of standardized surveys suggest that many Internet users have been exposed to hateful content online, they reveal little about how specific user groups perceive hate speech and which contextual factors influence their perceptions. Beyond that, few studies have investigated whether, not to mention how, social media users perceive different forms of online incivility differently (e.g. Kenski et al., 2020). Moreover, because the research to date has focused on confronting participants with isolated statements, it has ignored the influence of the social media environment and the specific form in which hate speech is expressed. Indeed, hate speech can have many faces that may be perceived, recognized, and/or understood differently among users with unique backgrounds and normative concepts (Bormann et al., 2021).

Despite the large body of research on the prevalence of hate speech on social media (Matamoros-Fernández and Farkas, 2021), we argue, along with Bormann et al. (2021), for a sharper focus on the audience and how individual characteristics and contextual factors shape their perceptions. That focus is essential to developing adequate intervention strategies (Rafael, 2021) and finding ways of counteracting the normalization of hate speech on social media (Bilewicz and Soral, 2020).

To address the mentioned gaps in the research, we first reviewed the literature on the topic before elaborating the following three sets of different characteristics that might influence perceptions of hate speech on social media: (1) personal characteristics of the social media users, (2) the presentation form of hate speech, and (3) content-related characteristics. In a preregistered qualitative multi-method study with 23 German-speaking social media users 18- to 67 years old, we next examined the role of those characteristics in perceptions of hate speech. Given our interest in both users' initial recognition of hate speech and their detailed engagement with and evaluation of such content, we distinguished first-level from second-level perceptions of hate speech. Our study revealed considerable differences regarding the presentation form and directness of hate speech, with direct, visual representations exerting the most decisive impact on social media users. Furthermore, differences in perceptions and reactions depended on participants' age and social media usage, which indicates potential desensitization to hate speech in cases of frequent exposure. Overall, however, perceptions of hate speech seemed to be remarkably individual.

The prevalence and dissemination of hate speech on social media

Hate speech is a type of hostile online communication targeting social groups to insult, degrade, or belittle their members (Hawdon et al., 2017). As such, hate speech is regarded as an extreme expression of online *incivility*, broadly defined as bearing “features of discussion that convey an unnecessarily disrespectful tone” (Coe et al., 2014: 660). Although numerous definitions of *hate speech* have emerged in recent years (Paasch-Colberg et al., 2021), our study focused on the public expression of hate or degrading attitudes toward a collective, whose targets are devalued based on group-defining characteristics (e.g. race and/or religion) instead of individual traits (Rieger et al., 2021). In Germany, most online hate speech recognized by Internet users is directed toward politicians and minorities and focuses on their race, religion, and/or sexual orientation (Geschke et al., 2019). Moreover, women are often victims of sexist hostility or stereotypical gender-based hate speech (Henry and Powell, 2018).

Although hate speech can be expected wherever discourses occur online, such speech thrives on social media platforms, where hatred and agitation are widespread (Matamoros-Fernández and Farkas, 2021; Zhang and Luo, 2019). Social media platforms seem to be the perfect place for disseminating hate speech due to not only several of their defining characteristics but also the characteristics of social media users. For one, social media platforms enable hate groups to develop, connect, and organize, even internationally, and the resulting clusters of hate (Johnson et al., 2019) facilitate the spread of hate speech across platforms (Nakamura, 2014). For another, actual or perceived anonymity in social media environments and the invisibility of other individuals can embolden users to “be more outrageous, obnoxious, or hateful in what they say” (Brown, 2018: 298). Moreover, social media networks act as “amplifiers and manufacturers of racist discourse by means of their affordances and users’ appropriation of them” (Matamoros-Fernández, 2017: 949). Together with their algorithmic recommendation systems, their interactive tools (e.g. comment and share functions) facilitate the mainstreaming of incivil content such that even uninvolved users find themselves confronted with hate speech (Matamoros-Fernández, 2017, 2018; Schulze et al., in press). Finally, the operators of social media platforms often act out of economic considerations, not considerations prioritizing community welfare (Matamoros-Fernández, 2018). In response, in Germany, as in many other countries, new measures to restrict hate speech on social media platforms have been implemented; however, those measures represent only a few of the strategies needed to combat hate speech. To adequately prepare individuals for measures against such incivility, it is equally important to understand how social media users perceive hate speech (Rafael, 2021).

Contextual factors of social media users’ perceptions of hate speech

The way in which social media content is perceived depends on a variety of contextual factors. As any other media perception, the perception of hate speech is not universal but hinges on individuals’ characteristics, the presentation of the message, and its content. As for what we mean by *perception*, in our study, we referred to Ohme and Mothes' (2020) systematization of selective exposure on social media and differentiated two levels of the perception of hate speech. *First-level perceptions* describe users' attention to single posts containing hate speech while scrolling through social media feeds. Accordingly, first-level perceptions involve recognizing hate speech and deciding "to slow down, or even stop scrolling . . . to look more carefully at a specific post, based on message cues that are immediately visible" (Ohme and Mothes, 2020: 1223). Thereafter, *second-level perceptions* are possible, which describe a user's more intensive engagement with the hate speech encountered. Those perceptions thus entail users' feelings, attitudes, and opinions regarding the content.

Personal characteristics

Among social media users exposed to hate speech, different first- and second-level perceptions of such speech may result from their social distance to and/or personal involvement with the targeted group. Individuals targeted by hate speech and who have thus experienced threats to their social identity may pay more attention to hate speech and be more emotionally affected by it than individuals who have not. Research has suggested that being confronted with hate speech can have the same consequences as other traumatic events (Leets, 2002), cause frustration, fear, and anger (e.g. Masullo Chen and Lu, 2017), and induce psychological stress or even depression (Gelber and McNamara, 2016). Along with individuals who have previously been targeted by hate speech, women seem to have a heightened sensitivity to incivility in general and typically experience it as being more severe than men do (Kenski et al., 2020; Stryker et al., 2016). Moreover, preexisting attitudes seem to factor into perceptions of hate speech, for individuals are less likely to judge statements consistent with their own opinions as being hate speech, and vice versa (Wojatzki et al., 2018).

