MediaWell

Research Review

Hate Speech, Information Disorder, and Conflict

April 7, 2020

The following is a condensed and edited version of a research review that the SSRC commissioned for its Academic Network on Peace, Security, and the United Nations. For the full version, please click here.

Speech as a Weapon

Social scientists have been studying hate speech and information disorder for a long time, investigating their role in enabling and triggering conflict and violence. More recently, researchers have been trying to explain how hate speech and information disorder interact with vitriolic and dehumanizing language and imagery powered by the spread of online communication.

Hate speech and information disorder are weapons of war and enablers of conflict, used to create and reinforce sentiments of mistrust, exclusion, fear, and anger toward perceived enemies, and simultaneously to unite allies. Their use and impact—under the labels of propaganda, information warfare, and psychological warfare—have been widely documented and researched (Taylor 2003).

While this has been true for a long time, new features of these phenomena and behaviors have encouraged new conceptual and methodological approaches. Hate speech online is particularly disturbing for its commonness. This kind of speech is appropriated and shared by ordinary citizens, is used to support open confrontations between nations or blocs, and can be approved or encouraged by state governments. While the involvement of governments and powerful organized groups (such as terrorist organizations) is striking in concerted disinformation campaigns (Richey 2018) and propaganda (Howard and Kollanyi 2016), these tactics also turn ordinary users into active participants in the spread of hate and disinformation. (For more on concerted disinformation campaigns, see our research review Producers of Disinformation.)

Hate speech: Scope and approaches

Academics have tended to treat hate speech as distinct from other forms of communication, as a specific type of emotional expression that has the ability to reduce empathy and trigger conflicts under specific conditions. Despite its uniqueness and its potential harms, definitions of hate speech vary widely. Ways of thinking about the causal links between hate speech and conflict also vary. Narrower conceptions like “dangerous speech” and “fear speech” attempt to focus on the ability of speech to cause harm and lead to violent outcomes (Benesch 2012; Buyse 2014).

Legal and regulatory studies have been concerned with defining hate speech in precise enough terms to enable legal and regulatory action, drawing a balance between freedom of speech and rights to dignity and safety. The vast majority of these studies have focused on the global North and on the divide that exists between American and European approaches to regulating hate speech (Rosenfeld 2003; Bleich 2014). To a much lesser extent, studies have scrutinized legal traditions in other countries. For example, researchers have explored the influence of customary laws or the role religion plays in enabling and restricting freedom of expression (D’Souza et al. 2018; Edge 2018). Examples of legal pluralism and diverse approaches to defining and regulating hate speech do exist. In Somalia, where poetry constitutes a popular vehicle for the dissemination of information and ideas, community elders prohibit poets from composing new work if they have a history of producing derogatory poems that slander individuals or groups (Stremlau 2012).

A relatively distinct approach toward defining hate speech has sought to focus not on its intrinsic content but on the functions it serves. Hate speech involves manipulation of social differences with two interlinked effects (Waltman and Mattheis 2017). One of those produces an out-group effect by targeting populations using dehumanizing terms. Target communities are positioned as threats to the communities that hate speakers claim to represent. On the other hand, hate speech also has an in-group function in terms of recruiting and socializing new members and strengthening in-group memory. By exchanging and repeating hateful expressions targeting an out-group, group solidarities are built through rhetorical means and memory politics (Perry 2001).

As Waldron (2012) writes, the warnings in hateful expressions aimed at out-groups may sound something like this:

> Don’t be fooled into thinking you are welcome here. […] You are not wanted, and you and your families will be shunned, excluded, beaten, and driven out, whenever we can get away with it. We may have to keep a low profile right now. But don’t

> get too comfortable. […] Be afraid.

The same expression can let allies of the speaker know they are not alone and reinforce a perceived threat to the in-group. In this case, the covert message may read:

> We know some of you agree that these people are not wanted here. We know that some of you feel that they are dirty (or dangerous or criminal or terrorist). Know now that you are not alone. […] There are enough of us around to make sure these people are not welcome. There are enough of us around to draw attention to what these people are really like.

Beyond legal scholarship and security studies, other academic disciplines have adopted a more eclectic approach. They have been less concerned with finding widely shared definitions and more with understanding hate speech as a phenomenon affecting specific groups, and one which points to wider societal challenges.

Communication studies, sociology, anthropology, and cultural studies consider hateful speech as a form of "constitutive rhetoric" in which a text calls its audience into being (Charland 1987). This means that written, auditory, or visual materials can construct audiences by creating relationships among strangers by *addressing* them and demanding their attention, and by simultaneously creating a discursive field for exchanging certain ideas (Warner 2002). Relatedly, text is approached as a "speech act" (Butler 1997) that can have perlocutionary effects (acts done by saying something) and illocutionary force (acts done in saying something) (Austin 1975). Illocutionary speech acts have the force to perform what they describe. For example, accusing someone of blasphemy can lead to *constituting* that person as a blasphemer (Schaflechner, in review). Perlocutionary effects are the consequences of such speech acts on the addressee (here, the person accused of blasphemy). Perlocutionary effects of words such as "run" can be the actual action of running. Sometimes perlocutionary effects are not indicated in the words themselves. For example, one may stop an action after someone exclaims, "What the hell?"

