The Communication Review

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gcrv20

Contextures of hate: Towards a systems theory of hate communication on social media platforms

Niklas Barth, Elke Wagner, Philipp Raab & Björn Wiegärtner

To cite this article: Niklas Barth, Elke Wagner, Philipp Raab & Björn Wiegärtner (2023) Contextures of hate: Towards a systems theory of hate communication on social media platforms, The Communication Review, 26:3, 209-252, DOI: 10.1080/10714421.2023.2208513

To link to this article: https://doi.org/10.1080/10714421.2023.2208513

THE COMMUNICATION REVIEW
2023, VOL. 26, NO. 3, 209–252
https://doi.org/10.1080/10714421.2023.2208513

Routledge
Taylor & Francis Group

Contextures of hate: Towards a systems theory of hate communication on social media platforms

Niklas Barth [a,b], Elke Wagner[c], Philipp Raab[c], and Björn Wiegärtner[c]

[a]Institut für SoziologieLudwig-Maximilians-Universität München Konradstr, 6, München 80801; [b]Kulturwissenschaftliches Institut Essen (KWI) Goethestraße, 31, Essen 45128; [c]Institut für Politikwissenschaft und SoziologieWittelsbacher Platz 1 Universität, Würzburg 97074

ABSTRACT

We inquire into different perspectives and patterns of problematizing online hate speech within the social sciences from a systems-theoretical perspective. Our results identify five different research perspectives adopted by studies on the issue: (1) systematic perspectives on problems of operationalizing (online) hate speech; (2) intentionalist perspectives on actors and their motives; (3) consequentialist perspectives on victims of online hate speech; (4) perspectives on media affordances, infrastructures, and strategies of online hate speech; and finally, (5) normative perspectives on the consequences of online hate speech. Additionally, we want to propose a functionalist perspective on hate communication and, for this purpose, develop a systems-theoretical and media-sociological framework for analyzing online hate speech. A systems-theoretical perspective connects to a process-oriented paradigm of doing hate speech. Instead of asking what hate speech is, a systems-theoretical framework focuses on how different communicative contextures empirically produce different understandings of hate communication. We will make four research proposals: We will (1) conceptualize hate as hate communication, then proceed to (2) analyze different communicative contextures, (3) develop media archeology of negation and conflict communication, and finally (4) focus on the function of conflict and hate communication for the emergence of (counter-)publics.

KEYWORDS

Hate speech; literature review; platform studies; social media; systems theory

Introduction

The advent of a digital public was celebrated as the *Pentecost of telematics*. Within the medium of social networks and platforms, the world would shrink spatially and temporally into a "global village" (McLuhan), scattered partial publics would reunite, and new forms of public speech would not only boost democratic inclusion and participation but also expand and improve opportunities for consensus, solidarity, and the integration of conflicts (Rheingold,

CONTACT Niklas Barth niklas.barth@soziologie.uni-muenchen.de Kulturwissenschaftliches Institut Essen (KWI) Goethestr, 31, Essen 45128

This text has been translated from German by John Koster and Stephan Elkins for *SocioTrans – Social Science Translation & Editing*.

2000). To date, these high hopes in the emancipatory potential of digital publics have met with disillusion: In public and in scholarly discourse, it is rather the dysfunctional phenomena entailed by the formation of digital publics that have become conspicuous. What we see is hate speech, radicalization, vulgarization, and brutalization of discourse, polemical and antagonistic communication, forms of insult and disparagement, as well as invectives and incivility. Digital publics that were once supposed to advance the integration of differentiated and plural societies into a rational unity have developed a mode of communication that has come down to being perceived as the irrational raging of a digital mob. They produce pathologies of public speech and undermine the foundations of democracy by polarizing and radicalizing conflicts. As a consequence, platforms increasingly appear to be media of decivilization (Koschorke, 2021).

The transformation of the public sphere has been widely discussed in debates on the role of digital publics. Jürgen Habermas continues to argue for a concept of the public sphere as an arena of rational discourse and dialogue. Digital publics release users from the "editorial guardianship of legacy media" (Habermas, 2021, p. 488; trans.[1]), as Habermas puts it in his *Reflections and Hypotheses on a New Structural Transformation of the Public Sphere*. At the same time, the "egalitarian and unregulated character" of public speech as well as its "emancipatory promise" are "drowned out today by desolate noises in fragmented echo chambers circling around themselves" (Habermas, 2021). Accordingly, the emancipatory and pathological consequences of digital public spheres are to be understood as complementary phenomena of a digital media evolution.

Media scholars have not only criticized the echo chamber thesis that Habermas is referring to (Bruns, 2019a, 2019b; Törnberg, 2022) but have also argued for a more differentiated concept of digital publics. As numerous studies have shown, there is no longer a singular public sphere, as the concept of the bourgeois public sphere assumes, but a heterogeneous constellation of different publics and different styles of communication emerging within these digital publics (Barth, 2020; Bruns & Burgess, 2011; Marwick & Boyd, 2010; Papacharissi, 2010, 2015; Wagner, 2019; Weller, Bruns, Burgess, Puschmann, & Mahrt, 2014; Bruns & Highfield, 2016). Undoubtedly, a pivotal feature in the transformation of digital publics is that platforms are currently providing the technological and media infrastructure for major parts of digital public communication (Bucher & Helmond, 2017; Gillespie, 2015; van Dijck, 2013). A key feature of platform infrastructures is that public communication is no longer unidirectional in the sense of one-to-many transmissions from editorial broadcasters to their media consumers. This mode has been superseded by a form of many-to-many communication within the plural public spheres of

[1]Hereinafter, original direct quotations translated into English will be tagged with the abbreviation *trans*.

decentralized networked platforms. Passive media consumers have taken now on the role of potentially active media producers ("produser," Bruns, 2009). As such, they claim the opportunity that digital media provide to speak for themselves, criticize, contradict, protest, form counter publics, and organize social movements. A consequence of this is that active media producers can fail to adapt content, topics, forms, and styles to the discourse-ethical standards of communication.

Empirical evidence shows the following: The US Anti-Defamation League (ADL) states in its report for the year 2021 that "forty-one percent of Americans said they had experienced online harassment over the past year, comparable to the 44% reported in last year's 'Online Hate and Harassment' report. Severe online harassment comprising sexual harassment, stalking, physical threats, swatting,[2] doxing[3] and sustained harassment also remained relatively constant compared to the prior year, experienced by 27% of respondents, not a significant change from the 28% reported in the previous survey" (https://www.adl.org/media/16219/download). Referring to the report by Shandwick, Tate, and KRC Research (2018), Chen, Muddiman, Wilner, Pariser, and Stroud (2019, p. 3) give even higher figures for the US: "More than 84% of U.S. adults say they've experienced incivility in online or offline life, and people report an average of 5.4 uncivil online encounters every week." And in 2019, users reported an average of 5.5 uncivil online encounters every week (Shandwick, Tate, & KRC Research, 2019). Likewise, in a 2021 quantitative study in Germany, about 76% of respondents said they had already been confronted with hate communication online; 39% had already had to deal with online hate very often, which is a new high compared to the previous year (i.e., 2020; Für Medien NRW, 2021).

Against this background, the communication practice of digital publics has not only become a subject of debates on the theory of democracy and the ethics of discourse but has also become a problem for normative regulation—for example, with regard to the question of what legal and actor status should be accorded to the platforms themselves. Should platforms have a stronger obligation to take editorial responsibility for the contents of public communication and to report it to state prosecutors? To what extent does this bear on freedom of speech? And to what extent, in turn, would this impinge on the right to privacy?

From a sociological point of view, these questions address the problem of the polycontexturality of modern society: What appears to be a legitimate public expression of opinion from one perspective meets antagonistic, disparaging, or discriminatory communicative acts from another. This study

[2] *Swatting* refers to faking an emergency just for fun to annoy the police and the person concerned, which is often a celebrity (our explanation added).

[3] *Doxing* refers to publishing personal data from someone on the net for his or her embarrassment (our explanation added).

provides a sociological literary review of the vast scholarly work on online hate speech by putting this problem of the polycontexturality of online hate speech at its core. The objective of this paper is to show how different scholarly perspectives on this subject of online hate speech address and highlight varying aspects and thus each constitute online hate speech as a different kind of problem. Additionally, this paper proposes a *functionalist perspective* on hate communication and, for this purpose, develops a systems-theoretical and media-sociological framework for analyzing online hate speech. A systems-theoretical perspective connects to a process-oriented paradigm of *doing hate speech*. Instead of asking what hate speech is, a systems-theoretical framework focuses on how different communicative contextures empirically produce different understandings of hate communication.

First, we will outline our data selection as well as methods and methodologies. Second, we will systematically differentiate five patterns of analyzing online hate speech. Third, we will discuss these conceptualizations of online hate speech from a functionalist systems-theoretical perspective. Fourth, we will derive four proposals for further research: They involve (1) conceptualizing hate as hate communication, (2) analyzing different communicative contextures, and (3) developing a media archeology of negation and conflict communication that (4) focuses on the function of conflict and hate communication for the emergence of (counter-)publics.

Material and methods

This sociological literature review does not undertake a methodologically rigorous discourse analysis, nor have we conducted a quantitative full-scale meta-analysis of scholarly work on online hate speech. Methodologically, we employ a systems-theoretical perspective and examine scientific communication on the phenomenon of online hate speech on social media platforms. We want to analyze the communicative modes of problematizing the phenomenon and not the phenomenon itself. We conduct a "second-order observation" (Luhmann, 1990, p. 68 ff.) of how scholarly work observes the phenomenon of online hate speech. From the viewpoint of a systems-theoretical epistemology, observing any given phenomenon requires *drawing a distinction*. These distinctions not only represent but constitute the object they observe. This being the case, different distinctions produce different perspectives on the object. From a second-order perspective, these perspectives become comparable as different *contextures*. Our aim is to identify and differentiate types and patterns of social-scientific problematizations of hate communication (Nassehi & Saake, 2002). When identifying *modes of problematization*, we refer to functionalist terminology and specifically inquire *how* a phenomenon is communicatively dealt with as *a problem* (Luhmann, 1970; Nassehi, 2008). This results in the following main research questions: (1) How does scholarly literature

frame online hate speech as a distinct type of problem? (2) What questions are typically and repeatedly asked in the existing literature? (3) Which modes of problematization are left out? For example, do scholars focus on the conceptual problems of defining the problem? Or do they plead for normative regulation or state that empirical data is missing to better understand the phenomenon? After reconstructing these modes of problematization within the framework of a functionalist approach, we typify them according to an inductive process aligned with grounded theory (Glaser & Strauss, 1967).

In selecting the texts, we proceeded as follows: On the one hand, we have selected papers that have been referred to in the debate time and again; these are, so to speak, *canonical* papers within the scholarly work on the issue. On the other hand, we focused on *recent* papers published between 2017 and 2022. We have included publications from German- as well as English-speaking countries. In particular, we searched relevant databases, such as Jstor, Web of Science, Semantic Scholar, and Google Scholar. Our research applied the following keywords: hate speech, incivility, hate crime, hate communication, abusive speech, online harassment, online hate, toxicity. The following journals were revisited in detail: *Pragmatics and Society*; *Journalism*; *Studies in Communication and Media*; *Big Data & Society*; *Journal of Hate Studies*; *Media, Culture & Society*; *Lodz Papers in Pragmatics*; *New Media & Society*; *Social Media + Society*; and *Journal of Communication*. The final sample comprises about 200 papers. In joint data sessions, we repeatedly compared and systematized the results of our research. To make it clear: This is not a systematic literature review. What we did is follow the arguments within the social-science discourse concerning the topic of hate speech on social media platforms. In that sense, we applied no clear-cut exclusion criteria. We pursued our research question by searching for papers that contained our keywords listed above.

Our literature review identifies five patterns of problematizing online hate speech. (1) systematic perspectives on problems of conceptually operationalizing online hate speech; (2) intentionalist perspectives on actors and their motives and strategies; (3) consequentialist perspectives on victims of online hate speech; (4) perspectives on media affordances, infrastructures, and strategies; and finally, (5) normative perspectives on the consequences of online hate speech.

Results

Systematic perspectives: what is online hate speech?

First and foremost, systematic perspectives have problematized the semantic blurriness and polysemy of the concept of "online hate speech." Systematic reviews in this vein have been published on scholarly work in legal and

communication studies (Paz, Montero-Díaz, & Moreno-Delgado, 2020), from the perspective of critical race studies (Bliuc, Faulkner, Jakubowicz, & McGarty, 2018; Matamoros-Fernández & Farkas, 2021), with regard to algorithms (Poletto, Basile, Sanguinetti, Bosco, & Patti, 2021; Yin & Zubiaga, 2021), with a view to (technical) interventions against online hate (Alkomah & Ma, 2022; Demilie & Salau, 2022; Windisch, Wiedlitzka, & Olaghere, 2021), with regard to children (Kansok-Dusche et al., 2022), and from a social-psychology perspective (Castaño-Pulgarín, Suárez-Betancur, Vega, & López, 2021). Moreover, the complexity of the endeavor to systematically capture online hate speech has attracted attention in various disciplines in the social sciences, such as communication studies (Boromisza-Habashi, 2013; Frischlich, Boberg, & Quandt, 2019; Keller & Askanius, 2021; Klein, 2017; Paz, Montero-Díaz, & Moreno-Delgado, 2020; Quandt, 2018), political science (Brown, 2018; Gelber, 2021), history (Goldberg, 2017), philosophy (Frick, 2017), criminology (Williams & Burnap, 2015), anthropology (Pohjonen & Udupa, 2017), psychology (Hellsten, Crespi, Hendry, & Fermani, 2021), computational social sciences (Cinelli et al., 2021; Cinelli, De Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021; Jigsaw, 2021), as well as theater studies (Bachmann, 2019) and literary studies (Wagner-Egelhaaf, 2020).

Online and offline hate speech

As critical race studies have emphasized, the concept of hate speech owes its existence to its specific historical context as well as genuine experiences of discrimination and oppression (for an overview, see Matsuda, 1989; Walker, 1994; Brown, 2018 on the legal background, see Eickelmann, 2017). Against this backdrop, the question arises to what extent hate speech practices differ in terms of whether they occur online or offline (Brown, 2018). Judith Butler has already posed the fundamental problem in linguistic terms of defining hate speech *offline*: "To decide the matter of what is a threat or, indeed, what is a word that wounds, no simple inspection of words will suffice" (Butler, 1997, p. 13). In Butler's perspective, words, beyond their use and their history, have no intrinsic characteristic that could explain their hurtful power. Yet even including the contexts of abuse does not suffice for adequately coming to grips with the phenomenon, according to Butler: "But the circumstances alone do not make the words wound" (Butler, 1997). What remains unexplained, she argues, is why some expressions are more easily detached from their power to wound than others.