Another factor that might influence individuals' perception of hate speech is their social media usage. Evidence suggests that exposure to hate speech increases with increased social media usage; in particular, the more often Internet users consume political news on social media platforms, the more regularly they notice and consequently react to hate speech (Ziegele et al., 2020). Thus, for users who use social media more frequently, both first- and second-level perceptions of hate speech are likely to be more pronounced. Recent studies also indicate that past experience with hate speech can increase the likelihood of intervening against it (Schmid et al., 2022). On one hand, we might therefore assume that frequent social media users perceive hate speech as being particularly severe because they regularly encounter it, see the "whole picture," so to speak, and are aware of the problems that it can cause. On the other hand, being frequently exposed to hate speech may encourage desensitization to and the normalization of hate speech (Bilewicz and Soral, 2020; Santos et al., 2020) and thereby, lead to less recognition of (i.e. first-level perceptions) and less engagement with (i.e. second-level perceptions) such content.

Altogether, competing assumptions thus exist about whether and how the perception of hate speech differs according to level of social media usage, experiences with hate speech, and users' overall awareness of the problem. Against that background, the first goal of our exploratory study was to examine the role that those personal characteristics and factors play when individuals are confronted with hate speech on social media. To that purpose, we developed our first research question (RQ):

> *RQ1*. What role do the personal characteristics of gender and age as well as social media usage and previous experiences with hate speech play in social media users' perceptions of hate speech?

Presentation form

Hate speech on social media is communicated in different ways, not only within textual posts or user comments but also in (audio)visual forms such as videos and memes. Concerning user comments, experimental research has revealed that hateful and/or incivil comments not only induce negative emotions and/or hostile cognitions (Masullo Chen and Lu, 2017) but also influence users' perceptions of the content commented upon (Prochazka et al., 2018). However, other research has shown that user comments at the bottom of webpages usually receives only minor attention, if they are read at all (Haßler et al., 2019), which may cause hate speech within user comments to be overlooked or recognized less than hate speech within original posts. Aside from comments, and in some contrast to the term hate *speech*, visual representations and text-image combinations are widespread forms of online hate speech, especially memes (Schmitt et al., 2020; Zannettou et al., 2018). The visual character of such forms serves hate speech well, for images capture users' attention and provoke their emotions better than text because they are processed faster, remembered more easily, and thus facilitate the association of specific information with concrete imagery (e.g. Powell et al., 2015; Stenberg, 2006). We might therefore assume that visual hate speech is not only recognized faster (i.e. first-level perceptions) but also more provocative (i.e. second-level perceptions) than textual hate speech, especially when both forms are present in a social media feed. Building on the above, we asked the following:

> *RQ2*. What role does the presentation form of hate speech content (i.e. textual or visual) play in social media users' perception of hate speech?

Content-related characteristics

Regardless of its presentation form, hate speech can either be direct and "in your face" (Borgeson and Valeri, 2004)—for example, when featuring insults and direct verbal attacks—or more indirect and subtle (Meibauer, 2013). Such indirect forms of hate speech express hate more covertly by spreading negative stereotypes, strategically elevating one's in-group, and/or cloaking prejudices in supposedly ordinary statements (Åkerlund, 2021; Ben-David and Matamoros-Fernández, 2016). Examining the prevalence of those different types of hate speech on selected news sites, social media pages, and blogs, Paasch-Colberg et al. (2021) identified indirect stereotypical terms and generalizations as being most prevalent in the context of anti-immigrant content than direct, more extreme forms (e.g. threats of violence). In alt-right fringe communities, indirect hate speech is also more prevalent than direct hate speech, even though those communities are known for being outspokenly hateful (Rieger et al., 2021). However, other evidence suggests that overt forms of hate speech containing threats of violence are perceived as being more threatening and harmful than hate speech without such threats (Leonhard et al., 2018). In general, compared with other forms of incivil communication, statements containing insulting language and name-calling (Kenski et al., 2020) as well as threats of violence and insults are rated as being the most incivil (Stryker et al., 2016). However, because hate speech was considered in isolation (i.e. not integrated into a website or social media feed), those studies ignored the broader environment and context of the occurrence of hate speech as well as different levels of perception.

Explicit hate speech is often associated with its relevance in criminal law, although that association is not congruent in every case. Added to offenses under Germany's Criminal Law Code ("Strafgesetzbuch; StGB") and Network Enforcement Act ("Netzwerkdurchsetzungsgesetz"), many incidents of hate speech are not prosecutable (Wolter, 2020), even if they are no less harmful and/or entail direct offenses (Ben-David and Matamoros-Fernández, 2016). Thus, despite a more or less clear division between punishable and non-punishable hate speech under German law,[1] it remains unclear whether punishable forms are also perceived as being worse or more hateful than non-punishable ones. That question is particularly pressing given recent research on different types of hate speech showing that audiences perceive antagonistic stereotypes as being similarly incivil and harmful as direct hateful expressions (Ziegele et al., 2020). In some contexts, utterances of implicit hate speech are more likely to be classified as hate speech than their explicit counterparts (Benikova et al., 2018). Although direct hate speech may appear to be worse at first glance, indirect statements can have strong long-term media effects, especially if observed with frequency and consonance (Paasch-Colberg et al., 2021).

Within indirect hate speech, humorous stylistic devices such as irony, sarcasm, and satire (Filibeli and Ertuna, 2021; Matamoros-Fernández, 2017) are common, particularly within visual presentation forms such as hate memes (Schmitt et al., 2020; Zannettou et al., 2018). That combination of hate speech with humorous elements can reduce the recognition of the hateful message as such because it is perceived as being only a joke (Billig, 2001). When using humorous phrases, communicators are less prone to be accused of discriminatory intentions (Woodzicka et al., 2015), which can cause audiences to judge humorous hate speech as being less severe. Otherwise, if the hostile intentions behind the humor are recognized, then audiences may judge such implicit forms as being even more severe. As Benikova et al. (2018) have suggested, the "sly, potentially deceiving nature of implicitness might be perceived as more hateful, whereas the same content expressed clearly might be perceived as more honest and thus less hateful" (p. 177). Taken together, contradictory assumptions about the perception of the directness of different forms of hate speech and their relevance in criminal law are evident. Thus, our final RQ was the following:

> *RQ3*. What role do the content-related characteristics of directness and relevance in criminal law regarding hate speech play in social media users' perceptions of hate speech?