These foundational concepts are important because they see a deeper role for hateful speech in establishing and perpetuating the conditions for symbolic and physical attacks on target populations. In the words of Keen (1986), groups that are excluded are first "rhetorically killed" before they may be physically killed. Townsend (2014) has offered a "negative language continuum" comprising hate speech (the least extreme) and incitement to genocide (the most extreme). In the middle of the spectrum, "genocidal discourse" involves "the escalation of a widely acceptable language of hatred into language that

proposes, promotes or justifies the destruction of a group as acceptable and/or necessary." Townsend's examination of the persecution of Roma communities in some Eastern European countries provides a telling example of the ways hateful speech facilitated "biological erasure through coercive and forced sterilizations" in Slovakia.

Hate speech and the internet

The expansion of internet-enabled media has made it even harder to understand the nature and effects of hate speech. Prominent studies and literature surveys have suggested that the internet "has had a revolutionizing influence on groups' use of hate speech" (Waltman and Mattheis 2017), but there is no consensus on the actual role played by the internet on processes of radicalization and hate mongering (O'Callaghan et al. 2015).

In public debates, claims that "hate speech is on the rise" have become a common refrain, but these claims are very difficult to prove for at least three reasons. The first is the sheer amount of speech that is produced on a daily basis. Some countries keep a record of hate crimes (EUFRA 2018), allowing them to map whether these are on the rise or in decline (and possibly exploring correlations with potential triggers). However, when it comes to speech, there are very few reliable statistics mapping whether this is indeed more pervasive than in the past, beyond case studies and catalyzing events (e.g., elections). The second, related challenge to understanding whether hate speech has been on the rise is that the publicity and persistence of text and images enabled by social media may have simply made common slurs and vitriol previously contained in private spheres more visible and accessible (Rowbottom 2012). Related to this is the complexity of defining clear boundaries across phenomena that have become constitutive of internet culture, such as trolling, doxing, swarming, and "lulz" (internet pleasure cultures). Finally, the few institutions that may be able to provide large-scale and reliable statistics—the most popular social networking platforms—have been very careful not to make this information public, as it may severely affect their image.

For these reasons, it is also difficult to assess the impact of online hate speech on conflict situations, except when the broader ambient and symbolic effects of such speech are considered or specific cases are examined.

Disinformation: New frameworks for the digital era

As a nascent field of interdisciplinary inquiry, disinformation studies has yet to find a coherent framework for theory, definitions, and methods, though Wardle and Derakhshan's

(2017) “information disorder” typology has gained traction. (For more on this definitional problem, see our live research review Defining “Disinformation.” Both “hate speech” and “information disorder” have been invoked in an interrelated way to examine the internet’s role in shaping conflicts that are specific to contexts and regions.

Focusing on contemporary alt-right movements in the US, Marwick and Lewis (2017) show how these groups have taken advantage of the digital media ecosystem to spread disinformation, influence public opinion, and shift political consensus. According to them, it is impossible to quantify how online disinformation influenced the outcome of the 2016 US presidential election, but the impact is observable in the discourse and narratives taken up by mainstream news outlets and politicians. Daniels (2018) has shown linkages between alt-right disinformation and online activity and events such as the Charlottesville rally and Charleston church shooting, in terms of online activity that accompanied these events.

Examining the impact of digital disinformation on intercommunity conflicts in Bangladesh, Al-Zaman (2019) has illustrated that digital media are impeding the peaceful coexistence of religious communities, playing a role in inciting aggressive behavior by dominant religious groups against religious minorities, and successfully staging communal violence along religious fault lines. In the first case he examines, coordinated mob violence by the Muslim majority population was spurred by a Facebook post allegedly created by a Hindu fisherman “defaming” Islam. Following the violence, it was found that the post was a fake and had been purposefully created to fan the flames of intercommunal religious tensions. In the second case, a fake Facebook account linked to a young Buddhist man was used to spread a post portraying the desecration of the Quran. Similarly, in India, studies have shown how digital rumors have spurred mob lynching of minority Muslims by majority Hindu nationalist groups (Mirchandani 2018).

Security and defense studies frame the emerging trends of information disorder as “information warfare,” arguing that imagination has become the primary target of manipulation in the information era (Araźna 2015; see also Lewandowsky et al. 2013; Richey 2018; Stengel 2019). The impact of manipulative actions is based on stimulating emotions such as enthusiasm or fear. In the context of modern hybrid warfare, disinformation and manipulation blur the term “war” and make it imprecise in the field of international law.

As with hate speech, the specific configuration of power and the actors involved in a disinformation campaign vary across cases. In some cases, disinformation can be seen as carefully directed from a—more or less disguised—central authority. In others, the role of bottom-up practices of citizens contributes to produce a form of disorder that benefits specific actors.

Disinformation is seen as a problem not only of ordinary media users and governments but also (primarily) of social media companies and digital influencers (Tactical Tech 2019). Social networking platforms play a role in extremist cyberspaces (O'Callaghan et al. 2015) and in creating "truth markets" (Harsin 2015). Platform recommendation algorithms progressively isolate users in ideological content bubbles. On YouTube in particular, users are very likely to become immersed in an algorithmically sustained extreme right ideological bubble after only a few clicks (O'Callaghan et al. 2015; Lewis 2018). (For more on the "bubble" issue, see our research review on Contexts of Misinformation.)