Accordingly, Sponholz has attempted to differentiate the concept of hate speech further from the perspective of speech act theory in order to develop an explanatory framework. She hereby refers to the origin of the term *hate speech* in critical race theory and concludes that, if this distinction or demarcation were not made, "the power definition or the essence of hate speech as a form of discrimination would be erased because 'hate on the internet' can impact

anyone" (Sponholz, 2020, p. 63; trans.). Sponholz concludes: "However, a term created to give a name to the symbolic subjugation of historically oppressed groups ('hate speech') cannot simply be used to refer to a problem that can affect anyone (online harassment). Here, scholarship needs to be more precise in its conception and use of the everyday expression if it wants to produce insight" (Sponholz, 2020). The operationalization of the phenomenon "online hate speech" thus proves to be not only epistemologically ambiguous. It also turns out that the concept must be situated within complex power relations. Within that context, Katherine Gelber (2021) has elaborated "a systemic discrimination approach to defining hate speech" (p. 407) that focuses on power relations, institutional scripts and speaking positions that enable online hate speech. Such systematic perspectives deal with the problem of whether concepts of online hate speech are sensitive to power relations within asymmetric institutional contexts of communication and whether they are suitable to capture and explain systemic forms of exclusion, subordination, and discrimination.

Operationalizing online hate speech

The concepts of "hate speech" (Burch, 2018; Jakubowicz et al., 2017; Klein, 2012; Lumsden & Morgan, 2017; Oksanen, Räsänen, & Hawdon, 2014; Peterson & Densley, 2017), "extreme speech" (Udupa & Pohjonen, 2019) "cyber" or "online hate" (Hawdon, Costello, Barrett-Fox, & Bernatzky, 2019; Olson, 2020), "online aggression" (Stahel & Weingartner, 2019), "incivility" (Papacharissi, 2004; Su et al., 2018), "toxicity" (Jigsaw, 2021), "online extremism" (Hawdon, Costello, Barrett-Fox, & Bernatzky, 2019), "cyberviolence" (Peterson & Densley, 2017), and "cyber racism" (Bliuc, Faulkner, Jakubowicz, & McGarty, 2018) tend to focus on group-based hate communication. Nevertheless, it remains not only unclear what phenomenona these different concepts are supposed to denote but also how these concepts can be differentiated from each other.

Pars pro toto for a problem that a vast amount of scholarly literature diagnoses, the concept of "hate speech lacks unique, discriminative features" (Zhang & Luo, 2019) and for this very reason is not only difficult to identify theoretically but also difficult to operationalize methodically and methodologically. Therefore, nuanced and rapidly growing multi-disciplinary research fields have addressed the operationalization and definition problem of online hate speech (Meibauer, 2014; see also Baider, 2020; Chetty & Alathur, 2018; Gelber, 2021; Sponholz, 2018; Eickelmann, 2017; Keipi, Oksanen, & Räsänen, 2017; Paasch-Colberg & Strippel, 2021; Castaño-Pulgarín, Suárez-Betancur, Vega, & López, 2021; Chen, Muddiman, Wilner, Pariser, & Stroud, 2019; Fortuna & Nunes, 2019 or earlier, Duffy, 2003).

Thorsten Quandt has developed a concept for the definition of online hate speech that is widely used within communication studies (Boberg, Schatto-

Eckrodt, Frischlich, & Quandt, 2018; Frischlich, Boberg, & Quandt, 2019). Quandt systematizes the blurriness of the concept along the lines of "(a) wicked actors, (b) sinister motives and reasons for participation, (c) despicable objects/targets, (d) intended audience(s), and (e) nefarious processes/actions" (Quandt, 2018; see also Quandt, 2021 and in reference to Quandt: Westlund, 2021; De Vreese, 2021). To give another example, Paasch-Colberg, Strippel, Trebbe, and Emmer (2021) identify five modes of communication of online hate speech as "racist othering" (us-against-them rhetoric), "racist criminalization" (immigrants as a threat), "dehumanization," "raging hate" (demanding physical violence against immigrants), and/or finally the "call for hate crimes" (p. 176 f.).

For the concept of *toxicity*, the following definition has been proposed: Toxicity characterizes "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" (Jigsaw, 2021). This, however, immediately raises the question as to what a "rude" comment is. Or regarding the incivility model: "Given this, we define incivility as features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics" (Coe, Kenski, & Rains, 2014, p. 660). But again, what is a "disrespectful tone"? To answer this question, Coe et al. refer to a number of studies (Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2013; Borah, 2012; Brooks & Geer, 2007; Carter, 1998; Papacharissi, 2004) that suggest that civil forms of discourse are indispensably linked to mutual respect.

Accordingly, hate speech is problematized as a form of communication in which individuals or groups of individuals are defamed on the basis of a collective, ascribed, and stigmatizing group characteristic (e.g., ethnicity, sex, religious affiliation, sexual orientation; Castaño-Pulgarín, Suárez-Betancur, Vega, & López, 2021; Hawdon, Oksanen, & Räsänen, 2017; Höntzsch, 2020; Sponholz, 2020; Wachs & Wright, 2019; Zhang & Luo, 2019; Álvarez-Benjumea & Winter, 2018). In contrast, the terms cyberbullying and cyberstalking in particular are used to describe individual strategies of communication (Hawdon, Costello, Barrett-Fox, & Bernatzky, 2019). The boundary between hate speech and, for example, cyberstalking is thus drawn primarily according to the distinguishing criterion of the attributed group reference of victims of hate communication (individual vs. collective/groups). The aim of cyberbullying is seen as directly harming the victims as individuals, to threaten or embarrass them, in order to exclude them socially (Ang, 2015; Hellsten, Crespi, Hendry, & Fermani, 2021; Peterson & Densley, 2017, p. 9).

Intentionalist perspectives: actors, motives, and strategies of hate speech

Intentionalist conceptions of online hate speech ask about the actors involved. Who typically spreads hate online? What motives guide the perpetrators? These questions come into view from popular scientific (Nagle, 2017),

theoretical (Douglas, 2016), and methodologically quantitative (Artime, 2016; Borgeson & Valeri, 2004; Mathew, Dutt, Goyal, & Mukherjee, 2019; Ozalp, Williams, Burnap, Liu, & Mostafa, 2020) as well as qualitative perspectives (Erjavec & Poler Kovačič, 2012; Lumsden & Morgan, 2017; Marwick & Caplan, 2018).

Actors: Of unemployed men, right-wing radicals, misogynists, trolls, bots, states

Artime (2016) identifies "unemployed, unmarried men" as the primary group of actors in comment sections (p. 8). In this regard, the *troll* has been established as the paradigmatic social figure for explaining online hate speech. Trolls are conceptualized as individuals who get enjoyment from confronting other (groups of) people with hate speech and by destroying and sabotaging discursive routines (for an overview, see Lumsden & Morgan, 2017). "Trolls just want to have fun," as, for example, Erin Buckels, Trapnell, and Paulhus (2014) explain with reference to the sadistic personality profiles of those engaging in such activity. Additionally, some scholars have elaborated on various actors involved in online hate speech and considered troll on grounds of their political and ideological orientation. For example, perpetrator profiles have been analyzed with focus on individuals expressing right, right-populist, and right-wing radical fascist views (see, e.g., Fielitz & Thurston, 2019). Their practices have been classified within broader conceptualizations such as "digital fascism" (Fielitz & Marcks, 2019), "cyber fascism" (Griffin, 2000), or "broadband terrorism" (Feldman, 2009). Accordingly, staunchly anti-feminist positions (Drüeke & Klaus, 2014) and/or anti-Islamist views have also been identified (Alcántara-Plá & Ruiz-Sánchez, 2017).

From an intentionalist perspective, scholarly work on online hate speech largely assumes that human actors are immediate or indirect causal agents. Kim, Graham, Wan, and Rizoiu (2019, p. 3) have also tagged states (e.g., Russia) as an additional group of actors to which they ascribe an interest in disrupting and subverting Western public discourses (Kim et al. also refer to Badawy, Ferrara, & Lerman, 2018; Broniatowski et al., 2018). Using tweets from the 2020 U.S. presidential election campaign, Chang, Chen, Zhang, Muric, and Ferrara (2021) have examined the emergence of social bots as a group of actors. They suggest that social bots will continue to be deployed prominently in online debates. (Chang, Chen, Zhang, Muric, & Ferrara, 2021; see also Gorodnichenko, Pham, & Talavera, 2018; Zhuravskaya, Petrova, & Enikolopov, 2020; Yang et al., 2019; Zelenkauskaite & Balduccini, 2017; Ferrara, Varol, Davis, Menczer, & Flammini, 2016; Aiello, Deplano, Schifanella, & Ruffo, 2012). Online hate speech, then, cannot simply be attributed to young, white (unemployed) males. What becomes visible is a disparate mix of actors, technological sets of instruments (bots), and institutions (states).

Motives for online hate speech: Ideologies and psychopathologies

Intentionalist perspectives examine the motives and reasons that animate individuals and/or groups to actively participate in various forms of hate communication (Burch, 2018; Jikeli, Cavar, & Miehling, 2019; Kopytowska & Baider, 2017; Oksanen, Räsänen, & Hawdon, 2014; Sorokowski, Kowal, Zdybek, & Oleszkiewicz, 2020). These contributions to the research prioritize uncovering individual psychological motivations or group dynamics. Studies in this vein highlight individual characteristics such as vengefulness, anger, and sadism (Beckert & Ziegele, 2020; Buckels, Trapnell, & Paulhus, 2014; Craker & March, 2016; Sest & March, 2017)

Just as numerous and heterogenous as the perpetrator profiles encountered in research are the motives that can be identified. While the above-mentioned studies argue in terms of individual psychology, some also refer to the motivational force of group dynamics as an explanatory variable (Wachs, Wettstein, Bilz, and Gámez-Guadix (2022). Important insights for the internal differentiation of ideological motive structures of online hate speech can be found for "racism" (Ben-David & Matamoros-Fernández, 2016; Bliuc, Faulkner, Jakubowicz, & McGarty, 2018; Chaudhry & Gruzd, 2020; Cohen, Holt, Chermak, & Freilich, 2018; Faulkner & Bliuc, 2016, 2018; Matamoros-Fernández & Farkas, 2021; Matamoros-Fernández, 2017), for "Islamophobia" (Evolvi, 2019; Froio, 2018; Hanzelka & Schmidt, 2017; Horsti, 2017), and for hate speech based on sex/gender or sexual orientation (Dragiewicz et al., 2018; KhosraviNik & Esposito, 2018; Sobieraj, 2018).

Consequentialist perspectives: different groups of victims

While intentionalist perspectives discuss which actors generate online hate speech, consequentialist perspectives problematize the consequences of hate speech for different groups of victims. Oksanen, Räsänen, and Hawdon (2014, p. 25 f.) assume that differentiation of an ingroup and an outgroup is decisive for defining online hate speech. The boundary between "us" and "them" is thereby maintained through targeted communicative acts. Hate speech thus manifests itself as an insulting, exclusionary, intimidating, or discriminatory act of communication against a specific outgroup Schulzke (2016); Simpson (2013).

The victims accounted for are above all women (Chen et al., 2020; Drüeke & Klaus, 2014; Ganz & Meßmer, 2015; Pritsch, 2011; Ringrose, Harvey, Gill, & Livingstone, 2013; Sobieraj, 2018; Sundén & Paasonen, 2018), children and adolescents (Wiederer, 2003), the LGBTIQ+ community (Cleland, Magrath, & Kian, 2018), members of various ethnic groups (Boromisza-Habashi, 2013; Daniels, 2009; Holt, 2018; Kettrey & Laster, 2014; Miškolci, Kováčová, & Rigová, 2020; Ortiz, 2019; Park, 2017), members of various religions (Muslims: Evolvi, 2018, 2019; Horsti, 2017; Froio, 2018; Farkas, Schou, &

Neumayer, 2018; Zempi & Awan, 2016; Hanzelka & Schmidt, 2017; Awan, 2016; Jews: Hardy & Chakraborti, 2020; Jikeli, Cavar, & Miehling, 2019; Schwarz-Friesel, 2019), and disabled people (Burch, 2018).

Attempting to exemplify this perspective on the victims of online hate speech, it has been noted in (feminist) scholarly work on the relationship between gender and digitality that the Internet is a space of free development, especially with regard to identity issues and politics: "Such spaces have proven to be valuable sources of solidarity for those from disadvantaged groups" (Sobieraj, 2018, p. 1,701). At the same time, however, a countermovement can be observed: "In terms of gender inequality, digital publics are rife with male resistance to women's involvement in public life. There is a steady drumbeat of sexism directed at many women who participate in public discourse" (Sobieraj, 2018, p. 1,701).

Kenski, Coe, and Rains (2020) point out that online hate speech is perceived differently, leading to correspondingly different victim stories and different ways of dealing with online hate: "Taken as a whole, the results related to the research question demonstrate that an audience's perceptions of uncivility are not likely to be uniform. Scholars and practitioners cannot assume that all audience members will interpret an uncivil act in the same way" (p. 809). This finding of internal differentiation in the perception of hate speech is explained from a psychologistic and causal analytic perspective. What is observed are the "individual characteristics that influence perceptions of incivility" (Kenski, Coe, and Rains (2020)). A study by Costello, Hawdon, Bernatzky, and Mendes (2019) shows that the probability of becoming a victim of symbolic violence online can depend on specific factors. Factors increasing the likelihood of seeing oneself as a victim of hate speech are, for example, the length of time spent online, the number of different sites and Internet services used, and one's belonging to ethnic minorities. We want to emphasize one particular assumption that Costello et al. express in the context of their study, namely, their finding of differences regarding the reception of hate online: "Many people view hate materials without experiencing negative consequences, and most hate messages do not directly advocate violence" (Costello, Hawdon, Ratliff, & Grantham, 2016, p. 312). In this regard, Costello et al. refer to various previous studies that have produced similar results (Douglas, McGarty, Bliuc, & Lala, 2005; Gerstenfeld, Grant, & Chiang, 2003; Glaser, Dixit, & Green, 2002; McNamee, Peterson, & Peña, 2010). This again shows that (hateful) appellation (online) does not have to be perceived and processed as such. It can also be ignored, in which case hate often simply becomes an act of communication no longer succeeded by connecting acts of communication. Nevertheless, consequences of online hate speech certainly can be identified, such as the generation of fear, the loss of social trust, the promotion of discrimination and extreme attitudes, and ultimately the stoking of violence and terror (Costello, Hawdon, Ratliff, & Grantham, 2016, p. 312).