Method

Design and participants

To answer our RQs, we conducted remote self-confrontation interviews with 23 German-speaking social media users (age: 18–67, women: $n=12$, educational degree of Abitur or higher: $n=12$) during June and July 2021.[2] Building upon a quota-based plan with the criteria of age, gender, and education (see preregistration for details), participants were recruited through the distribution of informative flyers and third parties (i.e. family members, colleagues, and former participants were asked to spread the word about the study). Lasting 45–80 minutes, the interviews were conducted on the videoconferencing platform Zoom, which allowed us to record the entire interviews and transcribe all relevant aspects afterward. Each interview was conducted with the participant's informed consent, and participants were made aware of their right to refuse participation and told about how the data would be used. The study was reviewed and approved by the first author's university ethics committee, and each interviewee received €50 as an incentive for participation.

To answer our RQs, we adapted the self-confrontation interview method (Lim, 2002). The method entails confronting participants with their behavior by means of an artifact (e.g. videos or screen capture recordings) and asking them to report the thoughts and feelings that they had while performing the behavior. The method's chief benefit is that it allows participants to perform the behavior of interest largely undisturbed and without interference of researchers. Only afterward is the behavior addressed and discussed in detail. Our adaptation of the method altered two important aspects. First, we conducted the interviews remotely. Second, given our research interest, we did not confront participants with video recordings of their entire browsing sessions but only with the manipulated social media feed that they were asked to browse. In doing so, we were able to direct discussions toward users' reactions to the (hate speech) content encountered.

Our specific design involved four steps that allowed granular insights into users' first- and second-level perceptions of social media content (see Figure 1 for an overview). First, we observed participants as they scrolled through the feed of a fictitious social media platform integrated with different forms of hate speech (Step 1: Observation; see Figure 2 and "Stimulus Material"). To create a plausible scenario for such activity, participants were not aware of the platform's or the content's fictitiousness but told that they were piloting a new social media platform free of content moderation and restrictions. Moreover, they were not told beforehand that the study focused on hate speech. No time limit was specified for the task, and participants were asked to view the content at their own pace. The observation focused on participants' dwell times, intensity of their engagement with posts in the feed, and their nonverbal reactions, including facial expressions. After participants completed the browsing session, the simulated feed was brought back up, with the researcher scrolling to pivotal posts on the feed to gather participants' comments and evaluations (Step 2: Self-confrontation). In a third step, additional semi-structured interviews were conducted to gain insights into participants' characteristics, attitudes, social media use, and awareness of and experiences with hate speech. Fourth, participants were asked

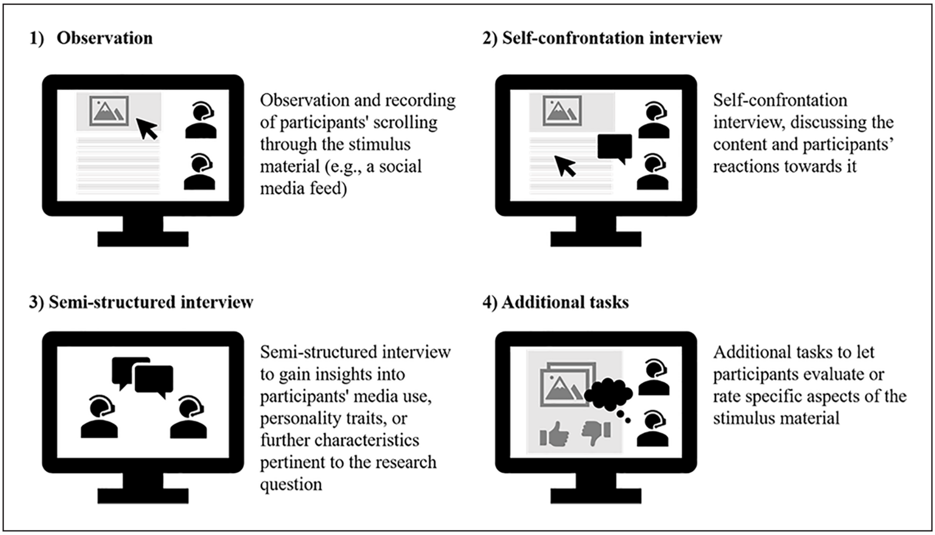
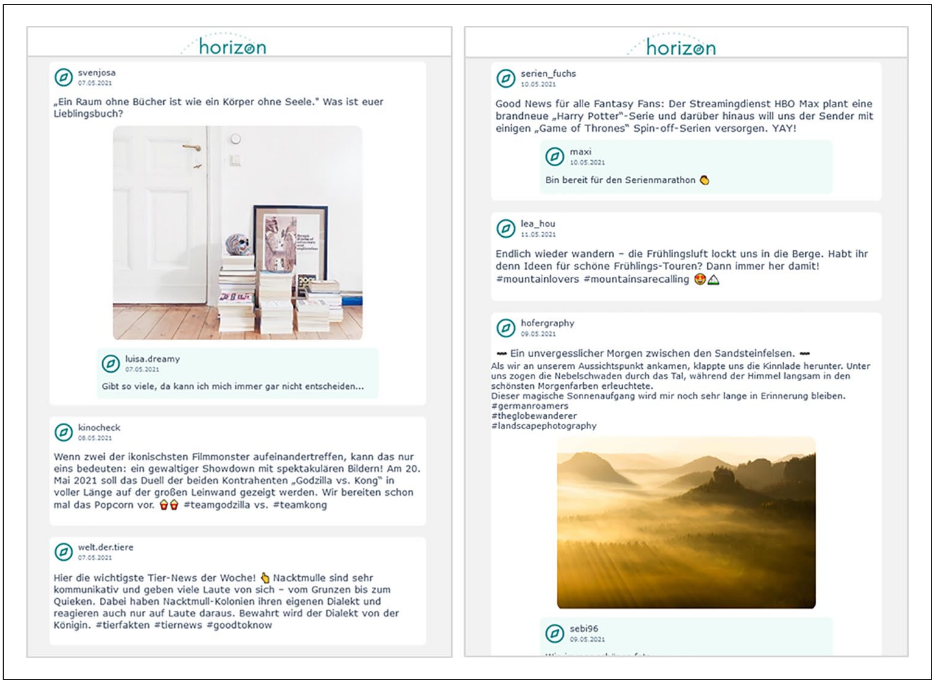

1) Observation
Observation and recording of participants' scrolling through the stimulus material (e.g., a social media feed)
2) Self-confrontation interview
Self-confrontation interview, discussing the content and participants' reactions towards it
3) Semi-structured interview
Semi-structured interview to gain insights into participants' media use, personality traits, or further characteristics pertinent to the research question
4) Additional tasks
Additional tasks to let participants evaluate or rate specific aspects of the stimulus material

Figure 1. Setup of the remote self-confrontation interview method.

whether they would report specific hateful posts in an interactive setup (i.e. rate posts with a thumbs-up or thumbs-down). After completing the interviews, participants were informed about the study's aim and debriefed regarding the fictitiousness of both the platform and the content displayed. We also provided participants with more information about hate speech and links to support websites.