Evolving debates around disinformation are conceptually rich, but empirical evidence that links disinformation with conflict situations is lacking. A majority of studies across disciplines as varied as psychology, peace and security studies, political science, media and conflict studies, political communication studies, and anthropology have used the case-study method to gather empirical evidence. They have closely analyzed the spread of disinformation within a selected set of conflict situations such as riots, hate crimes and elections (Forelle et al. 2015; Howard and Kollanyi 2016; Kajimoto and Stanley 2019; Lewandowsky et al. 2013; Persily 2017; Richey 2018).

Actors, actions, and target groups

A rich body of research has highlighted new dynamics emerging around online hate and disinformation. Looking through the narrow lens of a causal link between online speech and physical conflict misses this nuance.

First, there are new kinds of actors that the internet has energized and facilitated, with direct consequences for how hate and aggression have spread online as a shared transnational practice. The roles of "ordinary users" as disseminators of disinformation as well as "disinformation innovators" who employ online freelance labor illustrate the new trend. These changes make it easier for foreign agents to tap into digital toxicity that transcends national boundaries, and these strategies directly benefit from digital communication that is built for instantaneous expression and reaction (Brown 2017).

Such conditions give rise to new paradigms of communication like "the shitstorm" which renders the public as a "swarm" that is trained on the hyper-present, unconcerned with the formulation of collective futures, and driven by affect (Han 2017).

Second, the processes that accompany hate speech have shifted. Online aggression and hateful speech are rendered pleasurable and enjoyable (Daniels 2018). People who call out racism are dismissed as "normies" (Nagle 2017) or "liberals who don't get the joke" (Hervik 2019). Wendling (2018) links this to internet cultures of lulz common in anonymous

imageboards such as 4Chan (see also Topinka 2017). Similarly, “muhei stickers” in China that circulate on online messaging apps target Muslim communities by reinforcing slanderous stereotypes through visual ethnic humor (de Seta, forthcoming). Udupa (2019) has defined this phenomenon as “fun as a meta-practice of exclusionary extreme speech.” Fun in this sense is not frivolity of action, but a serious political activity that consolidates communities of supporters for exclusionary ideologies. In digital environments, fun instigates collective pleasures of identity that can mitigate risk and culpability for hateful speech. Banalization of online hate has become a new enabling ground for exclusionary politics to stabilize, complementing conventional strategies of “serious” appeal and dissemination. Siapera, Moreo, and Zhou (2018) show that racist hate speech on Twitter and Facebook within the Irish context varies between “crude racism” (insults, slurs, profanity, animal comparisons, appeals to racial stereotypes, etc.) and “coded racism” (superficial appearance of rationality that appeals to cultures, values, ethnicity, and common-sense arguments).

Online hate speech is also itinerant and migratory. Even when the content is removed it can recur on the same platform under a different name or in different online spaces. For instance, responding to greater restrictions by social networking platforms, violent Jihadi groups moved to encrypted channels such as Telegram or file-sharing sites such as Pastebin, while the extreme right migrated to platforms such as VKontakte or Gab. Ganesh (2018) has argued that three formal features of digital hate cultures make them ungovernable. First, their swarm structures are characterized by decentralized networks. Second, they exploit inconsistencies in web governance between different social media companies, as well as between private and government actors. Third, they use coded language to evade content moderation.

If we look at the targets of online hate speech and disinformation, we see both disturbing continuities and surprising new victims. Racist banter continues to target people of color. Stereotypes against Jews portray them as stingy, conniving, and greedy. Vehement hate is directed at the newly invigorated category of “immigrants” denigrated as “refugees” and “asylum seekers,” and applied just as much to second-generation and mixed-background citizens (Siapera, Moreo, and Zhou 2018). Muslims in particular continue to be treated as canvasses for projecting fears of cultural conquest and displacement (Mårtensson 2014; Tanner and Campana 2019; Tell MAMA 2014; Stewart 2019). Online Islamophobia makes civilizational arguments that Muslim values are fundamentally in opposition to European and North American values (Bangstad 2014; Hervik 2019; Mårtensson 2014; Sponholz 2016). Muslim minorities are also frequent targets in India and Sri Lanka (George 2016). Online misogyny broadly attacks women and has become an integral part of contemporary alt-right ideology (Lyons 2017).

White supremacy that cuts through these targeted speech forms has threatened to roll back values of racial equality established in the post–civil rights era (Daniels 2018; Back 2002). Such exclusions are given a veneer of serious theoretical deliberation by invoking ideas of ethnopluralism that argue against mobility of people by framing it as people's "right" to live in their places of origin, and that forcing them out of their native lands is an act of violence. Jihadist extremism online propagates a religious war against all non-Muslims seen as *haram* (Conway et al. 2019).

Studying hate speech, disinformation, and conflict

Online communication has been offering an unprecedented amount of data for researchers to study mediated social interactions. Strides in computational and quantitative techniques are promising as well as necessary, considering the vast volumes of data generated each day and their systematic use by vested interest groups. Despite their rapid evolution and encouraging results, there are important limitations to these approaches. Social media companies have placed restrictions on how much data can be accessed for research; archival data comes with high price tags and lack of transparency in selection. Publicly available datasets differ vastly in size, scope, and characteristics of annotated data (Freelon 2018; MacAvaney et al. 2019).