Perspectives on media affordances, infrastructures, and strategies

Media infrastructures and practices are problematized in the context of online hate speech from at least three perspectives. (1) A first perspective problematizes the genesis, dissemination, and promotion of online hate speech via platform architectures, business models, and media infrastructures. (2) A second perspective problematizes how the identification of acts of communication as online hate speech is possible via technological devices (e.g., algorithms as regulatory machines [Katzenbach, 2018], buttons/flagging [Crawford & Gillespie, 2016; Kalch & Naab, 2017]). (3) A third perspective elaborates on the media practices that result from platforms as contexts of communication.

Media affordances and technological infrastructures of hate speech

The importance of affordances for the usage of media has been highlighted as the fundamental issue in media theory and media sociology (for an overview, see Bucher & Helmond, 2017; for an introduction, see Gibson, 2015, 1982) and has also been stressed with regard to the use and processing of social media platforms (Gillespie, 2015). A seminal study on the media infrastructures of social media platforms has been conducted by Van Dijck and Poell (2013). They observe a "social media logic" consisting of four elements that complements the logic of mass media. These elements are "programmability, popularity, connectivity, and datafication" (Van Dijck and Poell (2013) p. 5). Accordingly, studies that view hate communication as a media-produced phenomenon problematize how a "network media logic" (Klinger & Svensson, 2015) increases the likelihood of online hate speech (Klein, 2012; Matamoros-Fernández, 2017; Sydnor, 2018). Here, media are conceptually understood as "virtual stages of hate" (Kilvington, 2021), as "toxic technocultures" (Massanari, 2017), "platformed racism" (Matamoros-Fernández & Farkas, 2021; Matamoros-Fernández, 2017), or "mediatized contempt" (Eickelmann, 2014, 2017). Hate speech is problematized as a "sociotechnical process" (Sponholz, 2021, p. 15; trans.), for instance, when "shitstorms" are made possible in the first place by digital media infrastructures (Gaderer, 2018; Stegbauer, 2018). Factors identified here include filter-bubble or echo-chamber effects and thus the self-reinforcing and radicalizing effects of network infrastructures on communication processes (Cinelli et al., 2021; Cinelli, De Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021). While reasonable arguments have been made that criticize the echo chamber thesis (Bruns, 2019a, 2019b; Törnberg, 2022), scholars have continued to differentiate contributing media infrastructural factors that promote online hate speech: for example, the anonymity of public communication (Black, Mezzina, & Thompson, 2016; Uth & Meier, 2018; Wekesa, 2019), the real-time nature

of communication, the length of communication processes on platforms as well as the number of comments (Cinelli, De Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021), the "context collapse" (Marwick & Boyd, 2010) of different contexts and publics mediated within the network, and the affective dynamics of digital publics (Stegbauer, 2018 ff.; Wagner, 2019).

Strategies: Agonistic publics

This perspective focuses on practices and strategies of online hate speech. What comes into view here are practices of trolling and flaming (McCosker, 2014; Morgan, 2022; Ortiz, 2020; Phillips, 2019), stereotyping (Sundén & Paasonen, 2018), rhetorics of devaluation, insult, name-calling, sexual harassment, as well as the communication of obscenities, slandering and defamation, threats, silencing (surveyed in Andersen, 2021; Anderson & Barnes, 2022), and aggressive lexis in general (Lingam & Aripin, 2017). These strategies and practices are understood as "in-your-face politics" (Mutz, 2015) and micro-aggressions that are applied strategically by users in comment sections to draw attention to topics in the first place and thus create public spheres.

On the basis of this understanding there is scholarly work that addresses the function of hate speech in forming agonistic digital publics (McCosker, 2014). Following McCosker (2014), Frances Shaw (2016) assumes that trolling and provocations on the Internet do not necessarily contribute to foreclosing and destroying public discourses. She uses the example of the Facebook and Instagram page "Bye Felipe," a feminist campaign that features screenshots of examples of online hate toward women to show how a feminist discourse space forms when provoked by misogynistic communication: "Understood in this way, provocation can provide opportunities for the articulation of political claims" (Shaw, 2016, p. 8). However, she argues that "conflict or provocation can also be productive of a discursive politics in which a political community is able to define itself in opposition to others" (Shaw, 2016). Thus, she claims, it is not only hate speech that comes into view on the webpages of the aforementioned campaign; it also becomes apparent that the confrontation between trolls, haters, and feminists can give rise to the formation of political alliances (e.g., through the use of hashtags) and the articulation of political demands. This shows that hate on the Internet can be morally, legally, and ethically reprehensible but may nonetheless prove quite functional for the emergence of public spheres. Hate comments foster debates insofar as they promote the connectivity of communication. These acts of communication do not even have to refer directly to one another. The public sphere that is created in this way does not constitute a space of rational discourse in the sense of Habermasian discourse theory. Hate comments do, however, generate public spheres to the extent that antagonistic parties are able to connect and contend with one another (Wagner, 2019; see also Rega & Marchetti, 2021 for political

discourse in Italy as well as Bratslavsky, Carpenter, & Zompetti, 2020; Montez & Brubaker, 2019 for U.S. presidency discourse).

Technological identification of online hate speech

Algorithms have emerged as technologies for detecting and sorting out online hate speech (Gorwa, Binns, & Katzenbach, 2020; Katzenbach, 2016, 2018). Thus, the ever-growing field of *computational social sciences* is devoted to the question of detecting and combating hate speech by means of self-learning and automated programs (Fortuna & Nunes, 2019; Jigsaw, 2021; Jurgens, Hemphill, & Chandrasekharan, 2019; Schmidt & Wiegand, 2017; Van Aken, Risch, Krestel, & Löser, 2018; Vidgen et al., 2019; Waseem, Davidson, Warmsley, & Weber, 2017).

MacAvaney et al. (2019) have shown from a computer science perspective that automated big data analyses also know limitations. In particular, automated algorithm-based methods for detecting hate speech systematically neglect "user intent and context" (MacAvaney et al. (2019) p. 13). In addition, the authors point out that algorithms inherently have an "interpretability problem – that is, it can be difficult to understand why the systems make the decisions that they do" (MacAvaney et al. (2019) p. 1). In a very similar vein, Gorwa, Binns, and Katzenbach (2020) also note with respect to platform-specific regulation and moderation algorithms that "these systems remain opaque, unaccountable and poorly understood" (p. 2). Bonilla and Rosa (2015) make a further methodological argument with respect to research on Twitter hashtags, which can also contribute to the spread of hate online. Even data-intensive hashtag ethnographies must be contextualized, otherwise they have no diagnostic value: "The only way to really know what these tweets were 'about' was to view them in the context of the individual tweeters themselves: when they were posting, what they had previously posted, who they had begun following, and what they were retweeting" (Bonilla and Rosa (2015) p. 7). Algorithms do indeed determine structures of relevance, yet it is not sufficient to establish algorithmic search operations for detecting online hate speech (Schmidt & Wiegand, 2017; Zhang & Luo, 2019). This is because algorithms cannot trace the context of communication offers. Data becomes information only in contexts; algorithms can therefore hardly function without context knowledge (see, e.g., Boyd & Crawford, 2012; Schmidt & Wiegand, 2017; Zhang & Luo, 2019). Therefore, significant swaths of the qualitative and particularly of the quantitative research point out that online hate speech needs to be studied in its differentiated and situated contexts, while at the same time noting that it is precisely this contextual data that is missing, thereby identifying this as a gap in the research record.

Normative perspectives: counternarratives, community moderation, normative regulations

Regulatory perspectives inquire into the consequences of online hate speech. These consequences can be differentiated according to whether they (1) are understood as secondary effects of online hate speech on other social fields, logics, practices, and discourses or whether (2) one asks what lessons can be drawn from the spread of online hate speech about how to combat it.

Co-evolution of hate speech and hate crime

The social-scientific discussion about online hate speech is particularly urgent, among other reasons, as it identifies the issue of hate speech and hate crime being connected to each other in a mutually constitutive and co-evolutionary relationship. On the one hand, research has been conducted on whether specific events (e.g., terror attacks) *trigger* hate communication (Scrivens et al., 2021 with references to Bliuc, Betts, Vergani, Iqbal, & Dunn, 2019; Kaakinen, Oksanen, & Räsänen, 2018; Miro-Llinares & Rodriguez-Sala, 2016; Williams, Burnap, & Sloan, 2017; Miro-Llinares, Moneva, & Esteve, 2018; Williams & Burnap, 2015; Burnap et al., 2014). On the other hand, online hate speech results in hate finding its way from the comment section to the streets (on this, see Marcks & Pawelz, 2022). In Germany, for example, acts of violence and terror attacks attributed to networking practices on social media have taken place in the cities of Halle, Hanau, and Bautzen (on Bautzen, see Laux & Schmitt, 2017). In the case of the Christchurch, New Zealand, attack, the terrorist broadcast a live video of the crime on the social media platform 8chan, commenting that he was now putting his "shitposting" into action (Williams, Burnap, Javed, Liu, & Ozalp, 2020). National and international studies can be found that point to the causal effects of online hate communication on offline *hate crimes* (Chan, Ghose, & Seamans, 2016; Müller & Schwarz, 2021). Schools in particular are the focus here, along with "the importance of teaching digital civility in schools" (Dishon & Ben-Porath, 2018; see also Simpson, 2019 for an overview, see Nisa & Setiyawati, 2019).

Consequences for theories of the public sphere and of democracy

At the same time, it is above all the negative consequences of online hate speech for the formation of the public sphere that are increasingly being discussed. From a normative point of view, hate communication can have grave consequences for liberal democracies (Papacharissi, 2004). What has been observed in this regard is first and foremost a (gradual) loss of trust in liberal civic institutions as well as an increase in intolerance (Goovaerts & Marien, 2020; Rossini, 2019). Rossini (2019) draws attention to the specific contexts of online hate speech: "Future research needs to shift away from the perception that incivility is intrinsically problematic. Instead, researchers

should further examine how different online platforms may constrain or facilitate expressions of intolerance to understand how platforms might mitigate these behaviors to prevent democratically harmful online discussions" (Rossini (2019) p. 18). Despite all the divergences in the social-scientific debates on online hate, one can identify a kind of mainstream discourse in this matter: Hostility to democracy is one of the important threat scenarios and thus an anchor point in the debate about online hate, which is taken up again and again – not least in contributions form prominent scholars (Habermas, 2021).

Identifying, deleting, and reporting hate speech: Moderating incivility

Users commenting under news articles as well as the moderation of these user comments by community managers has become established communicative practices in digital public spheres (Loosen & Schmidt, 2012; Reich, Domingo, Paulussen, Quandt, & Reich, 2011; Ziegele, 2016; Ziegele, Jost, Bormann, & Heinbach, 2018). With a view to the active combating of hate speech, community managers have appeared in editorial teams as a new and important group of actors. The task and function of these actors is to moderate or comment on the public communication streams on the platform and to delete posts that are flagged as "hate communication" and/or (in Germany) report them to the public prosecutor's office.

From the perspective of communication studies and journalism studies, a nuanced field of research has opened up on this topic in recent years (Boberg, Schatto-Eckrodt, Frischlich, & Quandt, 2018; Chen & Pain, 2017; Domingo, Domingo, Paulussen, Quandt, & Reich, 2011; Ferrucci & Wolfgang, 2021; Frischlich, Boberg, & Quandt, 2019; Gillespie, 2018; Megarry, 2017; Paasch-Colberg & Strippel, 2021; Reich, Domingo, Paulussen, Quandt, & Reich, 2011; Wintterlin, Schatto-Eckrodt, Frischlich, Boberg, & Quandt, 2020; Ziegele & Jost, 2020; Ziegele, Jost, Bormann, & Heinbach, 2018). Moderation practices are differentiated by function (e.g., "gatekeeping," Chen & Pain, 2017), by strategies, styles, and types ("interactive vs. non-interactive moderation," Boberg, Schatto-Eckrodt, Frischlich, & Quandt, 2018; "authority," Wintterlin, Schatto-Eckrodt, Frischlich, Boberg, & Quandt, 2020; Ziegele, Jost, Bormann, & Heinbach, 2018; see also section 3.1 above), as well as by factors that influence the selection criteria of moderation practices (such as individual reasons, work routines, or organizational-institutional and social-discursive variables, see Paasch-Colberg, Strippel, Laugwitz, Emmer, & Trebbe, 2020, p. 112; Shoemaker & Reese, 2014).

Boberg, Schatto-Eckrodt, Frischlich, and Quandt (2018) used Quandt's model to conduct an "automated content analysis" of hate in online comments sections. Their algorithmic determination of hate on the Internet is guided by Quandt's model of "dark participation" as well as by the deduction of terms of abuse from dictionaries. One of their conclusions is that "to understand

moderation decisions, further contextual factors must be considered" (Boberg, Schatto-Eckrodt, Frischlich, & Quandt, 2018, p. 64). Correspondingly, Frischlich, Boberg, and Quandt (2019) deduce from their qualitative-deductive content analysis, again based on Quandt's model, "that it was less community managers' experiences with but more so their subjective theories about (strategic) dark participation that shaped their moderation practices" (p. 2,028). The argument here is based on causalities and a deductive search for motives. The authors assume that "dark participation" (Quandt) exists and then ask about the internal editorial handling of it (on this deductive study design, see also Frischlich, Schatto-Eckrodt, Boberg, & Wintterlin, 2021; Wintterlin, Schatto-Eckrodt, Frischlich, Boberg, & Quandt, 2020; Ziegele, Jost, Bormann, & Heinbach, 2018; Ziegele, Naab, & Jost, 2019).

In this context, a study by Muddiman and Stroud (2017) has shown that user comments containing terms of abuse or offensive talk are more likely to be deleted. It has also been observed how "toxic talk" (Anderson, Yeo, Brossard, Scheufele, & Xenos, 2018) and forms of "incivility" in comment sections are directed not only against participants in debates but also against the institution of mass media (Anderson, Yeo, Brossard, Scheufele, & Xenos, 2018) as well as against journalists (Chen et al., 2020; Sarikakis et al., 2021), thus bringing about and fueling social radicalization. Normative regulation of toxic, incivil, or offensive talk requires operationalization of hate communication. Such operationalizations function, however, by making strong normative assumptions, as we have already elaborated above (Jigsaw, 2021; Kim, Guess, Nyhan, & Reifler, 2021; Anderson, Yeo, Brossard, Scheufele, & Xenos, 2018; Chen & Pain, 2017; Coe, Kenski, & Rains, 2014; Xiang, Fan, Wang, Hong, & Rose, 2012; for an overview, Schwertberger & Rieger, 2021, p. 56; see also section 3.1 above).

