HATE SPEECH, INFORMATION DISORDER, AND CONFLICT

FEBRUARY 2020

SAHANA UDUPA, IGINIO GAGLIARDONE,
ALEXANDRA DEEM, AND LAURA CSUKA

Research Review

About the SSRC

The Social Science Research Council (SSRC) is an independent, international, nonprofit organization founded in 1923. It fosters innovative research, nurtures new generations of social scientists, deepens how inquiry is practiced within and across disciplines, and mobilizes necessary knowledge on important public issues.

Introduction

The SSRC Academic Network on Peace, Security, and the United Nations, an initiative of the Council's Conflict Prevention and Peace Forum (CPPF) and its Understanding Violent Conflict (UVC) Program, was established in 2019 out of a request from the United Nations Secretariat to provide UN entities and departments charged with responsibility for peace and security with better, more systematic access to new and emerging research in the academy. The Academic Network also aims to facilitate collaborative engagements between the UN and various academic institutions, research networks, and professional associations working on conflict-management relevant research.

On January 30, 2020, the SSRC convened the second research workshop of its Academic Network on Peace, Security, and the United Nations in New York. This workshop, on "Disinformation, Democratic Processes, and Conflict Prevention," examined the frameworks, findings, and debates in emerging scholarship on information disorder and the linkages between disinformation, elections, hate speech, and identity-based violence around the world. The workshop also explored the ways in which disinformation affects the UN conflict prevention agenda, and how the UN system can better identify, track, and respond to the negative impacts of disinformation in countries and regions where the UN is engaged.

This workshop was organized by CPPF in collaboration with the Council's MediaWell and Next Generation Social Sciences in Africa program.

CPPF commissioned Professor Sahana Udupa of Ludwig Maximilians-Universitat (LMU) Munich to produce a literature review on the intersection of hateful speech, information disorder, and conflict. The review was distributed to workshop participants, together with several MediaWell research reviews.

Academic inquiries on hate speech and information disorder can be approached both as part of a long-term history of researching enablers and triggers of conflict and violence, and as more recent attempts to explain vitriolic and dehumanizing language and imagery powered by the spread of online communication.

Hate speech and information disorder have long been weapons of war and enablers of conflict, used to create and reinforce sentiments of mistrust, exclusion, fear, and anger toward perceived external and internal enemies, and simultaneously to unite allies. Their instrumental use and impact—under the labels of propaganda, information warfare, and psychological warfare—have been widely documented and researched (Taylor 2013).

Recent manifestations of these phenomena and behaviors, however, have been characterized by some unique features, and have encouraged new conceptual and methodological approaches. Hate speech online has emerged as particularly disturbing for its commonness, appropriated and shared by ordinary citizens, rather than its exceptional nature, used to support open confrontations between nations or blocs, and powered by the apparatuses of the state. While the involvement of governments and powerful organized groups (e.g., terrorist organizations) is striking in concerted disinformation campaigns (Richey 2017) and propaganda (Howard and Kolanyi 2016), these tactics have resulted in attempts to exploit fragilities and polarizations within specific national polities, targeting ordinary users as active participants in the spread of hate and disinformation.

From a methodological standpoint, the sheer amount of text and images available through social media has promoted new ways to understand the spread of hateful messages and disinformation across different communities. As we explain below, however, the relative ease of accessing and following what is being said has also tilted research toward mapping how expressions of hatred and disinformation arise and travel, rather than explaining why they emerge and what consequences they may have beyond the spaces where they occur.