Moreover, the opportunity to access volumes of online data has been seized in distinct ways by different disciplines, deepening our collective understanding of specific mechanisms (e.g., how and why specific messages spread), but also leaving other pressing questions—especially those requiring deeper engagements with communities beyond their online manifestations—underresearched and unanswered. The primary focus of machine-learning models and computational linguistics has been on detection and labeling of data, with no sufficient contextual knowledge of actors, networks, and meanings underpinning hateful content.

Internet discourses cannot be isolated from other media channels and communication structures that exist in societies. Across all the cases of hateful speech and disinformation examined by academic studies, internet technologies have always influenced public discourse in connection with older media forms and existing animosities based on religion, migration status, gender, nationality, race, ethnicity, and caste. In Myanmar, hatred against Rohingya Muslims is perpetrated not only via Facebook but also state-controlled newspapers (Lee 2019). Timmermann (2008) has similarly shown that systematic, state-orchestrated hate speech was a direct cause of genocidal killing in Rwanda. Studying the case of hate speech against the Kurds in Turkey, Onbasi (2015) has illustrated how attempts

to curb such speech did not succeed because the state used the framework of "national security" to portray Kurds as threats to the nation, thereby undermining protection to Kurds from hateful speech.

The vast majority of these studies have had a narrow focus on what is being said or displayed, and how and why messages emerge and spread. They have offered few insights into the speakers as individuals, why they engage in these types of behaviors, and how these forms of language may contribute to violence beyond digital spaces.

There are a few exceptions. For example, Ong and Cabañes (2018) have revealed a complex business network that has emerged around "disinformation services" in the Philippines. They caution that the stockpile of digital weapons in the Philippines, with its highly organized online freelance labor force, may have far-reaching consequences for fragile democracies in the global South as well as more established democracies in the West.

Research emerging from Kenya, Uganda, and Somalia has illustrated how callers to radio stations have learned to exploit audiences' belief that new spaces of interactions are supposedly freer from power, allowing them to manipulate discussions in ways that favor partisan agendas (Brisset-Foucault 2016; Gagliardone 2016; Livingston 2011; Stremlau, Fantini, and Gagliardone 2015). In India, disinformation agents are not only well-paid techies and influencers but also underemployed youth who make opportunistic arrangements through networks of patronage politics and those drawn to precarious conditions of disinformation labor. Moreover, politically partisan groups have attempted to consolidate their agendas by presenting online discussions as user autonomy and voluntary work, concealing both online labor and top-down propaganda (Udupa 2019). These studies show how flagging these actors simply as self-serving manipulators risks missing complex realities on the ground, and potentially ignores the responsibilities of media organizations, networking platforms, and political systems.

More to the point, online speech—in its aggressive and antagonistic forms—has also been critical for political contestations. In their research on online communication in Ethiopia, Gagliardone et al. (2016) have located hate speech in the context of the broad variety of communicative practices enabled by social networking platforms. This approach has highlighted how antagonistic messages can also attack those in power in ways that can lay the foundations for other kinds of contestation. Livingston (2011) has found that across the African continent, older technologies like radio and newspapers are hubs of politically motivated disinformation. Digital communication technologies are positioned as a means to level the field, giving citizens access to information that could serve as a corrective against disinformation.

Social psychologists have arguably developed the most systematic strategies to test how hate speech can promote behaviors connected to violence and conflict, including prejudice, desensitization, and dehumanization (Rai, Valdesolo, and Graham 2017; Soral, Bilewicz, and Winiewski 2018). They have illustrated, for example, that repetitive exposure to hate speech does lead to lower evaluations of the victims and greater distancing. The resulting dehumanization may increase the likelihood of violence. Limitations, however, also exist in these cases. These studies have relied on small groups of individuals tested in controlled environments and exposed to selected inputs, which are often removed from what occurs in real-world scenarios. Using survey methods, a small number of studies have investigated the impact of disinformation in terms of differences in cultural perceptions and political views that exist between national communities. For instance, Gerber and Zavisca (2016) have shown that there was widespread acceptance of the Russian narrative regarding the conflict with Ukraine in Krygzstan, but people in Azerbaijan were more skeptical.

Responses and future directions

As scholarship on the impact of digital communication on hate speech and disinformation expands, one pressing question is how researchers should approach the vexing issue of finding solutions to ongoing developments.

Responses to violent speech have largely been in the form of content takedowns and prefiltering (Conway et al. 2019; Pohjonen 2018). Governmental agencies such as the US State Department, the UK Foreign and Commonwealth Office, and international organizations such as the United Nations, are frequent funders of projects that seek to counter violent extremism, recruitment, and radicalization (Ferguson 2016). Increasingly, tech platforms have adopted this language as they have come under increasing pressure by the US and European governments to address extremist incitements to violence (Andrews and Seetharaman 2016). AI-assisted systems are the latest effort in this direction. However, the problem of "black-boxing," where algorithmic decisions can no longer be interpreted or challenged by human appeal, is an unresolved issue (Davidson et al. 2017). Studies have also raised concerns over algorithmic bias in identifying hate speakers and hateful lingos because of the homogenous work force of technology companies with disproportionately few women and people of color (Noble 2018).