Stimulus material

To investigate the impact of various forms of hate speech, we created stimulus material based on real-world hate speech by combining different forms with the abovementioned factors of content. Overall, nine posts—11% of the content in the simulated social media feed—contained elements of hate speech, which approximated the amount of hate speech on existing social media platforms (i.e. Twitter: Zhang and Luo, 2019). Considering different target groups, we selected hate speech addressing Jews, (immigrant) Muslims, women, and the lesbian, gay, bisexual, transgender, queer or questioning, intersex, and asexual (LGBTQIA+) community and thus covered a broad range of hate speech directed against people's religion, ethnicity, gender, and sexual orientation, which are the most common targets of hate speech topics in Germany (Geschke et al., 2019). Regarding hate speech's characteristics of directness and relevance in criminal law, we created the stimulus material with reference to theoretical considerations and real-world hate speech.[3] As shown in Table 1, a mixture of (non-)prosecutable, (in)direct, and textual and/or visual hate speech was created and integrated into the social media feed. The feed's remaining content was designed based on real-world examples from social media platforms and featured various content, including pictures of nature, users' status updates, and inspirational images.



svenjosa
07.05.2021

„Ein Raum ohne Bücher ist wie ein Körper ohne Seele." Was ist euer Lieblingsbuch?

luisa.dreamy
07.05.2021

Gibt so viele, da kann ich mich immer gar nicht entscheiden...

kinocheck
08.05.2021

Wenn zwei der ikonischsten Filmmonster aufeinandertreffen, kann das nur eins bedeuten: ein gewaltiger Showdown mit spektakulären Bildern! Am 20. Mai 2021 soll das Duell der beiden Kontrahenten „Godzilla vs. Kong" in voller Länge auf der großen Leinwand gezeigt werden. Wir bereiten schon mal das Popcorn vor. 🍿🍿 #teamgodzilla vs. #teamkong

welt.der.tiere
07.05.2021

Hier die wichtigste Tier-News der Woche! 👆 Nacktmulle sind sehr kommunikativ und geben viele Laute von sich – vom Grunzen bis zum Quieken. Dabei haben Nacktmull-Kolonien ihren eigenen Dialekt und reagieren auch nur auf Laute daraus. Bewahrt wird der Dialekt von der Königin. #tierfakten #tiernews #goodtoknow

serien_fuchs
10.05.2021

Good News für alle Fantasy Fans: Der Streamingdienst HBO Max plant eine brandneue „Harry Potter"-Serie und darüber hinaus will uns der Sender mit einigen „Game of Thrones" Spin-off-Serien versorgen. YAY!

maxi
10.05.2021

Bin bereit für den Serienmarathon 👏

lea_hou
11.05.2021

Endlich wieder wandern – die Frühlingsluft lockt uns in die Berge. Habt ihr denn Ideen für schöne Frühlings-Touren? Dann immer her damit! #mountainlovers #mountainsarecalling 😍⛰

hofergraphy
09.05.2021

〰 Ein unvergesslicher Morgen zwischen den Sandsteinfelsen. 〰
Als wir an unserem Aussichtspunkt ankamen, klappte uns die Kinnlade herunter. Unter uns zogen die Nebelschwaden durch das Tal, während der Himmel langsam in den schönsten Morgenfarben erleuchtete.
Dieser magische Sonnenaufgang wird mir noch sehr lange in Erinnerung bleiben.
#germanroamers
#theglobewanderer
#landscapephotography

sebi96
09.05.2021

Figure 2. Extract from the (German) stimulus material.

Table 1. Investigated combinations of presentation form and content-related characteristics of the hate speech content in the stimulus material.

	Presentation form	
	Textual	Visual
Content-related characteristics	Anti-Islam (user comment) Indirect (through prejudice) Not prosecutable	Anti-LGBTQIA+ (meme) Indirect (through humor) Not prosecutable
	Anti-LGBTQIA+ (post) Direct Threat of violence §241 StGB	Anti-feminist (text-image) Indirect (through stereotypes) Not prosecutable
	Anti-Islam (user comment) Direct Reward and approval of criminal acts §140 StGB	Anti-semitic (meme) Direct Incitement of the people §130 StGB + Swastika §86a StGB
	Anti-feminist (user comment) Direct Insult §185 StGB	Anti-semitic (meme) Direct Incitement of the people §130 StGB
		National socialistic (picture) Direct (+indirect) Swastika §86a StGB

Data analysis

The data from interviews and observations were scrutinized using qualitative content analysis, which integrates inductive category formation and deductive category assignment (Mayring, 2015). Accordingly, we relied on a predefined coding scheme that we developed based on our study's focus as well as theoretical considerations and added categories during analysis to account for new themes and perspectives prevalent in the data. Both coding and analysis were performed in MAXQDA version 20.4.1. To ensure comparability and consensus, we organized a mutual exchange in which three researchers analyzed and discussed the data. Because the interviews were conducted in German, all quotations have been translated into English. The parenthetical expressions after the quotations—for example "(Sarah, 28, w)"—inform readers about the anonymized names, age, and gender of the quoted participants.