Hate Speech: Scope and Approaches

Across disciplinary boundaries, academic scholarship has tended to treat hate speech as distinct from other forms of interpersonal communication, as a specific type of emotional expression that has the ability to reduce empathy and trigger conflicts under specific conditions. Despite recognition of its uniqueness and the potential harms it can produce, however, both the definitions of hate speech and the assessments of the causal links between hate speech and conflict have varied significantly, even more so if we expand the focus beyond academia. Hate speech as a concept has been contested as too wide-ranging and open to manipulation, and narrower conceptions, including "dangerous speech" and "fear speech," have been advanced 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 the divide between American and European approaches to regulating hate speech (Rosenfeld 2001; Bleich 2013). In the United States, the First Amendment protection of freedom of expression stretches well beyond the boundaries of speech tolerated in Europe. The First Amendment approach places an emphasis on identifying a necessary clear and present danger in order to ban or punish certain forms of speech. Numerous European countries, including Germany and France, have adopted instead an approach that bans speech not only because of its likelihood to lead to harm but also for its intrinsic content. For example, with the passage of the controversial Network Enforcement Act (NetzDG) in Germany, social media platforms are required to take down content like hate symbols and derogatory references to immigrants and other vulnerable populations.

To a much lesser extent, studies have scrutinized legal traditions in other countries, exploring, for example, the influence of customary laws or the role religion plays in enabling and restricting freedom of expression (D'Souza, Griffin, and Walt 2018; Edge 2018). Yet, 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). It produces an out-group effect by targeting populations using dehumanizing terms. Target communities are seen as a threat to the safety and values of communities hate speakers claim to represent. 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) exemplifies, a hateful expression aimed at reinforcing out-group dynamics, targeting—and threatening—a specific population, may contain warnings sounding more or less like the following:

> 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 (Waldron 2012).

The same expression can serve to let another category of individuals—those sharing similar views with the speaker—know they are not alone, reinforcing a sense of an in-group that is (purportedly) under threat. 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 (Waldron 2012).

Beyond legal scholarship and security studies, other disciplines have adopted a more eclectic approach, concerned not as much with finding widely shared definitions as with understanding hate speech as a phenomenon affecting specific groups and indicative of 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 2009). This

means written, auditory, or visual texts can construct audiences by creating a relation 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* the addressee 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 perpetrating 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 killed physically. Townsend (2014) has offered a "negative language continuum" comprising hate speech (the least extreme), incitement to genocide (the most extreme), and "genocidal discourse" in the middle of the spectrum, involving "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." His 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.

The expansion of internet-enabled media has compounded the problem of comprehending 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, in actuality, they 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 hate 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 texts 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 aspect 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 owners of 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 have yet to find a coherent framework for theory, definitions, and methods. Wardle and Derakhshan's (2017) three-part typology has gained traction. In their analysis, "information disorder" consists of three types: *dis*information—"information that is false and deliberately created to harm a person, social group, organization or country"; *mis*information—"information that is false, but not created with the intention of causing harm"; and *mal*information—"factual information released to discredit or harm a person or institution, such as doxing, leaks, and certain kinds of hate speech" (2017, 20).

In these definitions, hate speech is associated with malinformation, disinformation, or both. When compared to hate speech, there is an implied sense that information disorder is somehow more closely linked to developments in digital technology. Rather than searching for definitional precision, it is important to recognize that 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 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 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 Muslims (dominant religious groups) against Hindus and Buddhists (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. This technique of framing a religious minority member resulted in mass mob violence even though the post merely tagged the alleged Buddhist perpetrator (i.e., did not even picture him) and featured a pair of white, apparently female feet with painted nails stepping on a Quran. Similarly, in India, studies have shown how digital rumors have spurred mob lynching of minority Muslims by Hindu nationalist groups (Mirchandani 2018). Riots and rumors are not always spontaneous or disorganized. Studies on Asian elections have revealed that political parties actively fund disinformation campaigns to stir unrest and influence voter loyalties (Kaur et al. 2018; Ong and Cabanes 2018; Tan 2019).

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 (Arazna 2015; see also Lewandowsky et al. 2013; Richey 2017; 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 terms of war and make it imprecise in the field of international law. Stengel (2019) identifies a malign chain of cause and effect between disinformation campaigns of ISIS, Russia, and Trump, who all used strategies of weaponizing the grievances of those who felt left out by modernity and globalization. The key strategy is not to "establish falsehoods as true, but rather [to] pollute political discourse such that news information consumers are led to doubt the very concepts of truth and objective political facts" (Richey 2017, 102). These trends are complex because of the involvement of nonstate actors who use information technologies to support asymmetric tactics that spark conflict.