Some studies have emphasized the value of counterspeech in combating online hateful speech and disinformation (ARTICLE 19 2019; Benesch 2014; Citron and Norton 2011; Faris et al. 2016; Mårtensson 2014; Roshani 2016). Scholars suggest that counterspeech is preferable to state interference because it can avoid governmental misuse of legal provisions to clamp down opposition. However, critics have pointed out several problems

with this solution. Counterspeech comes with the risk of providing hateful speech with "relevance, discussability and better discourse quality" by turning objectionable content into a newsworthy controversy (Sponholz 2016). Examining the case of Italian intellectual Oriana Fallaci's Islamophobic pamphlet, *The Rage and the Pride*, which was published in newspapers, Sponholz argues that counterspeech did not lead to refutation of hate speech but contributed toward transforming it into a legitimate controversy deserving media attention. Other studies have argued that counterspeech and grassroots activism have gone hand in hand to generate several positive outcomes. In Brazil and Colombia, counterspeech activism has increased public awareness around racism, provided free legal advice to victims, and led to greater enforcement of laws criminalizing racism as well as promoting inspiring public personalities through online media (Roshani 2016). These efforts resonate with the longer tradition of building societywide counternarratives to combat hate.

There is a glaring need to bring historical context to hate speech and information disorder in the digital age. On the one hand, digital landscapes in the global South are underexplored, despite the fact that these regions constitute the fastest-growing digital markets in the world, with a vast plurality of political systems (Milan and Treré 2019). On the other hand, existing studies of online hate and disinformation in the global North are constrained by over-emphasis on contemporary developments in technology while overlooking longer postcolonial histories of racial construction (see Deem 2019; de Genova 2010). There is a related conceptual problem that undergirds these issues. With notable exceptions, studies on the global North implicitly assume that "emotionality" of hateful speech is an aberration that stands in contrast to calm rationality as a default value of the postwar Western world. Studies on Africa, South Asia, and Southeast Asia, on the other hand, consider conflict as a propensity exacerbated by emotionally charged verbal cultures that are further amplified by long-standing ethnic, religious, and caste divisions. This heuristic division between the North and the South, and the accompanying conceptual construction of the rational center and emotional periphery, do not account for vast disparities inflicted upon societies through the colonial encounter. In an ironic twist, the expansion of the internet media has had an equalizing effect in terms of recognizing that North America and Europe are no longer "exceptional" in terms of violent emotionality of hate speech. The broader policy agenda would then be to inquire how a global approach to hateful speech, disinformation, and conflict might recognize enduring hierarchies and emerging exclusions within and across societies.

Historical contextualization, attention to everyday online user cultures, and global comparative models are important in developing a non–digital media centric analysis of hate speech and disinformation—an approach advocated by the "extreme speech" framework (Udupa and Pohjonen 2019). This framework emphasizes understanding specific cultural contexts and connecting key debates on hateful speech and disinformation with decolonial

perspectives. Among other things, this entails systematic inquiry into longer histories of racial construction and hierarchies shaped by colonial rule that have been revived and weaponized by current regimes, including those aimed against people within one's own national communities.

To address these challenges, we urgently need interdisciplinary collaboration between computational scientists and scholars who study media practices, societies, histories, and cultures. We also need concerted pressure on social media companies to provide data access to researchers. Such interdisciplinary efforts can advance beyond the currently limited focus on detecting and labeling hate speech and disinformation, and move us toward holistic, context-sensitive solutions.

Works Cited

Al-Zaman, Sayeed. 2019. "Digital Disinformation and Communalism in Bangladesh." *China Media Research* 15 (2). https://doi.org/10.31235/osf.io/8s6jd.

Andrews, Natalie, and Deepa Seetharaman. 2016. "Facebook Steps Up Efforts AgainstTerrorism." *Wall Street Journal*. https://www.wsj.com/articles/facebook-steps-up-efforts-against-terrorism-1455237595.

Araźna, Marzena. 2015. "Conflicts of the 21st Century Based on Multidimensional Warfare – 'Hybrid Warfare,' Disinformation and Manipulation." *Security and Defence Quarterly* 8 (3): 103–29. https://doi.org/10.5604/23008741.1189421.

Article 19. 2019. "'Hate Speech' Explained: A Toolkit." ARTICLE 19. Accessed March 5, 2020. https://www.article19.org/resources/hate-speech-explained-a-toolkit/.

Austin, John Langshaw. 1975. *How to Do Things with Words*. Harvard University Press.

Back, Les. 2002. "Wagner and Power Chords: Skinheadism, White Power Music, and the Internet." In *Out of Whiteness*. University of Chicago Press. https://www.press.uchicago.edu/ucp/books/book/chicago/O/bo3641103.html.

Bangstad, Sindre. 2014. *Anders Breivik and the Rise of Islamophobia*. Zed Books Ltd.

Benesch, S. 2012. "Dangerous Speech: A Proposal to Prevent Group Violence." World Policy Institute. http://www.worldpolicy.org/sites/default/files/Dangerous%20Speech%20Guidelines%20Benesch%20January%202012.pdf.

Benesch, Susan. 2014. "Countering Dangerous Speech – New Ideas for Genocide

Prevention.” United States Holocaust Memorial Museum.

Bleich, Erik. 2014. “Freedom of Expression versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the USA and Europe.” *Journal of Ethnic and Migration Studies* 40 (2): 283–300. https://doi.org/10.1080/1369183X.2013.851476.