It becomes apparent that the significance of these new moderating gatekeepers for theories of democracy is a subject of controversial public and scholarly debates. Is it necessary to moderate the communication practices of these historically rather novel media producers until they have learned to implement a civilized communication practice (see Habermas, 2021)? Or do we have to conceptualize the filtering activities of community managers and algorithms as a form of non-state censorship (Gonçalves, Weber, Masullo, Torres da Silva, & Hofhuis, 2021; Guo & Johnson, 2020; Kalsnes & Ihlebæk, 2021)?

Normative regulations and counternarratives

Normative perspectives encounter different forms of legal regulation of online hate speech. From a legal perspective, the phenomenon of online hate speech is often framed as a problem of balancing freedom of opinion against freedom of speech (Pöyhtäri, 2014). The German Criminal Code specifies what is legally to be considered as hate crimes – namely, communicative forms of

expressing opinions (online and offline) that fall into the categories of, for example, sedition, defamation, libel, or slander – but fails to define the gray area of hate speech (Brugger, 2003a, 2003b). Baldauf, Ebner, and Guhl (2019) conclude, "The term 'hate speech' is the subject of public controversy. It is important to state that hate speech is not a legally specified category in Germany. The German Netzwerkdurchsetzungsgesetz [Network Enforcement Act] primarily targets hate crime and legally punishable misinformation. Various definitions exist for hate speech" (p. 7). The Network Enforcement Act obliges for-profit platforms in Germany to delete "manifestly unlawful content" after receiving a complaint and to report it to the public prosecutor's office (Tworek & Leerssen, 2019). The vagueness of the codification of "manifestly unlawful content" thus institutionalizes the problem of defining hate speech on a legal, institutional, and operational level, for example, in the practice of community moderators on social media platforms or on an institutional level (German Press Council, 2017; United Nations, 2020). To address this problem, the new legislative package for combating hate crime online attempts to deal with hate speech more systematically but also states: "Insults, malicious gossip and defamation are not covered by the reporting obligation, because it can sometimes be difficult to delimit them from those statements which are covered by the freedom of speech" (https://www.bmj.de/EN/FocusTopics/Legislative-package-combat-hate-hate-speech.html).

In addressing this, the literature refers to national or regional differences of classifying and regulating online hate speech (Groebel, Metze-Mangold, van der Peet, & Ward, 2001, p. 65). There is also what has been referred to as a "cyberhate divide" (Daniels, 2009, p. 176), which essentially describes the difference between the U.S. approach to hate speech, which is based on the protection of the freedom of speech enshrined in the First Amendment, and a "strong European stance against hate speech online" Daniels, 2009 p. 22). Hawdon, Oksanen, and Räsänen (2017) conclude in their quantitative study comparing the four countries USA, Finland, Great Britain, and Germany that such different approaches to regulation can in fact have an impact on the exposure of online hate speech. It is therefore not surprising that various commentators argue for a more uniform regulation of hate speech (for analog contexts, see Downing, 1999; Iganski, 1999; Gelber & McNamara, 2015; explicitly for digital contexts, see Beausoleil, 2019; Höntzsch, 2020). In this discussion, however, the question of the actual mechanisms of demarcation between legitimate/acceptable and illegitimate/unacceptable communication practices as well as of the actors involved in this demarcation often remains unanswered.

Complementing normative regulatory issues is the elaboration of counterstrategies and counternarratives to hate speech. These researchers ask about the possibilities of producing more civility and digital literacy (Pöttker, 2016, p. 352; Saputra & Al Siddiq, 2020). At the same time, ethical codes of conduct

have also been formulated (Prinzing, 2017, p. 340 f.). Porten-Cheé et al. have described different degrees of "online civic intervention" (Porten-Cheé, Kunst, & Emmer, 2020 citing among others Kalch & Naab, 2017). Katharina Benesch (2020) has developed a proposal for regulating online hate speech for online service providers (OSPs) and counternarratives and hashtag movements (summarized by Kuehn & Salter, 2020 with reference to Jakubowicz et al., 2017; Ray, Brown, Fraistat, & Summers, 2017 among others) have emerged as means of combating online hate speech.

Discussion: Towards a functionalist perspective on hate communication

Our literature review has identified five perspectives on online hate speech. The *systematic perspective* diagnoses the fuzziness of concepts and semantics of (online) hate speech, attempts to order the phenomenon of online hate speech conceptually, and tries to integrate it coherently into existing conceptual frameworks (3.1). The *intentionalist perspective* inquiries into the actors and their motives and intentions when uttering online hate speech (3.2). The *consequentialist perspective* focuses on differentiating victim groups and the consequences that they experience when they are confronted with online hate speech (3.3). The *normative perspective* examines above all the pathological and dysfunctional impact of online hate speech for political institutions, cultures, and values and inquires into strategies and modes of regulating, countering, and moderating online hate speech (3.5). Within these perspectives, online hate speech is often conceived as the result of intentions and causalities, for example, by narrowing the question down to the perpetrators' motive or the victims' experiences. Then again, online hate speech is analyzed by making strong theoretical-deductive assumptions about what online hate speech is and what it is not, as we have shown in our discussion of the concepts of "incivility" (Su et al., 2018), "toxicity" (Jigsaw, 2021), "negative speech" (Ben-David & Matamoros-Fernández, 2016), or "dark participation" (Quandt, 2018). In these studies, normative assumptions imply the dysfunctionality of online hate speech for digital publics. In our assessment, hate speech studies fail to meet the task of understanding the function of online hate speech in its full complexity when proceeding as described above. They usually define it away by referring to a deductive or normative definition. These attempts at conceptual operationalization narrow the problem to the question of what online hate is.

In contrast to *intentionalist* takes on online hate speech, *infra-structural perspectives* analyze the media logic of platforms and emphasize how media contexts and strategies can enable and hinder online hate speech (3.4). Scholarly work on media contexts and strategies addresses a blind spot within a wide range of hate speech studies when it applies perspectives that show us *how* online hate speech is practically constituted in different media contexts of

communication. Building on this idea of such *infra-structural perspectives*, we want to propose a functionalist systems-theoretical perspective on hate communication and on this basis develop a media-sociological framework for analyzing online hate speech (Barth & Wagner, 2021). A systems-theoretical perspective connects with the process-oriented paradigm of *doing hate speech* that infra-structural perspectives have employed. In our understanding, a "conceptual language of relations," as Goffman put it in his sociological study of stigma (Goffman, 1963/1986/1986, p. 11), and less a language of essentialisms is needed to understand online hate speech.

In the following, we outline four research proposals for a systems-theoretical perspective: (1) Such a perspective must conceptualize hate as hate communication, (2) analyze different communicative contextures, and 3) develop a media archeology of negation and conflict communication that (4) focus on the function of conflict and hate communication for the emergence of (counter-)publics.

(1) Hate as hate communication

From our everyday perspective, it is often the speaker's motives that supposedly guide us in understanding meaning: What did the speaker say? What did she mean exactly? From a feminist and critical language-theory perspective, however, it is the other way around: It should be the victim who decides whether a comment was insulting, invective, or hateful. The aforementioned questions already imply that a communicative act is attributed to a person only after the completion of the act.

A systems-theoretical perspective on online hate speech is not interested in singular speech acts, motives, intentions, or experiences of hate speech but in communicative processes. This perspective does not address the question of *what* online hate speech is but *how* communicative acts can be understood as hate communication within a social system. In this context, Niklas Luhmann distinguished three "thresholds of discouragement" (Luhmann, 1981, p. 124) of communication: 1) The improbability of an identical understanding of meaning, which follows from the operational closure and the related mutual non-transparency of psychic systems. 2) The improbability of reaching a greater number of recipients of communication beyond those present. 3) The improbability of successful communication, as "[e]ven if a communication is understood, there can be no assurance of its being accepted" (Luhmann, 1981). Solving "the problem of improbability in the process of communication" is the function of media (Barth, 2020, p. 32) Luhmann thus assigned the three basic problems to three types of media: 1) Language solves the problem of the improbability of understanding; 2) dissemination media (writing, print, mass media) solve the problem of reaching absent others; 3) success media or symbolically generalized communication media (e.g., power, love, truth, law, money) solve the problem of the

improbability of successful communication by eliciting the motivation to accept specific selections. Instead of treating communication as the *pursuit of understanding* or the *transmission of meaning* and thus as a solution to the problems of communication, systems theory conceives of communication itself as a problem calling for a solution. Social order becomes possible because meaningful patterns of the mutual adaption of behavior and coordination of action can be stabilized *in the process of* communication.

In the perspective that we are proposing here, we do not observe "hate" as an individual motive, nor as an emotion of resentment, hostility, or contempt but as a code of communication by which actors express, form, and simulate emotions or impute them to others and form expectations as to the consequences of their attributions for communication (Luhmann, 1986, p. 20). This implies a low-threshold and empirically open concept of hate communication that is interested in *how* communicative acts are *related* and *connected* to each other to facilitate the emergence of patterns of communication. Here, we explicitly draw on the concept of invective communication practices (Ellerbrock et al., 2017). This concept conceives of communication as hate communication when those involved in the communication attribute a polemogenous, transgressive, or "invective" (Ellerbrock et al., 2017) character to a communicative act and it is therefore understood and experienced as a form of disregard, devaluation, or discrimination: "The term [invectivity] shall center on all aspects of communication (verbal or nonverbal, oral, written, gestural or graphic) that are geared toward degrading, hurting, or marginalizing others." (Ellerbrock et al., 2017 trans.) This concept of invective communication connects extremely well with our concept of the polycontexturality of hate communication. It remains an open empirical question how, by whom, and in which context a breach of the boundaries of respect, and which ones, are characterized as a transgression. Yet, while this line of inquiry cannot help to find a standardized definition of hate communication for research, it does, however, make it possible to answer questions about the practical attribution of meaning to the phenomenon of "hate communication" within different communicative contextures.

(2) Analyzing differing communicative contextures

From a systems-theoretical perspective, modern society is differentiated into incommensurable systems (medical, legal, religious, scientific, economic, etc.) all of which simultaneously structure their communication processes according to their own functional logics (Luhmann, 2012). For example, from a political perspective, the problem arises whether the stronger political regulation of hate speech on social media platforms can be an opportune political issue to secure political majorities or whether engaging in hate speech itself could be an opportune political strategy to gain power. An economic perspective might, for instance, require considering the question if online hate

speech is a risk to the business model of a platform or perhaps stoking it promises to be a profitable business model itself. From a religious perspective, the attractiveness of hate speech can be related to questions about the sinful nature of human beings. From a normative-legal perspective, hate speech is a problem of balancing conflicting legal interests (freedom of speech versus personal rights) and integrating new laws into existing formal legal corpora. And from a scientific perspective, hate speech raises issues, for example, of methodologically ensuring that our concepts of hate speech actually measure the phenomenon in order to learn to distinguish true from false statements about it. From the perspective of the mass media, hate speech makes for a very good story and topic for reporting on hate in mass media as the conflicts and scandals associated with it generate high informational value. From a medical perspective, by contrast, users displaying persistent invective communication patterns might be seen as pointing to mental pathologies and behavioral disorders that may require medical consultation. From an aesthetic-artistic perspective, new forms of public speaking appear as a medium for the creation of new aesthetic forms and formats, for example, when parodic meme cultures form around figures like Pepe the Frog. And from a pedagogical perspective, unbridled hateful speech might draw attention primarily as a problem of a lack of socialization and a failure to meet the affectual standards of civilization. These brief examples show that contexts have their own logics and interrelationships with other contexts. This "simultaneity of the different" (Nassehi, 2006, p. 429; trans.) constitutes a polycontextual world. Every context is only a context within the context of other contexts and must therefore be examined as a *communicative contexture* (Günther, 1979). As a consequence, a functionalist perspective on hate communication does not ask what online hate is or what motives drive it. What matters from a systems-theoretical perspective is rather the *empirical contextures of hate communication* and the particular, specific *communicative connections that could always also be otherwise*. In this vein, an intended insult does not have to be understood and processed as such and responded to accordingly. It could always potentially be interpreted and responded to in ways that deviate from what had been intended. We therefore propose to focus on the communicative processes that reveal points at which different and incommensurable options for subsequent communication to connect with become available.

The framework sketched here suggests some questions for empirical research in media sociology and hate speech studies. The dichotomy of freedom of speech versus hate speech must be translated sociologically into the problem of multiple communicative contextures: What appears to ego as a legitimate public expression of opinion, represents an invective form of communication from the perspective of alter egos. For instance, one might look at disputes between platform moderators and the community of users to analyze the situated practice of that online community in terms of different

communicative contextures. Paasch-Colberg and Strippel (2021) have made an important contribution by taking an interest in the practical contexts of community moderators in defining online hate speech. To this end, they conducted 20 problem-focused interviews to understand the types, styles, forms, and functions of moderation practices in their genesis and thereby also shifted the focus to questions of the practical generation of online hate speech. They concluded: "With regard to our empirical findings, it is necessary to assume that comment moderators have different viewpoints on hate comments or hate speech. In some cases, they are even critical of these terms or reject them. Hate speech in user comments is a multifaceted and at times subtle phenomenon, which is why its identification and moderation demands high context sensitivity" (Paasch-Colberg and Strippel (2021) p. 17; see also Wagner, 2019). They went on to identify a need for further research, specifically for the analysis of *communicative contextures of hate*, as we want to call it.

To give another example, Sahana Udupa and Matti Pohjonen (2019) refer to the problem of the culturally relative meaning of "extreme speech" (p. 3051) from an ethnological perspective. They "argue that the production, circulation, and consumption of online vitriol should be approached as much as a cultural practice and social phenomenon as a legal or regulatory concept." They claim accordingly that "there is no self-evident category of hate speech" (Sahana Udupa and Matti Pohjonen (2019). From our perspective, however, this is a cultural argument with regard to online hate speech that puts its finger precisely on the problem of the polycontexturality of hate speech and the fact that, in a global society, hate speech operates within different communicative contextures. In contrast to such an approach, much scholarly work attempts to systematize definitions of online hate speech and, in so doing, at least implicitly treats the different perspective as being only a *theoretical problem* that could be solved by means of ever more specific definitions. As opposed to this line of thought, we believe it to be more fruitful to approach the different perspectives on hate speech as an *empirical problem*. When we propose to focus on the necessarily different empirical communicative contextures of online hate speech within different (media) contexts, this calls for operationalizing the ethnographic principle of "following the actor" within modern society. The key question then is, how is online hatred expressed and experienced in the situated practice of different actors within different media contexts? Empirically analyzing different contextures of hate communication within medially situated practices requires more qualitative studies as well as research in digital media ethnography.