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.

During the MH17 plane crash in Ukraine, for instance, citizen users acted as *curators* of pro-Kremlin disinformation, producing, selecting, and spreading the most popular content about the event on Twitter (Golovchenko, Hartmann, and Adler-Nissen 2018). It is not only the state-supported media monopoly that produces and disseminates propaganda in the context of Russia-Ukraine, but citizens themselves who further their own disenfranchisement by using social media to generate, consume, or distribute disinformation (Mejias and Vokuev 2017). Studies have argued that these developments have undermined the autonomy and agency of civil society in the region.

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).

Cautioning against isolating social media as the site of disinformation, Bennett and Livingston (2018) have argued that disinformation is a systemic problem that reverberates through the interlinked mainstream news media and alternative media ecosystem. In this sense, it is more ordered than disordered and

compounded by disinformation-amplification-reverberation (DAR) cycles. They describe the phenomenon as "disinformation order." This inversion of the more commonly used terminology of "information disorder" highlights the way in which incidents such as the Russian TV crew that attempted to film a fake immigration-related riot in Sweden stem from a modern media landscape where digital and broadcast media are interlinked.

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 Kolanyi 2016; Kajimoto and Stanley 2019; Lewandowsky et al. 2013; Persily 2017; Richey 2017).

Actors, actions, and target groups

To review existing studies through the narrow lens of a causal link between online speech and physical conflict misses a rich body of research that has highlighted new dynamics emerging around online hate and disinformation.

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 role of "ordinary users" as disseminators of disinformation as well as "disinformation innovators" who employ online freelance labor illustrates the new trend. The more horizontal distribution of information agency makes it easier for foreign agents to tap into digital toxicity that transcends national boundaries. These strategies directly benefit from digital communication that is built for instantaneous expression and reaction (Brown 2018). The new communicative paradigm of "the shitstorm" is native to these infrastructural conditions, rendering 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, activities, practices, and 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 2018). Similarly, "muhei stickers" in China that circulate on online messaging apps target Muslim communities by reinforcing slanderous stereotypes through visual ethnic humor (de Seta 2018). Udupa (2019) has defined this phenomenon as "fun as a meta-practice of exclusionary extreme speech." Fun 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 may find expression elsewhere, possibly 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. Daniels (2009) has exposed "cloaked websites" published by white supremacist groups or individuals who conceal authorship in order to disguise their cyber-racism. Ganesh (2018) has argued that three formal features of digital hate cultures make them ungovernable: swarm structure characterized by decentralized networks; exploitation of inconsistencies in web governance between different social media companies as well as between private and government actors that allows hate content to migrate when detected; and the use of coded language to evade content moderation. An important characteristic of digital hate speech is that there are shared language patterns and rhetorical styles among xenophobic groups in different parts of the globe. Siapera, Moreo, and Zhou (2018) have found that there are common vocabularies of hateful speech between "international" alt-right groups and parts of the Irish digital sphere.

Third, groups and communities targeted by online hateful speech and disinformation reveal disturbing continuities as well as

surprising new victims. Racist banter continues to target people of color. Stereotypes against Jews portray them as stingy, conniving, and greedy. A target of vehement hate is the newly invigorated category of "immigrants" denigrated as "refugees" and "asylum seekers," 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 targets Muslims in general, making 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 a frequent target in India and Sri Lanka (George 2016). Online misogyny attacks women more broadly, reflecting the far-reaching consequences of a resurgent Men's Rights Movement that has taken shape as "feminism's doppelganger" (Hodapp 2017) and 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). In Africa, South America, and Asia, groups targeted by online hate speech are ethnic or religious minorities within the nation-state, and immigrants marked by their ethnic and religious identities. For instance, studies have shown that digital racist speech targeting Afro-descendant youth in Brazil and Colombia intensifies discrimination, creates a context of fear and terror, negatively impacts a sense of dignity and self-esteem, silences and isolates victims, fosters a sense of mistrust toward the police and government, and excludes them from economic and political processes, resulting in poverty (Roshani 2016). Similar observations have been made in the case of Bolivian immigrants in northern Chile who are subject to racist stereotypes online (Haynes 2019).