Brisset-Foucault, Florence. 2016. “Serial Callers: Communication Technologies and Political Personhood in Contemporary Uganda.” *Ethnos* 83 (2): 255–73. https://doi.org/10.1080/00141844.2015.1127984.

Brown, Alexander. 2017. “What Is so Special about Online (as Compared to Offline) Hate Speech.” *Ethnicities*.http://journals.sagepub.com/doi/10.1177/1468796817709846.

Butler, Judith. 1997. *Excitable Speech: A Politics of the Performative*. Psychology Press.

Buyse, Antoine. 2014. “Words of Violence: ‘Fear Speech,’ or How Violent Conflict Escalation Relates to the Freedom of Expression.” *Human Rights Quarterly* 36 (4): 779–97. https://doi.org/10.1353/hrq.2014.0064.

Charland, Maurice. 1987. “Constitutive Rhetoric: The Case of the ‘Peuple Quebecois.’” *Quarterly Journal of Speech* 73 (2): 133–50.

Citron, Danielle, and Helen Norton. 2011. “Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age.” *Boston University Law Review* 91: 1435–84.

Conway, Maura, Moign Khawaja, Suraj Lakhani, Jeremy Reffin, Andrew Robertson, and David Weir. 2019. “Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts.” *Studies in Conflict & Terrorism* 42 (1–2): 141–60. https://doi.org/10.1080/1057610X.2018.1513984.

Daniels, Jessie. 2018. “The Algorithmic Rise of the ‘Alt-Right.’” *Contexts*, April. http://journals.sagepub.com/doi/10.1177/1536504218766547.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language,” March. http://arxiv.org/abs/1703.04009v1.

De Genova, Nicholas. 2010. “Migration and Race in Europe: The Trans-Atlantic Metastases of a Post-Colonial Cancer.” *European Journal of Social Theory* 13 (3): 405–19. https://doi.org/10.1177/1368431010371767.

Deem, Alexandra. 2019. “Extreme Speech | The Digital Traces of #whitegenocide and Alt-Right Affective Economies of Transgression.” *International Journal of Communication* 13

(July). https://ijoc.org/index.php/ijoc/article/view/9631.

D'Souza, Tanya, Laura Griffin, Nicole Shackleton, and Danielle Walt. 2018. "Harming Women with Words: The Failure of Australian Law to Prohibit Gendered Hate Speech." *UNSW Law Journal* 41: 38.

Edge, Peter W. 2018. "Oppositional Religious Speech: Understanding Hate Preaching." *Ecclesiastical Law Journal* 20 (3): 278–89. https://doi.org/10.1017/S0956618X18000467.

EUFRA. 2018. *Hate Crime Recording and Data Collection Practice across the EU*. EU Agency for Fundamental Rights. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-hate-crime-recording_en.pdf.

Faris, Robert, Amar Ashar, Urs Gasser, and Daisy Joo. 2016. "Understanding Harmful Speech Online." Networked Policy Series, Berkman Klein Center Publication Series. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2882824.

Ferguson, Kate. 2016. *Countering Violent Extremism Through Media and Communication Strategies: A Review of the Evidence*. Partnership for Conflict, Crime, and Research Security. http://www.paccsresearch.org.uk/wp-content/uploads/2016/03/Countering-Violent-Extremism-Through-Media-and-Communication-Strategies-.pdf.

Forelle, Michelle, Phil Howard, Andrés Monroy-Hernández, and Saiph Savage. 2015. "Political Bots and the Manipulation of Public Opinion in Venezuela." Social Science Research Network. https://ssrn.com/abstract=2635800.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35 (4): 665–68. https://doi.org/10.1080/10584609.2018.1477506.

Gagliardone, Iginio. 2016. "'Can You Hear Me?' Mobile–Radio Interactions and Governance in Africa." *New Media & Society*18 (9): 2080–95. https://doi.org/10.1177/1461444815581148.

Gagliardone, Iginio, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright, et al. 2016. "Mechachal: Online Debates and Elections in Ethiopia – From Hate Speech to Engagement in Social Media." SSRN Scholarly Paper ID 2831369. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2831369.

Ganesh, Bharath. 2018. "The Ungovernability of Digital Hate Culture." *Journal of*

International Affairs 71 (2): 30–49.

George, Cherian. 2016. *Hate Spin: The Manufacture of Religious Offense and Its Threat to Democracy*. Cambridge: MIT Press.

Gerber, Theodore P., and Jane Zavisca. 2016. "Does Russian Propaganda Work?" *The Washington Quarterly* 39 (2): 79–98.

Han, Byung-Chul. 2017. *In the Swarm: Digital Prospects*. MIT Press.

Harsin, Jayson. 2015. "Regimes of Posttruth, Postpolitics, and Attention Economies." *Communication, Culture and Critique* 8 (2): 327–33. https://doi.org/10.1111/cccr.12097.

Haynes, Nell. 2019. "Extreme Speech | Writing on the Walls: Discourses on Bolivian Immigrants in Chilean Meme Humor." *International Journal of Communication* 13 (July). https://ijoc.org/index.php/ijoc/article/view/9109.

Heath, Chip, Chris Bell, and Emily Sternberg. 2001. "Emotional Selection in Memes: The Case of Urban Legends." *Journal of Personality and Social Psychology* 81 (6): 1028–41. https://doi.org/10.1037/0022-3514.81.6.1028.