(3) Media archeology of negation communication

Modern society is not only functionally differentiated but also socially in terms of different social actors and publics that observe each other as different cultures. Consequently, this leads to the differentiation of binding and

collectively shared moral values as well as latent cultural patterns in terms of what is considered as (dis)respectful (Luhmann, 1978; 2012/2013). From a systems-theoretical perspective, communication therefore represents a fundamentally risky undertaking. "Communication is risky" (Luhmann, 1995, p. 115) because the information communicated does not guarantee immediate understanding and positively connecting subsequent communication or even consent. In modern differentiated society, people are far more likely to oppose, deny, criticize, or even actively reject a communicative act. A systems-theoretical perspective focuses on the systematic risks of negation and conflict that is inherent in every communicative act.

From a media-archaeological perspective, one can ask how different dissemination media contexts enable and produce different probabilities of negation, conflict, as well as polemogenous and invective forms of communication at a specific point and time. Hate communication can be understood as a highly selective and strong form of negation and conflict communication that combines the negation of a communicative act with a subsequent act of invective communication. Here, we want to briefly outline how one might structurally distinguish three periods in the evolution of public communication media.

(1) In the salons of 17th- and 18th-century societies characterized by stratificatory differentiation, public communication took place in the form of *interaction among those present* and within the medium of orality. In such settings, modes of tactful communication were functional in terms of an interaction ethics to reduce the risk of negation and conflict in communication, as face-to-face interaction involves the risk of sparking conflict and turning public interaction into a conflict system in its own right (Kieserling, 1999, p. 257 ff.).
(2) The emergence of the media infrastructure of the *Gutenberg Galaxy* and printing facilitates at around the same time put *communication with those absent* on a broader social basis. If we think of the spread of heresies or the various forms of circulating pamphlets, printed communication enables a form of polemogenous communication that can take the form of criticism in substantive matters but can also be criticism ad hominem (Darnton, 1982). Deliberative publics and an associated ethics emerged to regulate such forms of critique.
(3) Today, social media platforms structure public communication in a highly selective way. Social media platforms combine the real-time nature of *communication among those present* with *communication with those absent* (Nassehi 2020, p. 124). They facilitate new forms of what we might call the *present absence* in digital public communication. An expanded communication radius corresponds with a reduced likelihood of achieving a consensus in communication (Luhmann, 1981, p. 31).

Thus, communication on digital platforms stands out systematically through a "particular dissemination and establishment of 'naysaying' in communication" (Nassehi, 2020, p. 122; trans.). Platforms as media infrastructures of digital publics structurally tend to develop polemogenous and antagonistic forms of communication, thus promoting the escalation of conflicts in communication. We therefore suggest combining systems theory with insights from media archeology (Barth, 2020). Today's media infrastructures must be placed within the context of other (historical) media infrastructures to fully understand the evolution of invective and polemogenous communication as an effect of different media contexts. Infrastructural perspectives on online hate speech have shown the importance of digital media-ethnographic analysis of platforms (see section 3.4 above). Van Dijk and Poell (2013) note: "The principles, mechanisms, and strategies underlying social media logic may be relatively simple to identify, but it is much harder to map the complex connections between platforms that distribute this logic: users that use them, technologies that drive them, economic structures that scaffold them, and institutional bodies that incorporate them" (p. 11). From a systems-theoretical perspective, these analyses should be combined with the empirical analysis of communication patterns within situated media contexts (Barth & Wagner, 2021). Further research could focus on (a) digital material infrastructures of platformed publics and their effects on communication (e.g. context collapses; differences between algorithmic newsfeed and interactive comment lists; multimediality of platforms, e.g. the analyzing the difference between text vs. image driven platforms in generating online hate speech); (b) elaborating the transformation of the factual, temporal and social dimension of platform communication; and (c) inquiring into the multimodality of communication (e.g. "small forms" (Balke, Siegert & Vogl 2021) of communication and their effect on public practices like commenting, "tweeting" or "posting;" many-to-many communication, triadic communication).

We want to shed some light on the very phenomenon of *triadic communication* as this mode of communication is, in our view, a highly relevant factor in the genesis of online hate speech. Within digital publics, communication between ego and alter ego often takes place before third parties, forming a triadic constellation (Nassehi, 2020, p. 124). Public communication is increasingly realized in the mode of "talking with others about others before others" (Nassehi, 2020) and indicates the circumstance that these others are not absent but *absent present* on the platform and can observe these communication processes. This specific mode of *triadic communication* combines the interactive logic of communication among those present with the logic of

printed communication with those who are absent. This process of communication is worth further investigation as moral communication bears considerable potential for "polemogenic" (Luhmann, 2008, p. 111) dynamics of mutual devaluation and disregard. Accordingly, hate communication as triadic communication is highly structured through the "invective triad of the one articulating an invective, the one subject to the invective, and the audience" (Ellerbrock et al., 2017, p. 9).

On the one hand, communicative techniques of conflict control based on interaction ethics do not work when people are subject to anonymous invectives on platforms. For example, audiences can stoke the conflict by teaming up with one of the opponents. On the other hand, the audience can also moderate existing conflicts in triadic communication. For example, if communication proceeds along the path of open moral disregard, then it creates its own incentives and motives to intensify the individual conflict as a group conflict (us vs. them) when disregard for an outgroup is itself experienced as respect and solidarity for the ingroup (Luhmann, 2008, p. 112). From a systems-theoretical perspective, communication takes on a moral nature when the "moralization of themes, symbols, structures, opinions, and expectations" is used to communicate "the conditions of respect and disregard" Luhmann, 2008 p. 108; trans.). The tricky thing is that communication about moral values develops its own polemical drive because moral communication "creates additional motives for feeling respected for 'maintaining one's position' or for 'punishing the other'" (ibid., p. 112; trans.).

As a consequence, the differentiation of digital publics results in different communicative styles that deviate from the deliberative model of public communication. Personal perceptions dominate communication within digital publics and constitute "affective publics" (Papacharissi, 2015). These new forms of digital publics do not lead to intellectual reasoning but to heated debates that we refer to as "intimate publics" (Wagner, 2019). Intimate publics feed an emotional and affective style of communication. Empirically, the wider effect of an emotional and affective style is not simply the demise of the public sphere but rather ongoing communication with regard to how communication interconnects (or disconnects). Often the subject matter of digital public speech is firsthand experience. These forms of authentic communication lay claim to emancipatory universalist ideals of bourgeois dialogic culture: Nearly everyone can participate in this style of discourse, not only those who are able to offer better arguments. But the consequence of affective communication in digital publics is antagonistic conflicts. Firsthand experiences, emotions, and affects cannot be refuted in the way (better) arguments can; they tend to offer opportunities for conflict and polemogenous forms of communication. For instance, communication is likely to result in conflict when *individual experience* is countered (e.g., the experience of gendered or racial discrimination) by referring to macro-statistical data.

(4) Function of conflict and hate communication

Unlike normative perspectives, a functionalist perspective is not only interested in online hate speech as a pathology of communication that undermines the deliberative model of public speech. It also inquires into the functional relationship between invective, agonistic, incivil, and even hateful forms of communication and the emergence of digital publics. In practical terms, this means that we propose not only to investigate the "extreme cases" but rather the "gray area" of everyday public communication that never gets to the stage of being evaluated under criminal law. In other words, we deem it necessary to focus on how media give rise to new implicit and tacit plausibilities for the emergence of "intimate publics" (Wagner, 2019). From a functionalist perspective, hate communication seems to establish itself as a (functional, not normative) solution to the problem of communicative connectivity (Barth & Wagner, 2021). Seen from this perspective, the troll is plainly the negative side of intimate publics.

Online hate speech is a highly "popular" form of communication (Döring et al., 2021; Van Dijck & Poell 2013) Why does "online hate speech" manage to be so connectable in the first place? From a system-theoretical perspective, hate speech represents a hyper-connective form of communication. On the one hand, the communication of hate is highly *performative*: Its *meaning* does not require much intricate interpretation. If a communicative statement is experienced *as hate speech*, the communicative context unambiguously changes into an asymmetrical form of (moral) disregard or hostility and is therefore produces highly selective communicative connections (e.g. "triggering" and instantaneously provoking affective counter reactions, insulting back (by the victim or the networked public).

On the other hand, the communication of hate is highly *informative*: As a violation of communication ethics and patterns of civilized dialogue, online hate speech generates high informational value so that the meaning of hate speech has to be interpreted within triadic communication networks and thus provokes subsequent communication (e.g. asking about the motives, backgrounds or causes of hate speech; negotiating the invective, commenting on the invective by others, defending the victim or showing solidarity with the victim/offender, rationalizing the behavior of the offender). It seems paradoxical: online hate speech is an unambiguous (performative) *and* an ambiguous (informative) form of communication at the same time. If the logic of social media is to ensure the "connectivity" (van Dijck, 2013) of communication within the network of communication, online hate speech is hyper-connective and therefore proves its functionality to platform contextures of communication. One way (performative) or the other (informative), hate speech does provoke communicative reactions and can temporarily set the center of attention within the decentralized flows of communication within the network.

From this it follows that the classical insight from *sociology of conflict* holds true for digital conflicts as well: Conflicts integrate because they reduce complexity, for example, by establishing clear-cut friend/enemy distinctions. Communicating conflicts is not a pathology of communication but a highly selective form of communication. As a strong form of conflict communication, hate communication does not disrupt communication processes but fosters them: "Conflicts serve to continue communication" (Luhmann, 1995, p. 398) by promoting possibilities of negation and therefore information. If we apply this perspective sociologically, it is for instance possible to identify that gendered digital hate speech does not always have to lead to victimization as various scholars discussed above have shown (Costello, Hawdon, Bernatzky, & Mendes, 2019; Costello, Hawdon, Ratliff, & Grantham, 2016, p. 312; Douglas, McGarty, Bliuc, & Lala, 2005; Gerstenfeld, Grant, & Chiang, 2003; Glaser, Dixit, & Green, 2002; Kenski, Coe, & Rains, 2020, p. 809; McNamee, Peterson, & Peña, 2010; Sobieraj, 2018, p. 1,701). Just like the practical meaning of the once derogatory term "queer" has been redefined and now stands for a non-heterosexual or non-cis-gender lifestyle, one can observe that progressive debates can also develop around incivil discourse. This demonstrates that trolling and provocation on the Internet do not automatically contribute to the prevention and/or destruction of specific discourses (McCosker, 2013; Shaw, 2016). Hate speech in digital publics can be morally, legally, and ethically reprehensible but may well prove to be functional nonetheless for the emergence of critical or "counterpublics" (Warner, 2002). Hate comments can promote debates insofar as they provide connecting points for subsequent communication – that is, they create communicative connectivity. The public sphere created in this way does not constitute a reasonable discourse space in the sense of Habermas' discourse theory. Hate comments, however, create a public sphere inasmuch as antagonistic parties can engage and grapple with one another there.

Conclusion

A study from a systems-theoretical perspective would inquire into (1) how communication is understood as invective communication within specific *communicative contextures* as well as within the situated practices of the defining actors, (2) which media infrastructures provide the possibilities for negation, conflict, and hate communication, (3) what function hate communication has for the emergence of digital publics, and (4) how patterns of communication stabilize when people express hate on the Internet. It remains an open empirical question how, by whom, and in which context a breach of the boundaries of moral respect, and which ones, is characterized as

a transgression as well as how, by whom, and in which context communication is understood as hateful.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Niklas Barth http://orcid.org/0000-0003-3188-1149

References

Aiello, L. M., Deplano, M., Schifanella, R., & Ruffo, G. (2012). People are strange when you're a stranger: Impact and influence of bots on social networks. *ArXiv*. doi:https://doi.org/10.48550/arXiv.1407.8134

Alcántara-Plá, M., & Ruiz-Sánchez, A. (2017). The framing of Muslims on the Spanish internet. *Lodz Papers in Pragmatics*, *13*(2), 261–283. doi:10.1515/lpp-2017-0013

Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information*, *13*(6), 273. doi:https://doi.org/10.3390/info13060273

Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. doi:10.1093/esr/jcy005

Andersen, I. V. (2021). Hostility online: Flaming, trolling, and the public debate. *First Monday*, *26*(3). doi:10.5210/fm.v26i3.11547

Anderson, L., & Barnes, M. (2022). Hate speech. In *The Stanford encyclopedia of philosophy (spring (2022 edition))*, E. N. Zalta (Ed.). https://plato.stanford.edu/archives/spr2022/entries/hate-speech/

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2013). The "nasty effect": Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. doi:10.1111/jcc4.12009

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, *30*(1), 156–168. doi:10.1093/ijpor/edw022

Ang, R. (2015). Adolescent cyberbullying: A review of characteristics, prevention, and intervention strategies. *Aggression and Violent Behavior*, *25*(1), 35–42. doi:10.1016/j.avb.2015.07.011

Artime, M. (2016). Angry and alone: Demographic characteristics of those who post to online comment sections. *Social Sciences*, *5*(4), 1–11. doi:10.3390/socsci5040068

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1–20. doi:10.5281/zenodo.58517

Bachmann, M. (2019). Hassmonologe. Die Stimme des Anderen im zeitgenössischen Dokumentartheater und -film. *Der Deutschunterricht*, *71*(5), 32–42.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. *ArXiv*. doi:https://doi.org/10.1109/ASONAM.2018.8508646

Baider, F. (2020). Pragmatics lost? Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society*, *11*(2), 196–218. doi:10.1075/ps.20004.bai

Baldauf, J., Ebner, J., & Guhl, J. (Eds.). (2019). *Hassrede und Radikalisierung im Netz. Der OCCI-Forschungsbericht*. London: ISD.

Balke, F., Siegert, B., & Vogl, J. (2021). *Kleine Formen. Archiv für Mediengeschichte* (Vol. 19, p. 8). Berlin: Vorwerk.

Barth, N. (2020). *Gesellschaft als Medialität. Studien zu einer funktionalistischen Medientheorie*. Bielefeld: Transcript.

Barth, N., & Wagner, E. (2021). Dinge als Medien denken. Was leistet eine funktionalistische Mediensoziologie. Peltzer, A. N. Zillien & M. Wieser (Eds.), *Medienjournal* Vol. *45*(1), Special Issue pp. 19–31.

Beausoleil, L. E. (2019). Free, hateful, and posted: Rethinking first amendment protection of hate speech in a social media world. *Boston College Law Review*, *60*(7), 2101–2144.

Beckert, J., & Ziegele, M. (2020). The effects of personality traits and situational factors on the deliberativeness and civility of user comments on news websites. *International Journal of Communication*, *14*, 3924–3945.

Ben-David, A., & 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.