Results

Being confronted with most types of hate speech triggered either negative associations and emotions (e.g. anger or fear) or a lack of understanding among participants. However, some of the posts containing hate speech went unnoticed. Upon recognizing hate speech, most participants stopped scrolling through the feed for a moment; in a few cases, their recognition was visible in the form of frowns, raised eyebrows, and/or headshaking. The motivation to deal more intensively with the hate speech was low, primarily due to the participants' self-protective avoidance of harmful content. Overall, qualitative data analysis revealed that personal and content-related characteristics as well as the presentation form of hate speech considerably shaped the perception of hate speech on social media platforms. In the following, we discuss the role of each of those factors on three levels: first-level perceptions (i.e. recognition of hate speech content), second-level perceptions (i.e. attitudes and opinions toward the hate speech), and presumed effects and intentions to engage in counterspeech or reporting.

The role of personal characteristics

Participants' gender hardly seemed to be relevant for the perception of hate speech. Few women gave indirect hate speech against their group more attention and drew more personal conclusions than men but nevertheless stated that such utterances did not affect them personally. As one young woman explained, "This [content] somewhat belongs to the category of 'If you think that, then I think it's a pity'. But in that case, it doesn't bother me very much" (Mara, 26, w).

Differences in perceptions based on participants' age and social media usage were more clear-cut, even though those factors were associated in our sample. Whereas the recognition of hate speech was relatively low in all age and user groups, younger and more experienced social media users tended to recognize hate speech even less (i.e. first-level perceptions). Moreover, upon identifying hate speech, those user groups dealt with it more cursorily (i.e. second-level perceptions were more superficial) because, for example, engaging with it was not perceived to be worthwhile or they simply did not want to involve themselves with harmful content. They neither did so in our study setting—"As soon as I realized that it was something about [hate speech], I moved on relatively quickly, because I thought to myself, 'I don't even want to see that'" (Anna-Lena, 19, w)—nor did they report doing so in their daily life: "If this post appeared on my Facebook feed, I would just shake my head and scroll on" (Markus, 28, m). Among older participants, who generally use social media less often, attention was somewhat greater and reactions observed through facial expressions more emotional, often due to their lack of familiarity with hate speech on social media: "I think I looked at this [user comment containing hate speech] longer than at this [main post], because I felt that it can't be true that someone would write something like that. I had to read it again" (Sandra, 66, w).

Taken together, desensitization to and the normalization of hate speech resulting from frequent exposure seem likely, as one participant also suspected, "When I see something like that for the first time, it really affects me. . . . But if you already know about it, then you're more resistant to it" (Katharina, 19, w). After being targeted with incivil comments online herself, one woman, for example, avoided further elaborating on hate speech in our study. She reported preferring to ignore it instead, for she had had the negative experience of personally confronting the perpetrators, failing to make any difference, and thus feeling worse afterward. Since then, she has concluded somewhat resignedly, "It would be nice if that no longer existed. But it would also be nice if there were no more war in the world" (Antonia, 28, w).

In addition to past experiences, individual character traits and users' social environments emerged as being relevant for both first- and second-level perceptions of hate speech. For example, a 23-year-old student stood out among the participants due to his commitment to equal rights, interest in political discussions, and awareness of hate speech. Whereas other participants of his age group were less attentive to hate speech content, he demonstrated relatively extreme concern and highlighted hate speech's adverse effects on society. Likewise, a 60-year-old participant was noticeably upset and shared insights from his personal experience, which may explain his emotional reaction to hate speech: "These [posts] make me notice how it [outrage] surges inside me. And that's what drove me to burnout. Such disdain. Such devaluations of people. For me, that's the worst" (Georg, 60, m).

The role of presentation form

Regardless of content, images attracted more attention both quantitatively and qualitatively, which explains why visual hate speech was recognized more often than textual hate speech (i.e. first-level perceptions). When we asked participants about their first impression of the social media platform, they frequently recalled images. Regarding posts containing visual hate speech, most participants identified the visual design as being the primary reason for the salience of the posts, especially when they presented direct hate speech. The extremity of such content was also emphasized, "You see a swastika like that—it may sound stupid—but you don't see it every day" (Markus, 28, m). By comparison, indirect hate speech memes were seldom noticed, though it later became clear that the memes were not fully understood or not seen as being hateful (see next section).

Unlike visual posts, hate speech within user comments was recognized only by a minority of participants, meaning that second-level perceptions rarely came to light. The reason most often stated for the lack of recognition was that comments are usually disregarded when scrolling through social media feeds, especially if the primary post's topic is not of interest, but also to intentionally avoid harmful content. One participant described, "It's a habit of mine [to avoid comments] . . . because for me they are totally devoid of content and often also very negative. They're just stupid. I don't read them anymore" (Luisa, 46, w). Along similar lines, posts consisting exclusively of text were rarely read completely and only when the first words piqued users' interest owing to, for example, their recency or personal relevance. On top of that, participants were more likely to be aware of hate speech in textual form when conspicuous features such as capital letters or unusual words were used.

Concerning second-level perceptions, visual hate speech again seemed to exert a more substantial influence on participants than textual hate speech. As described, the content was often intentionally ignored as soon as the hateful intent was recognized. However, because images could be grasped far more readily, they quickly aroused emotional reactions and lingered in participants' minds because they had been "burnt in" (Daniel, 42, m). Participants especially described being shocked by visual hate speech: "That's . . . phew . . . blatantly bad" (Carolin, 42, w) and "That knocked my socks off" (Alexander, 37, m). However, the decision to report hate speech depended less on its presentation form than on its content-related characteristics.

The role of content-related characteristics

We observed the most considerable differences in first- and second-level perceptions of hate speech depending on its directness. Whether hate speech was recognized at all primarily hinged on its presentation form. However, directness was also a decisive factor. Indeed, some indirect hate speech was noticed but not perceived as being incivil or problematic by some participants. That dynamic was particularly true for humorous indirect hate speech targeting women that some men but also some women seem to have enjoyed: "Some may feel attacked, but I think it's quite funny" (Annette, 28, w). We also noted that some of the posts with indirect hate speech in our stimulus material were not understood due to unknown meme designs or the fact that some participants simply did not know specific terms (e.g. "pride flag"). In those cases and others, indirect hate speech often was skimmed over and perceived to be harmless. However, evidence suggests that it may influence audiences subliminally, as one participant also suspected,

> Let's say this was actually my feed, and I would read through it . . . and in between, there would always be something like this. . . . Subconsciously, it would have an effect on me. Subconsciously, it will do something to me. (Lars, 23, m)