Hervik, Peter. 2019. "Extreme Speech | Ritualized Opposition in Danish Practices of Extremist Language and Thought." *International Journal of Communication* 13 (July): 18.

Howard, Philip N., and Bence Kollanyi. 2016. "Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2798311.

Kajimoto, Masato and Samantha Stanley, eds. Kanchan Kaur, Shyam Nair, Yenni Kwok, Masato Kajimoto, Yvonne T. Chua, Ma Diosa Labiste, Carol Soon, et al. 2018. "Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3134581.

Keen, Sam. 1986. *Faces of the Enemy: Reflections of the Hostile Imagination*. HarperCollins Publishers.

Lee, Ronan. 2019. "Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis." *International Journal of Communication* 13 (July): 22.

Lewandowsky, Stephan, Werner G. K. Stritzke, Alexandra M. Freund, Klaus Oberauer, and

Joachim I. Krueger. 2013. “Misinformation, Disinformation, and Violent Conflict: From Iraq and the ‘War on Terror’ to Future Threats to Peace.” *American Psychologist* 68 (7): 487–501. https://doi.org/10.1037/a0034515.

Lewis, Rebecca. 2018. *Alternative Influence: Broadcasting the Reactionary Right on YouTube.* New York: Data & Society Research Institute.

Livingston, Steven. 2011. “Africa’s Evolving Infosystems: A Pathway to Security & Stability.” Africa Center Research Paper No. 2, Africa Center for Strategic Studies. https://africacenter.org/publication/africas-evolving-infosystems-a-pathway-to-security-and-stability/.

Lueders, Adrian, Mike Prentice, and Eva Jonas. 2019. “Refugees in the Media: Exploring a Vicious Cycle of Frustrated Psychological Needs, Selective Exposure, and Hostile Intergroup Attitudes.” *European Journal of Social Psychology* 49 (7): 1471–79. https://doi.org/10.1002/ejsp.2580.

Lyons, Matthew N. 2017. *Ctrl-Alt-Delete: The Origins and Ideology of the Alternative Right. Political Research Associates*. https://www.politicalresearch.org/2017/01/20/ctrl-alt-delete-report-on-the-alternative-right.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. “Hate Speech Detection: Challenges and Solutions.” *PLOS ONE* 14 (8): e0221152. https://doi.org/10.1371/journal.pone.0221152.

Mårtensson, Ulrika. 2014. “Hate Speech and Dialogue in Norway: Muslims ‘Speak Back.’” *Journal of Ethnic and Migration Studies* 40 (2): 230–48. https://doi.org/10.1080/1369183X.2013.851473.

Marwick, Alice, and Rebecca Lewis. 2017. “Media Manipulation and Disinformation Online.” New York: Data & Society Research Institute.

Milan, Stefania, and Emiliano Treré. 2019. “Big Data from the South(s): Beyond Data Universalism.” *Television & New Media*20 (4): 319–35. https://doi.org/10.1177/1527476419837739.

Mirchandani, Maya. 2018. “Digital Hatred, Real Violence: Majoritarian Radicalisation and Social Media in India.” Observer Research Foundation.

Nagle, Angela. 2017. *Kill All Normies: Online Culture Wars From 4Chan And Tumblr To Trump And The Alt-Right*. John Hunt Publishing.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://nyupress.org/9781479837243/algorithms-of-oppression/.

O'Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2015. "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems." *Social Science Computer Review* 33 (4): 459–78. https://doi.org/10.1177/0894439314555329.

Onbaşi, Funda Gençoğlu. 2015. "Social Media and the Kurdish Issue in Turkey: Hate Speech, Free Speech and Human Security." *Turkish Studies* 16 (1): 115–30. https://doi.org/10.1080/14683849.2015.1021248.

Ong, Jonathan, and Jason Vincent Cabañes. 2018. "Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines." Communication Department Faculty Publication Series, University of Massachusetts Amherst, January. https://doi.org/10.7275/2cq4-5396.

Perry, Barbara. 2001. *In the Name of Hate: Understanding Hate Crimes*. New York: Routledge.

Persily, Nathaniel. 2017. "The 2016 U.S. Election: Can Democracy Survive the Internet?" *Journal of Democracy* 28 (2): 63–76. https://doi.org/10.1353/jod.2017.0025.

Pohjonen, Matti. 2018. "Horizons of Hate – A Comparative Approach to Social Media Hate Speech." VOX Pol.

Rai, Tage S., Piercarlo Valdesolo, and Jesse Graham. 2017. "Dehumanization Increases Instrumental Violence, but Not Moral Violence." *Proceedings of the National Academy of Sciences* 114 (32): 8511–16. https://doi.org/10.1073/pnas.1705238114.

Richey, Mason. 2018. "Contemporary Russian Revisionism: Understanding the Kremlin's Hybrid Warfare and the Strategic and Tactical Deployment of Disinformation." *Asia Europe Journal* 16 (1): 101–13. https://doi.org/10.1007/s10308-017-0482-5.

Rosenfeld, Michel. 2003. "Hate Speech in Constitutional Jurisprudence: A Comparative Analysis." *Cardozo Law Review* 24 (January): 1523.