Benesch, S. (2020). Proposals for improved regulation of harmful content online. Reducing online hate speech: Recommendations for social media companies and internet intermediaries, 247–306. https://ssrn.com/abstract=3686826

Black, E. W., Mezzina, K., & Thompson, L. A. (2016). Anonymous social media – Understanding the content and context of Yik Yak. *Computers in Human Behavior*, *57*, 17–22. doi:10.1016/j.chb.2015.11.043

Bliuc, A. -M., Betts, J., Vergani, M., Iqbal, M., & Dunn, K. (2019). Collective identity changes in far-right online communities: The role of offline intergroup conflict. *New Media & Society*, *21*(8), 1770–1786. doi:10.1177/1461444819831779

Bliuc, A. -M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75–86. doi:10.1016/j.chb.2018.05.026

Boberg, S., Schatto-Eckrodt, 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. doi:10.17645/mac.v6i4.1493

Bonilla, Y., & Rosa, J. (2015). #ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist*, *42*(1), 4–17. doi:10.1111/amet.12112

Borah, P. (2012). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research*, *41*(6), 809–827. doi:10.1177/0093650212449353

Borgeson, K., & Valeri, R. (2004). Faces of hate. *6*(2) *Journal of Applied Sociology/Sociological Practice*, *21*(2), 99–111. doi:10.1177/19367244042100205

Boromisza-Habashi, D. (2013). *Speaking hatefully: Culture, communication, and political action in Hungary*. University Park: Pennsylvania State Univ. Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Bratslavsky, L., Carpenter, N., & Zompetti, J. (2020). Twitter, incivility, and presidential communication: A theoretical incursion into spectacle and power. *Cultural Studies*, *34*(4), 593–624. doi:10.1080/09502386.2019.1656760

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A. . . . Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384. doi:10.2105/AJPH.2018.304567

Brooks, D. J., & Geer, J. G. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science*, *51*(1), 1–16. doi:10.1111/j.1540-5907.2007.00233.x

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. doi:10.1177/1468796817709846

Brugger, W. (2003a). The treatment of hate speech in German constitutional law (Part I). *German Law Journal*, *4*(1), 1–22. doi:10.1017/S2071832200015716

Brugger, W. (2003b). The treatment of hate speech in German constitutional law (Part II). *German Law Journal*, *4*(1), 23–44. doi:10.1017/S2071832200015728

Bruns, A. (2009). 'Anyone can edit': Vom Nutzer zum Produtzer. *kommunikation @ gesellschaft*, *10*, 1–23.

Bruns, A. (2019a). Are filter bubbles real?. Hoboken: Polity Press.

Bruns, A. (2019b). Filter bubble. *Internet Policy Review*, *8*(4), 1–14. doi:10.14763/2019.4.1426

Bruns, A., & Burgess, J. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In A. Bruns & P. De Wilde (Eds.) *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*. The European Consortium for Political Research (ECPR), United Kingdom (pp. 1–9). 10.14763/2019.4.1426

Bruns, A., & Highfield, T. (2016). Is Habermas on Twitter? Social media and the public sphere. In G. Enli, A. Bruns, A. O. Larsson, E. Skogerbo, & C. Christensen (Eds.), *The Routledge companion to social media and politics. Routledge, United States of America* (pp. 56–73).

Bucher, T., & Helmond, A. (2017). The affordances of social media platforms. In J. Burgess, T. Poell, & A. Marwick (Eds.), *The SAGE handbook of social media*. SAGE Publications. doi:10.4135/9781473984066.n14

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

Burch, L. (2018). 'You are a parasite on the productive classes': Online disablist hate speech in austere times. *Disability & Society*, *33*(3), 392–415. doi:10.1080/09687599.2017.1411250

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A. . . . Voss, A. (2014). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 1–14. doi:https://doi.org/10.1007/s13278-014-0206-4

Butler, J. (1997). *Excitable speech. A politics of the performative*. London: Routledge.

Carter, S. L. (1998). *Civility: Manners, morals, and the etiquette of democracy*. New York: Basic Books.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 1–7. doi:10.1016/j.avb.2021.101608

Chang, H. -C.H., Chen, E., Zhang, M., Muric, G., & Ferrara, E. (2021). Social bots and social media manipulation in 2020: The year in review. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of computational social science*. Routledge. doi:https://doi.org/10.48550/arXiv.2102.08436

Chan, J., Ghose, A., & Seamans, R. (2016). The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly*, *40*(2), 381–403. doi:10.25300/MISQ/2016/40.2.05

Chaudhry, I., & Gruzd, A. (2020). Expressing and challenging racist discourse on Facebook: How social media weaken the 'spiral of silence' theory. *Policy & Internet*, *12*(1), 88–108. doi:10.1002/poi3.197

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019, July-September). We should not get rid of incivility online. *Social Media + Society*, 1–5. doi:10.1177/2F2056305119862641

Chen, G. M., & Pain, P. (2017). Normalizing online comments. *Journalism Practice*, *11*(7), 876–892. doi:10.1080/17512786.2016.1205954

Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, *21*(7), 877–895. doi:10.1177/1464884918768500

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. doi:10.1016/j.avb.2018.05.003

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), 1–8. doi:10.1073/pnas.2023301118

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, *11*(1), 1–12. doi:https://doi.org/10.1038/s41598-021-01487-w

Cleland, J., Magrath, R., & Kian, E. (2018). The internet as a site of decreasing cultural homophobia in association football: An online response by fans to the coming out of Thomas Hitzlsperger. *Men and Masculinities*, *21*(1), 91–111. doi:10.1177/1097184X16663261

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. doi:10.1111/jcom.12104

Cohen, S. J., Holt, T. J., Chermak, S. M., & Freilich, J. D. (2018). Invisible empire of hate: Gender differences in the Ku Klux Klan's online justifications for violence. *Violence and Gender*, *5*(4), 209–225. doi:10.1089/vio.2017.0072

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*(3), 427–452. doi:10.1111/soin.12274

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*, 311–320. doi:10.1016/j.chb.2016.05.033

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, *102*, 79–84. doi:10.1016/j.paid.2016.06.043

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. doi:10.1177/1461444814543163

Daniels, J. (2009). *Cyber racism. White supremacy online and the new attack on civil rights*. Washington: Rowman & Littlefield.

Darnton, R. (1982). *The literary underground of the old regime*. Cambridge: Harvard University Press.

Demilie, W., & Salau, A. (2022). Detection of fake news and hate speech for Ethiopian languages: A systematic review of the approaches. *Journal of Big Data*, *9*(66), 1–17. doi:10.1186/s40537-022-00619-x

De Vreese, C. (2021). Beyond the darkness: Research on participation in online media and discourse. *Media and Communication*, *9*(1), 215–216. doi:10.17645/mac.v9i1.3815

Dishon, G., & Ben-Porath, S. (2018). Don't @ me: Rethinking digital civility online and in school. *Learning, Media and Technology*, *43*(4), 434–450. doi:10.1080/17439884.2018.1498353

Domingo, D., Domingo, D., Paulussen, S., Quandt, T., & Reich, Z. (2011). Managing audience participation. Practices, workflows and strategies. In J. B. Singer, A. Heinonen, A. Hermida, & M. Vujnovic (Eds.), *Participatory journalism: Guarding open gates at online newspapers* (pp. 76–95). Hoboken: Wiley-Blackwell.

Döring, J., Werber, N., Albrecht-Birkner, V., Gerlitz, C., Hecken, T., Paßmann, J., & Venus, J. (2021). Was bei vielen Beachtung findet: Zu den transformationen des Populären. *Kulturwissenschaftliche Zeitschrift*, *6*(2), 1–24. doi:10.2478/kwg-2021-0027

Douglas, D. M. (2016). Doxing: A conceptual analysis. *Ethics Information Technology*, *18*(3), 199–210. doi:10.1007/s10676-016-9406-0

Douglas, K. M., McGarty, C., Bliuc, A. -M., & Lala, G. (2005). Understanding cyberhate: Social competition and social creativity in online white supremacist groups. *Social Science Computer Review*, *23*(1), 68–76. doi:10.1177/0894439304271538

Downing, J. (1999). 'Hate speech' and 'first amendment absolutism' discourses in the US. *Discourse & Society*, *10*(2), 175–189. doi:10.1177/0957926599010002003

Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, *18*(4), 609–625. doi:10.1080/14680777.2018.1447341

Drüeke, R., & Klaus, E. (2014). Öffentlichkeiten im Internet: Zwischen Feminismus und Antifeminismus. *Femina Politica – Zeitschrift für feministische Politikwissenschaft*, *23*(2), 59–71. doi:10.3224/feminapolitica.v23i2.17614

Duffy, M. E. (2003). Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups online. *The Journal of Communication Inquiry*, *27*(3), 291–312. doi:10.1177/0196859903252850

Eickelmann, J. (2014). Mediatisierte Missachtung und die Verhandlung von Gender bei Empörungswellen im Netz. Der Fall Anita Sarkeesian. *onlinejournal kultur & geschlecht*, *13*, 1–19.

Eickelmann, J. (2017). *'Hate Speech' und Verletzbarkeit im digitalen Zeitalter*. Bielefeld: Transcript.

Ellerbrock, D., Koch, L., Müller-Mall, S., Münkler, M., Scharloth, J., Schrage, D., & Schwerhoff, G. (2017). Invektivität – Perspektiven eines neuen Forschungsprogramms in den Kultur- und Sozialwissenschaften. *Kulturwissenschaftliche Zeitschrift*, *2*(1), 2–24. doi:10.2478/kwg-2017-0001

Erjavec, K., & Poler Kovačič, M. (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. doi:10.1080/15205436.2011.619679

Evolvi, G. (2018). Hate in a tweet: Exploring internet-based Islamophobic discourses. *Religions*, *9*(10), 1–14. doi:10.3390/rel9100307

Evolvi, G. (2019). #islamexit: Inter-group antagonism on Twitter. *Information, Communication & Society*, *22*(3), 386–401. doi:10.1080/1369118X.2017.1388427

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

Faulkner, N., & Bliuc, A. M. (2016). 'it's okay to be racist': Moral disengagement in online discussions of racist incidents in Australia. *Ethnic and Racial Studies*, *39*(14), 2545–2563. doi:10.1080/01419870.2016.1171370

Faulkner, N., & Bliuc, A. M. (2018). Breaking down the language of online racism: A comparison of the psychological dimensions of communication in racist, anti-racist, and non-activist groups. *Analyses of Social Issues and Public Policy*, *18*(1), 307–322. doi:10.1111/asap.12159

Feldman, M. (2009, September 10). Broadband terrorism. A new face of fascism. *History & Policy*. http://www.historyandpolicy.org/opinion-articles/articles/broadband-terrorism-a-new-face-of-fascism

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. doi:10.1145/2818717

Ferrucci, P., & Wolfgang, J. D. (2021). Inside or out? Perceptions of how differing types of comment moderation impact practice. *Journalism Studies*, *22*(8), 1010–1027. doi:10.1080/1461670X.2021.1913628

Fielitz, M., & Marcks, H. (2019). Digital fascism. Challenges for the open society in times of social media. *CRWS Working Papers*, 1–25. https://escholarship.org/uc/item/87w5c5gp

Fielitz, M., & Thurston, N. (Eds.). (2019). *Post-digital cultures of the far right. Online actions and offline consequences in Europe and the US*. Bielefeld: Transcript.

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. doi:10.1145/3232676

Frick, M. -L. (2017). *Zivilisiert streiten. Zur Ethik politischer Gegnerschaft*. Ditzingen: Reclam.

Frischlich, L., Boberg, S., & Quandt, T. (2019). Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments. *Journalism Studies*, *20*(14), 2014–2033. doi:10.1080/1461670X.2018.1556320

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., & Wintterlin, F. (2021). Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication*, *9*(1), 195–208. doi:10.17645/mac.v9i1.3360

Froio, C. (2018). Race, religion, or culture? Framing Islam between racism and neoracism in the online network of the French far right. *Perspectives on Politics*, *16*(3), 696–709. doi:10.1017/S1537592718001573

Für Medien NRW, L. (2021). *Ergebnisbericht: forsa-Befragung zu: Hate Speech 2021*. https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html

Gaderer, R. (2018). Shitstorm. Das eigentliche Übel der vernetzten Gesellschaft. *ZMK Zeitschrift für Medien- und Kulturforschung Alternative Fakten*, *9*(2), 27–42. doi:10.28937/1000108173

Ganz, K., & Meßmer, A. -K. (2015). Anti-Genderismus im Internet. Digitale Öffentlichkeiten als Labor eines neuen Kulturkampfes. In S. Hark & P.-I. Villa (Eds.), *Anti-Genderismus: Sexualität und Geschlecht als Schauplätze aktueller politischer Auseinandersetzungen* (pp. 59–78). Bielefeld: Transcript.

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4), 393–414. doi:10.1080/13698230.2019.1576006

Gelber, K., & McNamara, L. (2015). The effects of civil hate speech laws: Lessons from Australia. *Law & Society Review*, *49*(3), 631–664. doi:10.1111/lasr.12152

German Press Council (2017). *German Press Code. Guidelines for journalistic work as recommended by the German Press Council – complaints procedure*. https://www.presserat.de/en.html

Gerstenfeld, P. B., Grant, D. R., & Chiang, C. -P. (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44. doi:10.1111/j.1530-2415.2003.00013.x

Gibson, J. J. (1982). Notes on affordances. In E. Reed & R. Jones (Eds.), *Reasons for realism. Selected Essays of James J. Gibson* (pp. 401–418). Hillsday, New Jersey, London: Lawrence Erlbaum Associates.

Gibson, J. J. (2015). *The ecological approach to visual perception* (*Classic edition* ed.). New York, London: Psychology Press.

Gillespie, T. (2015). Platforms intervene. *Social Media + Society*, *April–June*1–2. doi:10.1177/2F2056305115580479

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, London: Yale University Press.

Glaser, J., Dixit, J., & Green, D. P. (2002). Studying hate crime with the internet: What makes racists advocate racial violence? *The Journal of Social Issues*, *58*(1), 177–193. doi:10.1111/1540-4560.00255

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New Brunswick, London: Aldine.

Goffman, E. (1963/1986). *Stigma: Notes on the management of spoiled identity*. New York, London, Toronto: Simon & Schuster.

Goldberg, A. (2017). Minority rights, honor, and hate speech law in post-Holocaust West Germany. *Law, Culture and the Humanities*, *17*(2), 1–22. doi:10.1177/1743872117702822

Gonçalves, J., Weber, I., Masullo, G. M., Torres da Silva, M., & Hofhuis, J. (2021). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *New Media & Society*, 1–23. doi:10.1177/14614448211032310

Goovaerts, I., & Marien, S. (2020). Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communication*, *37*(6), 768–788. doi:10.1080/10584609.2020.1753868

Gorodnichenko, Y., Pham, T., & Talavera, O. (2018). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *NBER Working Papers*, 1–39. http://www.nber.org/papers/w24631

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15. doi:10.1177/2053951719897945

Griffin, R. (2000). Interregnum or endgame? The radical right in the 'post-fascist' era. *Journal of Political Ideologies*, *5*(2), 163–178. doi:10.1080/713682938

Groebel, J., Metze-Mangold, V., van der Peet, J., & Ward, D. (2001). *Twilight zones in cyber-space: Crimes, risk, surveillance and user-driven dynamics*. Bonn: Friedrich-Ebert Stiftung.