Exposing emerging networks and older patterns, these studies have shown that there are new kinds of perpetrators as well as new ways to target communities through hateful speech and disinformation after the expansion of digital communication. These studies have also emphasized the importance of innovative research methods in examining hate speech and disinformation in the digital age.

Studying hate speech, disinformation, and conflict

Online communication has been offering an unprecedented amount of data for researchers to study mediated social interactions. Critical events can be observed unfolding in real time, or archives can be collected for retrospective analyses, using digital traces—directly—to understand how messages, actors, and platforms interact or—as proxies—to explore broader societal changes and shifting power dynamics.

Legal scholarship as well as disciplines that have increasingly relied on computer-assisted methods to detect and map hate speech and disinformation have been particularly concerned with defining and operationalizing concepts in ways that can enable legal action and machine learning. Many of these studies begin with the premise that hate speech "employs well-known stereotypes which can be distinguished from each other" (Warner and Hirschberg 2012) and are driven by the ambition to create and consolidate "efficient and effective hate speech detectors" (Djuric et al. 2015, 29). In recent years, they have moved beyond relatively simplistic key-words-based approaches—which struggle to detect speech cloaked as humor or satire—and incorporated community based systems to identify hateful messages and train algorithms for further detection (Nobata et al. 2016; Saleem et al. 2017).

These types of studies are offering increasing granularity to test how specific attributes of the speakers—e.g., anonymity (Mondal, Silva, and Benevenuto 2017)—or of the targets (Wei et al. 2017) influence the emergence and spread of hateful messages, and are proving particularly promising for companies and institutions struggling with handling increasingly massive amounts of information. The spirit informing these studies also resonates with similar attempts to research information disorder, detecting which elements allow rumors, fake news, and other forms of disinformation to spread (Bounegru et al. 2018; Heath, Bell, and Sternberg 2001; Petersen, Osmundsen, and Arceneaux 2018), even if computer-assisted methods in the latter case are still in their infancy.

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—under-researched and unanswered. The primary focus of machine-learning models and computational linguistics has been on detection and labeling of data, with insufficient 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, hate speech against Rohingya Muslims is perpetrated via not only 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, Onbaşi (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.

Even more important for the purpose of this review, the vast majority of these studies have had a narrow focus on what is being said or displayed, how and why messages emerge and spread, but have offered very little insight into understanding who the users are (as complete individuals rather than bearers of specific attributes), why they engage in these types of behaviors, and how these forms or language may contribute to creating or sustaining conflict and violence beyond digital spaces.

A better understanding of these important components of digital communication has been offered by disciplines and scholars who have prioritized engagement with specific communities and behaviors rather than with a particular medium of expression. Anthropologists, communication ethnographers, and area study specialists have provided key contributions in this regard. Ong and Cabanes (2018) have revealed a complex business network that has emerged around "disinformation services" in the Philippines. Hundreds of disinformation workers are employed by politicians across the party line to create and spread false and outrageous content online. Adopting the perspective of digital labor, they argue that disinformation innovations such as operating fake accounts for politicians involve similar always-on, flexible, and (self-) exploitative arrangements that characterize other online freelance work. It is accompanied by the added emotional labor of having to justify this work to others and themselves. Ong and Cabanes 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.