Moreover, when elaborating on the content and recognizing the intent behind the indirect hate speech, a few participants rejected it even more strongly. One participant described the sequence of her reactions toward a hate meme: "Here I laughed briefly, because I didn't really read the text but only saw the picture. . . . And then I read the text and thought to myself, ‘Ugh. Okay. Nah’” (Maria, 34, w). Being more unambiguously incivil, direct hate speech incited stronger emotional reactions, particularly outrage and disgust. In very few cases, direct hate speech resulted in a mirroring of aggression: “I’m outing myself now, but I really think that he [the poster] should be punched in the face” (Georg, 60, m). Those emotional reactions also transferred to participants’ reporting intentions. According to statements in the interviews, direct hate speech would have been reported more often, whereas many participants were unsure about the reporting of indirect hate speech. Participants often considered the hate speech’s relevance in criminal law, which for them was a clear, objective indication of whether posts should be removed from social media platforms, sometimes regardless of how they personally felt about the content: “I personally think that it’s horrendous . . . but I think that it probably doesn’t violate any law” (Josephine, 22, w). However, as participants stated, the intention to intervene with counterspeech depended more on one’s individual situation, although they also highlighted that it would be more likely if they knew the perpetrator.

Discussion

The results of our qualitative multi-method study highlight the relevance of considering both the context and individual users’ characteristics when examining hate speech. Combining observations of users’ behavior in response to a simulated social media feed, self-confrontation techniques, semi-structured qualitative interviews, and problem-focused tasks, our exploratory investigation of 23 German-speaking social media users with various backgrounds produced valuable insights into how different forms of hate speech on social media are perceived. To gain a holistic picture, we examined both participants’ first-level perceptions (i.e. recognition of hate speech) and second-level perceptions (i.e. attitudes and opinions toward it).

Overall, our results emphasize that the question of whether hate speech is recognized at all is fundamental. If social media users do not see hate speech, or do not identify it as such, then they obviously cannot intervene against it. It became apparent during our study that direct and visual hate speech is particularly conspicuous in first- and second-level perceptions and lingers in users’ memories, whereas textual and/or indirect hate speech often remains under the radar, which confirms assumptions of the picture superiority effect (Stenberg, 2006). Likewise, user comments were rarely read, and hate speech within them thus went unrecognized. Concerning second-level perceptions, visual and direct hate speech was evaluated as being more drastic than textual and indirect forms. Likewise, hate speech relevant in criminal law was perceived as being more severe as well. That result is consistent with past findings indicating that direct hate speech containing insulting language or threats of violence is perceived as being more hateful (Kenski et al., 2020; Stryker et al., 2016). Due to our study design, we could not assess whether indirect hate speech might indeed become more impactful when it occurs with consonance and high frequency, as Paasch-Colberg et al. (2021) have suggested. However, some participants who recognized the hidden hateful intent behind indirect hate speech after a more intensive elaboration perceived it in a particularly negative way and reasoned about its subliminal effects. Whether hate speech would be reported depended heavily on whether the content was considered to be punishable. Otherwise, the effort did not seem to be worthwhile.

Based on our data, we cannot make any clear-cut conclusions about gender's influence on perceptions of hate speech. Personal attitudes, however, considerably influenced first- and second-level perceptions of hate speech (Wojatzki et al., 2018). Moreover, younger and more frequent social media users among our participants tended to react less strongly to hate speech content, which could indicate tendencies of desensitization and normalization (Bilewicz and Soral, 2020; Santos et al., 2020). Such normalization carries the risk of becoming less involved with unpleasant content in general or even withdrawing from political discussion (Barnidge et al., 2019). However, given the nature of our data and that age and social media usage were somewhat confounded in our sample, we were unable to untangle which factor is more decisive in that regard. Nevertheless, it can be concluded that the perception of hate speech is a highly individual matter and thus cannot be traced back fully to sociodemographic characteristics, the content of hate speech, or presentation form. In agreement with Bormann et al.'s (2021) findings, our results suggest that perceptions of incivility and, more specifically, hate speech are heavily dependent on the specific (social media) context, normative attitudes, past experiences with hate speech, and individual users' characteristics. That finding also corresponds with the problem of gaining reliable annotations of hate speech when studying incivility with automated text analysis methods, for each annotator has different thoughts about what is more or less hateful (Benikova et al., 2018).

Limitations and directions for future research

Several limitations of our study warrant consideration. First, the results are limited to our qualitative sample of 23 German-speaking social media users. However, considering our research's focus, it was necessary to concentrate on individual experiences and reactions to clarify whether and how different hate speech content is perceived in realistic social media environments. Future (quantitative) research could build on our findings by experimentally investigating the influence of certain factors. Second, studies investigating the perception of hate speech likely depend on the participants' home country due to specific national laws, which are somewhat more restrictive in Germany, especially relative to other democratic countries (Hawdon et al., 2017). In a similar way, the cultural background of the participants' home country, which is historically unique in Germany, undoubtedly influences the perception of hate speech. Therefore, future research should investigate different legal regulations and country-specific cultural differences and their role in hate speech perceptions. Third, because we worked with a fictitious social media feed, our findings cannot entirely be transferred to real-world situations of social media usage characterized by a more personalized, more social experience. Nevertheless, because the feed was fully operational and perceived as being authentic, it was at least possible to study perceptions of hate speech in a more natural way than in past research. Due to ethical considerations, we told participants beforehand that the content on the platform was not moderated or restricted by regulations; thus, participants could have expected to be confronted with more extreme content. However, as our results show, even that priming did not induce participants to recognize all of the hate speech integrated into the feed. Fourth, qualitative research settings are vulnerable to social desirability effects, especially when examining sensitive topics such as hate speech. In our study, all interviews were conducted by a young woman, which may have further increased social desirability, particularly concerning perceptions of sexist hate speech.

Conclusion

Our findings highlight the importance of examining individual and contextual perceptions of hate speech, which have implications for its dissemination and subsequent effects. As our study has shown, some forms of hate speech simply do not seem to be recognized by most users. However, targeted users may especially notice and be affected by it. According to our results, images and humorous elements seem to be the ideal means of spreading hate speech through social media. Accordingly, programs to foster media literacy should emphasize those indirect forms and call attention to how they can be recognized, understood, and effectively eradicated.