Günther, G. (1979). *Beiträge zur Grundlegung einer operationsfähigen Dialektik. Band 2*. Stuttgart: Meiner.

Guo, L., & Johnson, B. G. (2020). Third-person effect and hate speech censorship on Facebook. *April–June Social Media + Society*, *6*(2), 1–12. doi:10.1177/2056305120923003

Habermas, J. (2021). Überlegungen und Hypothesen zu einem erneuten Strukturwandel der politischen Öffentlichkeit. *Leviathan*, *49*(Sonderband 37), 470–500. doi:10.5771/9783748912187-470

Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, *11*(1), 143–160. doi:10.5281/zenodo.495778

Hardy, S. -J., & Chakraborti, N. (2020). *Blood, threats and fears. The hidden worlds of hate crime victims*. Cham: Palgrave Macmillan.

Hawdon, J., Costello, M., Barrett-Fox, R., & Bernatzky, C. (2019). The perpetuation of online hate: A criminological analysis of factors associated with participating in an online attack. *Journal of Hate Studies*, *15*(1), 157–181. doi:10.33972/jhs.166

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266. doi:10.1080/01639625.2016.1196985

Hellsten, L. M., Crespi, I., Hendry, B., & Fermani, A. (2021). Extending the current theorization on cyberbullying: Importance of including socio-psychological perspectives. *Italian Journal of Sociology of Education*, *13*(3), 85–110. doi:10.14658/pupj-ijse-2021-3-5

Holt, L. F. (2018). Dropping the "N-word": Examining how a victim-centered approach could curtail the use of America's most opprobrious term. *Journal of Black Studies*, *49*(5), 411–426. doi:10.1177/0021934718756798

Höntzsch, F. (2020). 'Hate Speech' als Gefahr für die Stabilität liberaler Demokratie oder: Warum 'Hate Speech' die individuelle Freiheit gefährdet. *Leviathan*, *36*, 181–196. doi:10.5771/9783748907565-181

Horsti, K. (2017). Digital Islamophobia: The Swedish woman as a figure of pure and dangerous whiteness. *New Media & Society*, *19*(9), 1440–1457. doi:10.1177/1461444816642169

Iganski, P. (1999). Legislating against hate: Outlawing racism and antisemitism in Britain. *Critical Social Policy*, *19*(1), 129–141. doi:10.1177/026101839901900102

Jakubowicz, A., Dunn, K., Mason, G., Paradies, Y., Mliuc, A. -M., Bahfen, N. , and Connelly, K. (2017). *Cyber racism and community resilience: Strategies for combating online race hate*. Cham: Palgrave Macmillan.

Jigsaw (2021). *Perspective: Using machine learning to reduce toxicity online*. https://www.perspectiveapi.com

Jikeli, G., Cavar, D., & Miehling, D. (2019). Annotating antisemitic online content. Towards an applicable definition of antisemitism. *ArXiv*. doi:https://doi.org/10.48550/arXiv.1910.01214

Jurgens, D., Hemphill, L., & Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In Association for Computational Linguistics, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3658–3666). 10.18653/v1/P19-1357

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. doi:10.1016/j.chb.2017.09.022

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *Studies in Communication and Media*, *6*(4), 395–419. doi:10.5771/2192-4007-2017-4-395

Kalsnes, B., & Ihlebæk, K. A. (2021). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society*, *43*(2), 326–342. doi:10.1177/0163443720957562

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2022). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*, *0*(0), 1–18. doi:https://doi.org/10.1177/15248380221108070

Katzenbach, C. (2016). *Die Regeln digitaler Kommunikation. Governance zwischen Norm, Diskurs und Technik*. Wiesbaden: Springer VS.

Katzenbach, C. (2018). Die Ordnung der Algorithmen: Zur Automatisierung von Relevanz und Regulierung gesellschaftlicher Kommunikation. In R. M. Kar, B. E. P. Thapa, & P. Parycek (Eds.), *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (pp. 315–338). Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS.

Keipi, M. N., Oksanen, A., & Räsänen, P. (2017). *Online hate and harmful content. Cross-national perspectives*. New York: Routledge.

Keller, N., & Askanius, T. (2021). Combatting hate and trolling with love and reason? A qualitative analysis of the discursive antagonisms between organized hate speech and counterspeech online. *Studies in Communication and Media*, *9*(4), 540–572. doi:10.5771/2192-4007-2020-4-540

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795–814. doi:10.1177/0093650217699933

Kettrey, H. H., & Laster, W. N. (2014). Staking Territory in the "World White Web". *Social Currents*, *1*(3), 257–274. doi:10.1177/2329496514540134

KhosraviNik, M., & Esposito, E. (2018). Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics*, *14*(1), 45–68. doi:10.1515/lpp-2018-0003

Kieserling, A. (1999). *Kommunikation unter Anwesenden. Studien über Interaktionssysteme*. Frankfurt a. Main: Suhrkamp.

Kilvington, D. (2021). The virtual stages of hate: Using Goffman's work to conceptualise the motivations for online hate. *Media, Culture & Society*, *43*(2), 256–272. doi:10.1177/0163443720972318

Kim, D., Graham, T., Wan, Z., & Rizoiu, M. -A. (2019). Analysing user identity via time-sensitive semantic edit distance (t-SED): A case study of Russian trolls on Twitter. *Journal of Computational Social Science*, *2*(2), 331–351. doi:10.1007/s42001-019-00051-x

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. doi:10.1093/joc/jqab034

Klein, A. (2012). Slipping racism into the mainstream: A theory of information laundering. *Communication Theory*, *22*(4), 427–448. doi:10.1111/j.1468-2885.2012.01415.x

Klein, A. (2017). *Fanaticism, racism, and rage online: Corrupting the digital sphere*. London: Palgrave Macmillan.

Klinger, U., & Svensson, J. (2015). The emergence of network media logic in political communication: A theoretical approach. *New Media & Society*, *17*(8), 1241–1257. doi:10.1177/1461444814522952

Kopytowska, M., & Baider, F. (2017). From stereotypes and prejudice to verbal and physical violence: Hate speech in context. *Lodz Papers in Pragmatics*, *13*(2), 133–152. doi:10.1515/lpp-2017-0008

Koschorke, A. (2021). Anpassung nach unten?: Versuch über Vulgarität. *Leviathan: Berliner Zeitschrift für Sozialwissenschaft*, *49*(2), 231–243. doi:10.5771/0340-0425-2021-2-231

Kuehn, K. M., & Salter, L. A. (2020). Assessing digital threats to democracy, and workable solutions: A review of the recent literature. *International Journal of Communication*, *14*, 2589–2610.

Laux, H., & Schmitt, M. (2017). Der Fall Bautzen: Eine Netzwerkanalyse zur Entstehung digitaler Öffentlichkeiten. *Berliner Journal für Soziologie*, *27*(3–4), 485–520. doi:10.1007/s11609-018-0354-x

Lingam, R. A., & Aripin, N. (2017). Comments on fire! Classifying flaming comments on YouTube videos in Malaysia. *Jurnal Komunikasi, Malaysian Journal of Communication*, *33*(4), 104–118. doi:10.17576/JKMJC-2017-3304-07

Loosen, W., & Schmidt, J. -H. (2012). (Re-)discovering the audience: The relationship between journalism and audience in networked digital media. *Information, Communication & Society*, *15*(6), 867–887. doi:10.1080/1369118X.2012.665467

Luhmann, N. (1970). Funktion und Kausalität. In N. Luhmann Soziologische Aufklärung. Aufsätze zur Theorie sozialer Systeme*, Bd. 1* (pp. 9–30) Opladen: Westdeutscher Verlag.

Luhmann, N. (1978). Soziologie der Moral. In N. Luhmann & S. H. Pfürtner (Eds.), *Theorietechnik und Moral* (pp. 8–116). Frankfurt a. Main: Suhrkamp.

Luhmann, N. (1981). Wie ist soziale Ordnung möglich? In N. Luhmann, *Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft* Vol. 2, pp. 195–286. Frankfurt a. Main: Suhrkamp.

Luhmann, N. (1986). *Love as passion. The codification of intimacy*. Cambridge: Harvard University Press.

Luhmann, N. (1990). *Die Wissenschaft der Gesellschaft*. Frankfurt a. Main: Suhrkamp.

Luhmann, N. (1995). *Social systems*. Redwood: Stanford University Press.

Luhmann, N. (2008). Soziologie der Moral. In N. Luhmann *Die Moral der Gesellschaft* (pp. 97–122) Frankfurt a. Main: Suhrkamp.

Luhmann, N. (2012). *Theory of society*. Redwood: Stanford University Press.

Lumsden, K., & Morgan, H. (2017). Media framing of trolling and online abuse: Silencing strategies, symbolic violence, and victim blaming. *Feminist Media Studies*, *17*(6), 926–940. doi:10.1080/14680777.2017.1316755

MacAvaney, S., Yao, H. -R., Yang, E., Russell, K., Goharian, N., Frieder, O., & Huang, M. (2019). Hate speech detection: Challenges and solutions. *PloS One*, *14*(8), 1–16. doi:10.1371/journal.pone.0221152

Marcks, H., & Pawelz, J. (2022). From myths of victimhood to fantasies of violence: How far-right narratives of imperilment work. *Terrorism and Political Violence*, *34*(7), 1415–1432. doi:10.1080/09546553.2020.1788544

Marwick, A. E., & Boyd, D. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. doi:10.1177/1461444810365313

Marwick, A. E., & Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies*, *18*(4), 543–559. doi:10.1080/14680777.2018.1450568

Massanari, A. (2017). #gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. doi:10.1177/1461444815608807

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communnication & Society*, *20*(6), 930–946. doi:10.1080/1369118X.2017.1293130

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. doi:10.1177/1527476420982230

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *WebSci '19: Proceedings of the 10thACM Conference on Web Science*, 173–182. 10.1145/3292522.3326034

Matsuda, M. J. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review*, *87*(8), 2320–2381. doi:10.2307/1289306

McCosker, A. (2014). Trolling as provocation: YouTube's agonistic publics. *Convergence*, *20*(2), 201–217. doi:10.1177/1354856513501413

McCosker, A. (2014). Trolling as provocation: YouTube's agonistic publics. *Convergence: The International Journal of Research into New Media Technologies*, *20*(2), 201–217. doi:10.1177/1354856513501413

McNamee, L. G., Peterson, B. L., & Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, *77*(2), 257–280. doi:10.1080/03637751003758227

Megarry, J. (2017). Under the watchful eyes of men: Theorising the implications of male surveillance practices for feminist activism on social media. *Feminist Media Studies*, *18*(6), 1070–1085. doi:10.1080/14680777.2017.1387584

Meibauer, J. (2014). Hassrede: Von der Sprache zur Politik. In J. Meibauer (Ed.), *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (pp. 1–16). Giessen: Gießener Elektronische Bibliothek.

Miro-Llinares, F., Moneva, A., & Esteve, M. (2018). Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. *Crime Science*, *7*(1), 1–12. doi:10.1186/s40163-018-0089-1

Miro-Llinares, F., & Rodriguez-Sala, J. J. (2016). Cyber hate speech on Twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design & Nature and Ecodynamics*, *11*(3), 406–415. doi:10.2495/DNE-V11-N3-406-415

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review*, *38*(2), 128–146. doi:10.1177/0894439318791786

Montez, D. J., & Brubaker, P. J. (2019). Making debating great again: U.S. Presidential candidates' use of aggressive communication for winning presidential debates. *Argumentation and Advocacy*, *55*(4), 282–302. doi:10.1080/10511431.2019.1672033

Morgan, A. (2022). When doublespeak goes viral: A speech act analysis of internet trolling. *Erkenntnis*. doi:10.1007/s10670-021-00508-4

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. doi:10.1111/jcom.12312

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

Mutz, D. (2015). *In-your-face politics: The consequences of uncivil media*. New Jersey: Princeton University Press.

Nagle, A. (2017). *Kill All Normies – Online culture wars from 4chan and Tumblr to Trump and the alt–right. Zero Books*. Winchester, Hampshire.

Nassehi, A. (2006). *Der soziologische Diskurs der Moderne*. Frankfurt a. Main: Suhrkamp.

Nassehi, A. (2008). Rethinking Functionalism. Zur Empiriefähigkeit systemtheoretischer Soziologie. In H. Kalthoff, S. Hirschauer, & G. Lindemann (Eds.), *Theoretische Empirie. Zur Relevanz qualitativer Forschung* (pp. 79–108). Frankfurt a. Main: Suhrkamp.

Nassehi, A. (2020). Das große Nein. In *Eigendynamik und Tragik des gesellschaftlichen Protestes*. Hamburg: Murmann.

Nassehi, A., & Saake, I. (2002). Kontingenz: Methodisch verhindert oder beobachtet? Ein Beitrag zur Methodologie der qualitativen Sozialforschung. *Zeitschrift für Soziologie*, *31*(1), 66–86.

Nisa, A. D., & Setiyawati, D. (2019). A systematic review of digital literacy training for high school students. *Advances in Social Science, Education and Humanities Research*, *353*, 376–381. doi:10.2991/icosihess-19.2019.65

Oksanen, A., Räsänen, P., & Hawdon, J. (2014). Hate groups: From offline to online social identifications. In J. Hawdon, J. Ryan, & M. Lucht (Eds.), *The causes and consequences of group violence: From bullies to terrorists* (pp. 22–48). London: Lexington Books.

Olson, G. (2020). Love and hate online. Affective politics in the era of Trump. In S. Polak & D. Trottier (Eds.), *Violence and trolling on social media* (pp. 153–177). Amsterdam: Amsterdam University Press.

Ortiz, S. M. (2019). "You Can Say I Got Desensitized to It": How Men of Color Cope with Everyday Racism in Online Gaming. *Sociological Perspectives*, *62*(4), 572–588. doi:10.1177/0731121419837588

Ortiz, S. M. (2020). Trolling as a collective form of harassment: An inductive study of how online users understand trolling. *Social Media + Society*, *April–June*1–9. doi:10.1177/2F2056305120928512

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020, April–June). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society*, *6*(2), 1–20. doi:https://doi.org/10.1177/2056305120916850

Paasch-Colberg, S., & Strippel, C. (2021). 'T he boundaries are blurry . . . ': How comment moderators in Germany see and respond to hate comments. *Journalism Studies*, *23*(2), 224–244. doi:10.1080/1461670X.2021.2017793

Paasch-Colberg, S., Strippel, C., Laugwitz, L., Emmer, M., & Trebbe, J. (2020). Moderationsfaktoren: Ein Ansatz zur Analyse von Selektionsentscheidungen im Community Management. In V. Gehrau, A. Waldherr, & A. Scholl Eds. *Integration durch Kommunikation (in einer digitalen Gesellschaft): Jahrbuch der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft 2019* Münster: Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft e.V (pp. 109–119). Deutsche Gesellschaft für Publizistik und Kommunikationswissenschaft e.V.