Studying interactions between radio and mobile phones, research emerging from Kenya, Uganda, and Somalia has illustrated the complexities of how callers have learned to exploit audiences' belief that new spaces of interactions are supposedly freer from power, allowing anybody to vent their anger and denounce wrongdoings, to actually 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). In the words of Harsin (2015), truth and power in "post-truth regimes" are demonstrated through popular participation and attention. These studies have examined the life worlds, motivations, and "career trajectories" of actors who peddle vitriol and disinformation, explaining how flagging these actors simply as self-serving manipulators risks missing complex realities on the ground, and 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. In addition to quantifying the prevalence and significance of hate speech in relation to other kinds of online communication, this approach has highlighted how antagonistic messages can also be used to attack those in power in ways that can lay the foundations for more pervasive and widespread forms of contestation. Livingston (2011) has found that across the African continent, older analog communication technologies like radio and newspapers are hubs of politically motivated disinformation, while digital communication technologies are positioned as a means for equalizing the media field, giving citizens access to information that could serve as a corrective against disinformation.

Across disciplines, few studies have asked direct questions on the connections between hate speech and/or information disorder and conflict. Among those that have done so, prevailing research strategies have included case studies (Al-Zaman 2019; Lewandowsky et al. 2013) and experimental methods (Lueders, Prentice, and Jonas n.d.; Rai, Valdesolo, and Graham, 2017; Soral, Bilewicz, and Winiewski, 2018). 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, and that the resulting dehumanization may increase the likelihood of instrumental violence. Limitations, however, also exist in these cases. Building on a long tradition of using experiments to test hypotheses on human behavior, these studies have relied on small cohorts of individuals tested in controlled environments and exposed to selected inputs, which are often removed from what occurs in real-world scenarios. Using survey-based methods, a small number of studies have investigated the impact of disinformation in terms of differences in cultural perception 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 (FCO), and international organizations such as the United Nations are frequent funders of projects that seek to counter violent extremism as well as attempts to use online media for recruitment and radicalization (Ferguson 2016). Increasingly, this has also become the language adopted by internet companies such as Facebook, Twitter, and Google as they have come under increasing pressure by the US and European governments to address extremist speech that incites violence online (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 that employ disproportionately fewer women and people of color (Noble 2018).

Some studies have emphasized the value of counterspeech in combating online hateful speech and disinformation (ARTICLE 19 2015; 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 (2016) 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 society-wide counternarratives to combat hate speech, which has been at the center of initiatives aimed at countering violent extremism (CVE). Counternarratives entail challenging the prevailing narrative that is used to promote violence and offering a different, more positive and inclusive narrative instead. In some cases, counternarratives are used explicitly to support certain groups (e.g., by encouraging and facilitating access to the airwaves of moderate religious leaders that may not be encouraging violence, or by certain political groups that subscribe to a particular political ideology or peacebuilding roadmap). A counternarrative can be articulated through speech (e.g., broadcasts, pamphlets, articles) or symbols (e.g., the use of dress that directly counters the codes of a group perpetrating violence, references to poetry that resonates with the ideas of earlier generations that may have had a more inclusive and peaceful vision of society, or songs that reflect overarching political goals the are attempting to supplant the more dangerous narratives). On a larger scale, the CVE approach to counternarratives has been adopted by governments, international organizations, and local organizations. For example, in Nigeria, the Partnership Against Violent Extremism (PAVE), a network of nearly 40 NGOs, is formed to provide a coordinated CVE approach to Boko Haram, the insurgency group in the north that has been responsible for widespread violence.

Such community-based solutions are now complemented with a range of online tools for detection and removal of online hate speech and disinformation. For instance, networks such as truly. media have developed platforms to integrate verification tools including Google Maps, TinEye, WolframAlpha, Google Reverse Image Search, Yandex, Snopes, and Pipl for quick collaboration among fact checkers. However, a peculiar problem has surfaced in this regard. The response to information disorder has largely been to generate more information in terms of verification and publication of "correct" information. Conceptually, this response is distinct from restrictive approaches to content such as prefiltering and takedowns. By spurring parallel processes of fact checking and publication, responses to information disorder in academic studies as well as public initiatives have fed the rationale of generating more information. Thereby, they have, if unwittingly, contributed to the market logic of data-hungry digital capitalism. Relatedly, despite the contributions of myriad actors to the issue, solutions to disinformation often point toward the journalistic profession and associated institutions such as newsrooms, journalism schools, and professional journalists' ethics boards. This is a strange approach considering how much of information disorder comes from institutions that are beyond the control of professional journalism. This is not to discount severe shortcomings ailing professional journalism worldwide. Some of these include low public trust in media, proclivity for sensationalism, and lack of resources for fact-checking and investigative journalism—problems that are also partly a result of disinformation campaigns (Marwick and Lewis 2017; see also Andrejevic 2015).