Acknowledgement

We would like to thank the Minerva Center for the Rule of Law under Extreme Conditions at the University of Haifa, which provided valuable suggestions that contributed to the final version of this manuscript. Moreover, we would like to thank Julian Unkel for his help with creating the stimulus for this study.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the German Federal Ministry of Education and Research (BMBF) under grant number 13N15340 (KISTRA project).

ORCID iDs

Ursula Kristin Schmid https://orcid.org/0000-0002-1892-002X

Anna Sophie Kümpel https://orcid.org/0000-0001-7184-4057

Diana Rieger https://orcid.org/0000-0002-2417-0480

Notes

1. Further information on the particularities of the German law can be found in the study's Open Science Framework (OSF) repository: https://osf.io/vsed9/. See also Paasch-Colberg and Strippel (2021: 7).
2. Using the "Qualitative Preregistrations" template provided by the OSF, we preregistered our study's aims, RQs, design, and information about the process of data collection and analysis before initiating data collection. The preregistration, along with the interview guide, can be accessed at https://osf.io/vsed9/registrations. Because this article focuses on the perceptions of ordinary social media users, the results of the interviews with police officers are not reported here.
3. We were provided with reported hate speech from the Democracy Center Baden-Württemberg, a German center for education, service, and networking in the fields of extremism, preventive education work, and human rights.

References

Åkerlund M (2021) Dog whistling far-right code words: the case of “culture enricher” on the Swedish web. *Information, Communication & Society*. Epub ahead of print 21 August. DOI: 10.1080/1369118X.2021.1889639.

Barnidge M, Kim B, Sherrill LA, et al. (2019) Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics and Informatics* 44: 101263.

Ben-David A and Matamoros-Fernández A (2016) Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication* 10: 1167–1193.

Benikova D, Wojatzki M and Zesch T (2018) What does this imply? Examining the impact of implicitness on the perception of hate speech. In: Rehm G and Declerck T (eds) *Language Technologies for the Challenges of the Digital Age*. Cham: Springer, pp: 171–179.

Bilewicz M and Soral W (2020) Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology* 41(Suppl. 1): 3–33.

Billig M (2001) Humour and hatred: the racist jokes of the Ku Klux Klan. *Discourse & Society* 12: 267–289.

Borgeson K and Valeri R (2004) Faces of hate. *Journal of Applied Sociology* 21(2): 99–111.

Bormann M, Tranow U, Vowe G, et al. (2021) Incivility as a violation of communication norms—A typology based on normative expectations toward political communication. *Communication Theory*. Epub ahead of print 6 October. DOI: 10.1093/ct/qtab018.

Brown A (2018) What is so special about online (as compared to offline) hate speech? *Ethnicities* 18(3): 297–326.

Coe K, Kenski K and Rains SA (2014) Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64(4): 658–679.

Common Sense (2018) Social media, social life: teens reveal their experiences. Available at: https://www.commonsensemedia.org/research/social-media-social-life-2018 (accessed 19 October 2021).

Filibeli TE and Ertuna C (2021) Sarcasm beyond hate speech: Facebook comments on Syrian refugees in Turkey. *International Journal of Communication* 15: 24.

Gelber K and McNamara L (2016) Evidencing the harms of hate speech. *Social Identities* 22(3): 324–341.

Geschke D, Klaßen A, Quent M, et al. (2019) #Hass im Netz: Der schleichende Angriff auf unsere Demokratie. eine Bundesweite repräsentative Untersuchung [#Hate on the net: the creeping attack on our democracy. A nationwide representative survey]. Available at: https://blog.campact.de/wp-content/uploads/2019/07/Hass_im_Netz-Der-schleichende-Angriff.pdf (accessed 9 November 2020).

Haßler J, Maurer M and Oschatz C (2019) What you see is what you know: the influence of involvement and eye movement on online users’ knowledge acquisition. *International Journal of Communication* 13: 3739–3763.

Hawdon J, Oksanen A and Räsänen P (2017) Exposure to online hate in four nations: a cross-national consideration. *Deviant Behavior* 38(3): 254–266.

Henry N and Powell A (2018) Technology-facilitated sexual violence: a literature review of empirical research. *Trauma, Violence, & Abuse* 19(2): 195–208.

Johnson NF, Leahy R, Restrepo NJ, et al. (2019) Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature* 573(7773): 261–265.

Kenski K, Coe K and Rains SA (2020) Perceptions of uncivil discourse online: an examination of types and predictors. *Communication Research* 47(6): 795–814.

Leets L (2002) Experiencing hate speech: perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues* 58(2): 341–361.

Leonhard L, Rueß C, Obermaier M, et al. (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.

Lim SS (2002) The self-confrontation interview: towards an enhanced understanding of human factors in web-based interaction for improved website usability. *Journal of Electronic Commerce Research* 3: 162–173.

Masullo Chen G and Lu S (2017) Online political discourse: exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media* 61(1): 108–125.

Matamoros-Fernández A (2017) Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20(6): 930–946.

Matamoros-Fernández A (2018) Inciting anger through Facebook reactions in Belgium: the use of emoji and related vernacular expressions in racist discourse. *First Monday* 23(9): 9.

Matamoros-Fernández A and Farkas J (2021) Racism, hate speech, and social media: a systematic review and critique. *Television & New Media* 22(2): 205–224.

Mayring P (2015) *Qualitative Inhaltsanalyse: Grundlagen und Techniken* [Qualitative Content Analysis: Basics and Techniques]. 12th ed. Beltz Verlag. Available at: https://www.beltz.de/fachmedien/sozialpaedagogik_soziale_arbeit/buecher/produkt_produktdetails/27650-qualitative_inhaltsanalyse.html (accessed 5 October 2021).

Meibauer J (2013) Hassrede—Von der Sprache zur Politik [Hate speech from language to politics]. In: *Hassrede. Interdisziplinäre Beiträge zu einer Aktuellen Diskussion* [Hate Speech. Interdisciplinary Contributions to a Current Discussion]. Gießener Elektronische Bibliothek, pp. 1–16. Available at: http://geb.uni-giessen.de/geb/volltexte/2013/9251/pdf/HassredeMeibauer_2013.pdf

Mondal M (2017) A measurement study of hate speech in social media. In: *Proceedings of HT '17*, Prague, Czech Republic, 4–7 July. New York: ACM.