Roshani, Niousha. 2016. "Grassroots Perspectives on Hate Speech, Race, and Inequality in Brazil and Colombia." Berkman Klein Center for Internet & Society. https://cyber.harvard.edu/publications/2016/GrassrootsPerspectives.

Rowbottom, Jacob. 2012. "To Rant, Vent and Converse: Protecting Low Level Digital

Speech." *The Cambridge Law Journal*71 (2): 355–83. https://doi.org/10.1017/S0008197312000529.

Schaflechner, J. Forthcoming. "Blasphemy Accusations as Extreme Speech Acts in Pakistan." In *Digital Hate: The Global Conjuncture of Extreme Speech*, edited by Sahana Udupa, Iginio Gagliardone, and P Hervik.

Seta, Gabriele de. Forthcoming. "The Politics of Muhei: Ethnic Humor and Islamophobia on Chinese Social Media." In *Digital Hate: The Global Conjuncture of Extreme Speech.*, edited by Sahana Udupa, Iginio Gagliardone, and P Hervik.

Siapera, Eugenia, Elena Moreo, and Jiang Zhou. 2018. *HateTrack: Tracking and Monitoring Racist Hate Speech Online*. Irish Human Rights and Equality Commission. https://www.ihrec.ie/documents/hatetrack-tracking-and-monitoring-racist-hate-speech-online/.

Soral, Wiktor, Michał Bilewicz, and Mikołaj Winiewski. 2018. "Exposure to Hate Speech Increases Prejudice through Desensitization." *Aggressive Behavior* 44 (2): 136–46. https://doi.org/10.1002/ab.21737.

Sponholz, Liriam. 2016. "Islamophobic Hate Speech: What Is the Point of Counter-Speech? The Case of Oriana Fallaci and The Rage and the Pride." *Journal of Muslim Minority Affairs* 36 (4): 502–22. https://doi.org/10.1080/13602004.2016.1259054.

Stengel, Richard. 2019. *Information Wars: How We Lost the Global Battle against Disinformation and What We Can Do about It*. Atlantic Books.

Stewart, James. 2019. "Anti-Muslim Hate Speech and Displacement Narratives: Case Studies from Sri Lanka and Australia." *Australian Journal of Social Issues* 54 (4): 418–35. https://doi.org/10.1002/ajs4.83.

Stremlau, Nicole. 2012. "Somalia: Media Law in the Absence of a State." *International Journal of Media & Cultural Politics* 8 (September): 159–74. https://doi.org/10.1386/macp.8.2-3.159_1.

Stremlau, Nicole, Emanuele Fantini, and Iginio Gagliardone. 2015. "Patronage, Politics and Performance: Radio Call-in Programmes and the Myth of Accountability." *Third World Quarterly* 36 (8): 1510–26. https://doi.org/10.1080/01436597.2015.1048797.

Tactical Tech. 2019. "Personal Data: Political Persuasion." Accessed March 5, 2020. https://cdn.ttc.io/s/tacticaltech.org/Personal-Data-Political-Persuasion-How-it-works.pdf.

Tanner, Samuel, and Aurélie Campana. 2019. "'Watchful Citizens' and Digital Vigilantism: A Case Study of the Far Right in Quebec." *Global Crime*, April. https://doi.org/10.1080/17440572.2019.1609177.

Taylor, Philip M. 2003. *Munitions of the Mind: A History of Propaganda, Third Edition*. Manchester University Press.

Tell MAMA. 2014. "Rotherham, Hate, and the Far Right Online." Tell MAMA. https://tellmamauk.org/rotherham-hate-and-the-far-right-online/.

Timmermann, Wibke. 2008. "Counteracting Hate Speech as a Way of Preventing Genocidal Violence." *Genocide Studies and Prevention*, December. https://utpjournals.press/doi/abs/10.3138/gsp.3.3.353.

Topinka, Robert J. 2017. "Politically Incorrect Participatory Media: Racist Nationalism on r/ImGoingToHellForThis." *New Media & Society*, June. http://journals.sagepub.com/doi/10.1177/1461444817712516.

Townsend, Emma. 2014. "Hate Speech or Genocidal Discourse? An Examination of Anti-Roma Sentiment in Contemporary Europe." *PORTAL Journal of Multidisciplinary International Studies* 11 (1). https://doi.org/10.5130/portal.v11i1.3287.

Udupa, Sahana. 2019. "Extreme Speech | Nationalism in the Digital Age: Fun as a Metapractice of Extreme Speech." *International Journal of Communication* 13 (July). https://ijoc.org/index.php/ijoc/article/view/9105.

Udupa, Sahana, and Matti Pohjonen. 2019. "Extreme Speech and Global Digital Cultures — Introduction." *International Journal of Communication* 13 (July). https://ijoc.org/index.php/ijoc/article/view/9102.

Waldron, Jeremy. 2012. *The Harm in Hate Speech*. Harvard University Press.

Waltman, Michael S., and Ashely A. Mattheis. 2017. "Understanding Hate Speech." *Oxford Research Encyclopedia of Communication*, September. https://doi.org/10.1093/acrefore/9780190228613.013.422.

Wardle, Claire, and Hossein Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making.* Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

Warner, Michael. 2002. "Publics and Counterpublics." *Public Culture* 14 (1): 49–90.

Wendling, Mike. 2018. *Alt-Right: From 4chan to the White House*. Pluto Press.