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180. doi:10.17645/mac.v9i1.3399

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

Papacharissi, Z. A. (2010). *A private sphere: Democracy in a digital age*. Polity Press.

Papacharissi, Z. A. (2015). *Affective publics*. Cambridge; Malden: Oxford University Press.

Park, S. (2017). Inventing aliens: Immigration control, 'xenophobia' and racism in Japan. *Race & Class*, *58*(3), 64–80. doi:10.1177/0306396816657719

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020, October–December). Hate speech: A systematized review. *SAGE Open*, *10*(4), 1–12. doi:10.1177/2158244020973022

Peterson, J. K., & Densley, J. (2017). Cyber violence: What do we know and where do we go from here? *Aggression and Violent Behavior*, *34*(1), 193–200. doi:10.1016/j.avb.2017.01.012

Phillips, W. (2019, July–September). It wasn't just the trolls: Early internet culture, 'fun,' and the fires of exclusionary laughter. *Social Media + Society*, 1–4. doi:10.1177/2F2056305119849493

Pohjonen, M., & Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 1173–1191.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55*(2), 477–523. doi:https://doi.org/10.1007/s10579-020-09502-8

Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online Civic Intervention: A New Form of Political Participation Under Conditions of a Disruptive Online Discourse. *International Journal of Communication*, *14*, 514–534.

Pöttker, H. (2016). Kommunikationsfreiheit im digitalen Zeitalter. *Communicatio Socialis*, *49*(4), 347–353. doi:10.5771/0010-3497-2016-4-347

Pöyhtäri, R. (2014). Limits of hate speech and freedom of speech on moderated news websites in Finland, Sweden, the Netherlands and the UK. *Annales*, *24*(3), 513–524.

Prinzing, M. (2017). Kompass, Kante, Kompetenz. Warum es nicht genügt, über die Verrohung des Umgangs im Netz zu klagen. *Comunicatio Socialis*, *50*(3), 334–344. doi:10.5771/0010-3497-2017-3-334

Pritsch, S. (2011). Verletzbarkeit im Netz – zur sexistischen Rhetorik des Trollens. *Feministische Studien*, *29*(2), 232–247. doi:10.1515/fs-2011-0207

Quandt, T. (2018). Dark participation: Manipulative user engagement in the news making process. *Media and Communication*, *6*(4), 36–48. doi:10.17645/mac.v6i4.1519

Quandt, T. (2021). Can we hide in shadows when the times are dark? *Media and Communication*, *9*(1), 84–87. doi:10.17645/mac.v9i1.4020

Ray, R., Brown, M., Fraistat, N., & Summers, E. (2017). Ferguson and the death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the evolution of collective identities. *Ethnic and Racial Studies*, *40*(11), 1797–1813. doi:10.1080/01419870.2017.1335422

Rega, R., & Marchetti, R. (2021). The strategic use of incivility in contemporary politics. The case of the 2018 Italian general election on Facebook. *The Communication Review*, *24*(2), 107–132. doi:10.1080/10714421.2021.1938464

Reich, Z., Domingo, D., Paulussen, S., Quandt, T., & Reich, Z. (2011). User comments: The transformation of participatory space. In J. B. Singer, A. Heinonen, A. Hermida, & M. Vujnovic (Eds.), *Participatory journalism: Guarding open gates at online newspapers* (pp. 96–117). Malden; Oxford: Wiley-Blackwell.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier*. Cambridge; London: MIT Press.

Ringrose, J., Harvey, L., Gill, R., & Livingstone, S. (2013). Teen girls, sexual double standards and 'sexting': Gendered value in digital image exchange. *Feminist Theory*, *14*(3), 305–323. doi:10.1177/1464700113499853

Rossini, P. (2019). Disentangling uncivil and intolerant discourse. In R. Boatright, D. Young, S. Sobieraj, & T. Shaffer (Eds.), *A crisis of civility? Contemporary research on civility, incivility, and political discourse* (pp. 142–157). New York: Routledge.

Saputra, M., & Al Siddiq, I. H. (2020). Social media and digital citizenship: The urgency of digital literacy in the middle of a disrupted society era. *International Journal of Emerging Technologies in Learning*, *15*(7), 156–161. doi:10.3991/ijet.v15i07.13239

Sarikakis, K., Kassa, B. E., Fenz, N., Goldschmitt, S., Kasser, J., & Nowotarski, L. (2021). "My haters and I ": Personal and political responses to hate speech against female journalists in Austria. *Feminist Media Studies*, *23*(1), 67–82. doi:10.1080/14680777.2021.1979068

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Association for Computational Linguistics, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). 10.18653/v1/W17-1101

Schulzke, M. (2016). The social benefits of protecting hate speech and exposing sources of prejudice. *Res Publica*, *22*(2), 225–242. doi:10.1007/s11158-015-9282-1

Schwarz-Friesel, M. (2019). *Judenhass im Internet. Antisemitismus als kulturelle Konstante und kollektives Gefühl*. Leipzig: Hentrich & Hentrich.

Schwertberger, U., & Rieger, D. (2021). Hass und seine vielen Gesichter: Eine sozial- und kommunikationswissenschaftliche Einordnung von Hate Speech. In S. Wachs, B. Koch-Priewe, & A. Zick (Eds.), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen* (pp. 53–77). Wiesbaden: Springer VS.

Scrivens, R., Burruss, G. W., Holt, T. J., Chermak, S. M., Freilich, J. D., & Frank, R. (2021). Triggered by defeat or victory? Assessing the impact of presidential election results on extreme right-wing mobilization online. *Deviant Behavior*, *42*(5), 630–645. doi:10.1080/01639625.2020.1807298

Sest, N., & March, E. (2017). Constructing the cyber-troll: Psychopathy, sadism, and empathy. *Personality and Individual Differences*, *119*, 69–72. doi:10.1016/j.paid.2017.06.038

Shandwick, W., Tate, P., & KRC Research (2018). *Civility in America 2018: Civility at work and in our public squares*. https://www.webershandwick.com/news/civility-in-america-2018-civility-at-work-and-in-our-public-squares/

Shandwick, W., Tate, P., & KRC Research (2019). *Civility in America 2019: Solutions for tomorrow*. https://www.webershandwick.com//wp-content/uploads/2019/06/CivilityInAmerica2019SolutionsforTomorrow.pdf

Shaw, F. (2016, October–December). "Bitch I Said Hi": The Bye Felipe Campaign and Discursive Activism in Mobile Dating Apps. *Social Media + Society*, *2*(4), 1–10. doi:10.1177/2056305116672889

Shoemaker, P., & Reese, S. (2014). *Mediating the message in the 21st Century. A media sociology perspective*. New York: Routledge.

Simpson, R. M. (2013). Dignity, harm, and hate speech. *Law and Philosophy*, *32*(6), 701–728. doi:10.1007/s10982-012-9164-z

Simpson, R. M. (2019). 'Won't somebody please think of the children?' Hate speech, harm, and childhood. *Law and Philosophy*, *38*(1), 79–108. doi:10.1007/s10982-018-9339-3

Sobieraj, S. (2018). Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society*, *21*(11), 1700–1714. doi:10.1080/1369118X.2017.1348535

Sorokowski, P., Kowal, M., Zdybek, P., & Oleszkiewicz, A. (2020). Are online haters psychopaths? Psychological predictors of online rating behavior. *Frontiers in Psychology*, *11*, 1–5. doi:10.3389/fpsyg.2020.00553

Sponholz, L. (2018). *Hate Speech in den Massenmedien. Theoretische Grundlagen und empirische Umsetzung*. Wiesbaden: Springer VS.

Sponholz, L. (2020). Der Begriff 'Hate Speech' in der deutschsprachigen Forschung. Eine empirische Begriffsanalyse. *SWS-Rundschau*, *60*(1), 43–65.

Sponholz, L. (2021). Hass mit Likes: Hate Speech als Kommunikationsform in den Social Media. In S. Wachs, B. Koch-Priewe, & A. Zick (Eds.), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen* (pp. 15–37). Wiesbaden: Springer VS.

Stahel, L., & Weingartner, S. (2019). Online aggression from a sociological perspective: An integrative view on determinants and possible countermeasures. In Association for Computational Linguistics, *Proceedings of the Third Workshop on Abusive Language Online* (pp. 181–187). 10.18653/v1/W19-3520

Stegbauer, C. (2018). *Shitstorms. Der Zusammenprall digitaler Kulturen*. Wiesbaden: Springer.

Sundén, J., & Paasonen, S. (2018). Shameless hags and tolerance whores: Feminist resistance and the affective circuits of online hate. *Feminist Media Studies*, *18*(4), 643–656. doi:10.1080/14680777.2018.1447427

Su, L.Y. -F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, *20*(10), 3678–3699. doi:10.1177/1461444818757205

Sydnor, E. (2018). Platforms for incivility: Examining perceptions across different media formats. *Political Communication*, *35*(1), 97–116. doi:10.1080/10584609.2017.1355857

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, *119*(42), 1–11. doi:10.1073/pnas.2207159119

Tworek, H., & Leerssen, P. (2019). An analysis of Germany's NetzDG law. *The Transatlantic Working Group*, 1–11. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf

Udupa, S., & Pohjonen, M. (2019). Extreme Speech — Extreme Speech and Global Digital Cultures — Introduction. *International Journal of Communication*, *13*(19). https://ijoc.org/index.php/ijoc/article/view/9102

Udupa, S., & Pohjonen, M. (2019). Extreme speech and global digital cultures — Introduction. *International Journal of Communication*, *13*(2019), 3049–3067.

United Nations (2020). *United Nations Strategy and Plan of Action on Hate Speech*. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

Uth, B., & Meier, K. (2018). Wie Redaktionen bessere Diskussionen fördern können. Einflussfaktoren auf die Qualität von Nutzerkommentaren. *Communicatio Socialis*, *51*(3), 331–345. doi:10.5771/0010-3497-2018-3-331

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. *ArXiv*. doi:10.18653/v1/W18-5105

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media* (online ed.). Oxford Academic. doi:10.1093/acprof:oso/9780199970773.001.0001

Van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, *1*(1), 2–14. doi:10.17645/mac.v1i1.70

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In Association for Computational Linguistics, *Proceedings of the Third Workshop on Abusive Language Online* (pp. 80–93). 10.18653/v1/W19-3509

Wachs, S., Wettstein, A., Bilz, L., & Gámez-Guadix, M. (2022). Adolescents' motivations to perpetrate hate speech and links with social norms. *Communicar*, *30*(71), 1–11. doi:10.3916/C71-2022-01

Wachs, S., & Wright, M. F. (2019). The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. *Cyberpsychology, Behavior and Social Networking*, *22*(5), 300–306. doi:10.1089/cyber.2018.0551

Wagner, E. (2019). *Intimisierte Öffentlichkeiten. Pöbeleien, Shitstorms und Emotionen auf Facebook*. Bielefeld: Transcript.

Wagner-Egelhaaf, M. (2020). Figuren des Hasses. Prolegomena zu einer Literatur- und Kulturgeschichte. *Jahrbuch für Internationale Germanistik*, *52*(1), 81–90. doi:10.3726/JA521_81

Walker, S. (1994). *Hate speech: The history of an American controversy*. Lincoln: University of Nebraska Press.

Warner, M. (2002). *Publics and counterpublics*. New York: Zone Books.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In Association for Computational Linguistics, *Proceedings of the First Workshop on Abusive Language Online* (pp.78–84). 10.18653/v1/W17-3012

Wekesa, M. (2019). Hate online: The creation of the 'other. *Lodz Papers in Pragmatics*, *15*(2), 183–208. doi:10.1515/lpp-2019-0011

Weller, K., Bruns, A., Burgess, J., Puschmann, C., & Mahrt, M. (Eds.). (2014). *Twitter & society*. Bern: Lang.

Westlund, O. (2021). Advancing research into dark participation. *Media and Communication*, *9*(1), 209–214. doi:10.17645/mac.v9i1.1770

Wiederer, R. (2003). Jugend, Gewalt und Internet: Risiken der Internetnutzung von Kindern und Jugendlichen. *Sozialwissenschaften und Berufspraxis*, *26*(2), 181–197.

Williams, M. L., & Burnap, P. (2015). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *The British Journal of Criminology*, *56*(2), 211–238. doi:10.1093/bjc/azv059

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117. doi:10.1093/bjc/azz064

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149–1168. doi:10.1177/0038038517708140

Windisch, S., Wiedlitzka, S., & Olaghere, A. (2021). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, *17*(1), 1–17. doi:10.1002/cl2.1133

Wintterlin, F., Schatto-Eckrodt, T., Frischlich, L., Boberg, S., & Quandt, T. (2020). How to cope with dark participation: Moderation practices in German newsrooms. *Digital Journalism*, *8* (7), 904–924. doi:10.1080/21670811.2020.1797519

Xiang, G., Fan, B., Wang, L., Hong, J. I., & Rose, C. P. (2012). Detecting offensive tweets via topical feature discovery over a large-scale Twitter corpus. *CIKM '12: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 1980–1984*. 10.1145/2396761.2398556

Yang, K. -C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. doi:10.1002/hbe2.115

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598. doi:https://doi.org/10.7717/peerj-cs.598

Zelenkauskaite, A., & Balduccini, M. (2017). "Information Warfare" and Online News Commenting: Analyzing Forces of Social Influence Through Location-Based Commenting User Typology. *Social Media + Society*, *3*(3), 1–13. doi:10.1177/2056305117718468

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*, *10*(5), 925–945. doi:10.3233/SW-180338

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, *12*(1), 415–438. doi:https://doi.org/10.1146/annurev-economics-081919-050239

Ziegele, M. (2016). *Nutzerkommentare als Anschlusskommunikation. Theorie und qualitative Analyse des Diskussionswerts von Online-Nachrichten*. Wiesbaden: Springer VS.

Ziegele, M., & Jost, P. B. (2020). Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research*, *47*(6), 891–920. doi:10.1177/0093650216671854

Ziegele, M., Jost, P. B., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *Studies in Communication and Media*, *7*(4), 525–554. doi:10.5771/2192-4007-2018-4-525

Ziegele, M., Naab, T. K., & Jost, P. (2019). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*(5), 731–751. doi:10.1177/1461444819870130