Nakamura L (2014) "I WILL DO EVERYthing that am asked": scambaiting, digital show-space, and the racial violence of social media. *Journal of Visual Culture* 13(3): 257–274.

Ohme J and Mothes C (2020) What affects first- and second-level selective exposure to journalistic news? A social media online experiment. *Journalism Studies* 21(9): 1220–1242.

Paasch-Colberg S and Strippel C (2021) "The boundaries are blurry . . .": how comment moderators in Germany see and respond to hate comments. *Journalism Studies* 23(2): 224–244.

Paasch-Colberg S, Trebbe J, Strippel C, et al. (2021) Insults, criminalization, and calls for violence: forms of hate speech and offensive language in German user comments on immigration. In: Monnier A, Boursier A and Seoane A (eds) *Cyberhate in the Context of Migrations. Postdisciplinary Studies in Discourse*. Cham: Palgrave Macmillan, pp. 137–163.

Pacheco E and Melhuish N (2018) Online hate speech: a survey on personal experiences and exposure among adult New Zealanders. Available at: https://www.netsafe.org.nz/wp-content/uploads/2019/11/onlinehatespeech-survey-2018.pdf (accessed 19 October 2021).

Powell TE, Boomgaarden HG, De Swert K, et al. (2015) A clearer picture: the contribution of visuals and text to framing effects. *Journal of Communication* 65(6): 997–1017.

Prochazka F, Weber P and Schweiger W (2018) Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies* 19(1): 62–78.

Rafael S (2021) Was wirkt gegen Hate Speech?—Erfahrungen aus über 10 Jahren zivilgesellschaftlicher Arbeit der Amadeu Antonio Stiftung [What works against hate speech?—Experiences from over 10 years of civil society work by the Amadeu Antonio Foundation]. In: Wachs S, Koch-Priewe B and Zick A (eds) *Hate Speech—Multidisziplinäre Analysen Und Handlungsoptionen: Theoretische Und Empirische Annäherungen an Ein Interdisziplinäres Phänomen* [Hate Speech—Multidisciplinary Analyses and Options for Action: Theoretical and Empirical Approaches to an Interdisciplinary Phenomenon]. Wiesbaden: Springer Fachmedien, pp. 339–351.

Rieger D, Kümpel AS, Wich M, et al. (2021) Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media + Society* 7(4): 20563051211052904.

Santos S, Amaral I and Simões R (2020) Hate speech in social media: perceptions and attitudes of higher education students in Portugal. In: *Proceedings of INTED2020 conferencem Valencia*, 2–4 March 2020, pp. 5681–5686. Valencia: IATED.

Schmid UK, Obermaier M and Rieger D (2022, May) *Who cares? Political characteristics that predict online counterspeech as civic participation against online hate speech*. Paper to be presented at the 72nd Annual Conference of the ICA, Paris, France.

Schmitt JB, Harles D and Rieger D (2020) Themen, Motive und Mainstreaming in rechtsextremen Online-Memes [Themes, motives, and mainstreaming in far-right online memes]. *Medien & Kommunikationswissenschaft* 68(1–2): 73–93.

Schulze H, Hohner J and Rieger D (in press) Soziale Medien und Radikalisierung [Social media and radicalization]. In: Rothenberger L, Krause J, Jost J, et al. (eds) *Terrorismusforschung—Interdisziplinäres Handbuch Für Wissenschaft Und Praxis* [Terrorism Research—interdisciplinary Handbook for Science and Practice]. Baden-Baden: Nomos.

Stenberg G (2006) Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology* 18(6): 813–847.

Steppat D (2021) Hate Speech Forsa-Studie 2021. Zentrale Untersuchungsergebnisse [Hate speech Forsa Study 2021: key findings]. Available at: https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2021_Praesentation.pdf

Stryker R, Conway BA and Danielson JT (2016) What is political incivility? *Communication Monographs* 83(4): 535–556.

Wojatzki M, Horsmann T, Gold D, et al. (2018) Do women perceive hate differently: examining the relationship between hate speech, gender, and agreement judgments. In: *Proceedings of the 14th conference on natural language processing (KONVENS 2018)*, Vienna, 19–21 September and 2 October.

Wolter D (2020) The Network Enforcement Act 2020 reloaded—improved as well? Available at: https://www.kas.de/en/kurzum/detail/-/content/the-network-enforcement-act-2020-reloaded-improved-as-well

Woodzicka JA, Mallett RK, Hendricks S, et al. (2015) It's just a (sexist) joke: comparing reactions to sexist versus racist communications. *HUMOR* 28(2): 289–309.

Zannettou S, Caulfield T, Blackburn J, et al. (2018) On the origins of memes by means of fringe web communities. arXiv:1805.12512 [cs]. Available at: http://arxiv.org/abs/1805.12512 (accessed 4 November 2020).

Zhang Z and Luo L (2019) Hate speech detection: a solved problem? The challenging case of long tail on Twitter. *Semantic Web* 10(5): 925–945.

Ziegele M, Naab TK and Jost P (2020) Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society* 22(5): 731–751.

Author biographies

Ursula Kristin Schmid (MA, LMU Munich) is a Research Associate in the Department of Media and Communication at LMU Munich, Germany. Her research focuses on digital media effects, specifically on the perception of (humorous) hate speech and online incivility as well as counter speech. Contact: ursula.schmid@ifkw.lmu.de

Anna Sophie Kümpel (PhD) is an Assistant Professor for Digital Media and Methods at TU Dresden, Germany. Her research focuses on a broad spectrum of media uses and effects, with most of her recent projects dealing with the use, dissemination, and perception of news and political information in algorithmically curated online environments. Contact: anna.kuempel@tu-dresden.de

Diana Rieger (PhD, University of Cologne), is Associate Professor in the Department of Media and Communication at LMU Munich, Germany. Her current work addresses characteristics and effects of hate speech, extremist online communication, and counter voices (e.g. counter narratives and counter speech). Furthermore, she focuses on entertainment research, investigating meaningful media content, for example, how meaning is portrayed in movies or in online content, for example, memes. Contact: diana.rieger@ifkw.lmu.de