There is also a glaring need to bring historical contextualization 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 & 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.

For instance, Al-Zaman (2019) has emphasized the need to take a broader contextual view of digital disinformation in Bangladesh. The unique contribution of digital disinformation to the manifestation of conflict on the ground is the way in which a single, relatively small piece of information can rapidly achieve a broad circulation and mobilize a large following through emotional appeal. Such mobilization and the perpetration of real-world violence, however, does not derive solely from the informational

infrastructure of digital technologies but is also dependent on coalition-building activities by interested parties. While digital technologies have facilitated disinformation, the reasons behind this pattern are manifold and include historical, geopolitical, and social factors such as the relationship between Bangladesh and neighboring countries (India, Pakistan), low information literacy in the context of a developing country, and a long history of ethnoreligious tensions and traumas related to the interweaving of religion and politics during colonial rule.

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 ethnographic sensibility to specific cultural contexts, 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 the colonial rule that have been revived and weaponized by current regimes, including those aimed against people within one's own national communities.

Recent studies on digital disinformation and election integrity have adopted a similar non-digital media centric perspective. Tan (forthcoming) has proposed electoral management digital readiness (EMDR) index for election management boards (EMB) in East and Southeast Asian countries. In developing this index, she has suggested that aspects that influence an EMB's ability to productively address the spread of disinformation include whether it is supported by a strong rule of law and adequate digital infrastructure, the degree to which it collaborates with social media companies and fact-checking organizations to flag and remove malicious content, and its investment in cybersecurity infrastructure and machine learning and data analytics tools (see also Goh and Soon 2019). Bennett and Livingston (2018) emphasize the role of changing media systems, calling for policy actions that account for different elements of the emerging "disinformation order" supporting the new right: political parties and politicians, foreign agents and governments inciting information warfare, and disinformation entrepreneurs. Similarly, Benkler, Faris, and Roberts (2018) have urged for reflection on larger systemic causes rather than myopic focus on Facebook algorithms or Russian interference as primary instigators of disinformation-related unrest in the US.

In order to address these gaps and challenges, what is urgently needed is interdisciplinary collaboration between computational scientists and scholars engaged in qualitative research on the practices, histories, and cultures of media and society. This should be coupled with concerted efforts to place pressure on social media companies to provide data access to researchers. Such interdisciplinary efforts can advance the current focus on the labeling and detection of hate speech and disinformation toward holistic comprehension and critical engagement for context-sensitive solutions.

References

Al-Zaman, Md. Sayeed. 2019. "Digital Disinformation and Communalism in Bangladesh." *China Media Research* 15 (2): 68–76.

Andrejevic, Mark. 2015. *Infoglut: How Too Much Information is Changing the Way We Think and Know*. New York: Routledge.

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

Arazna, Marzena. 2015. "Conflicts of the 21st Century Based on Multidimensional Warfare: 'Hybrid Warfare,' Disinformation and Manipulation." *Security and Defense Quarterly* 8 (3): 103-129, http://doi.org/10.5604/23008741.1189421.

ARTICLE 19. 2015. *'Hate Speech' Explained: A Toolkit*. Retrieved from https://www.article19.org/resources/hate-speech-explained-a-toolkit/.

Austin, J. L. 1975. *How to Do Things with Words*. Oxford University Press.

Back, Les. 2002. "Wagner and Power Chords: Skinheadism, White Power Music, and the Internet." In *Out of Whiteness*, edited by Vron Ware and Les Back, 94–132. Chicago, IL: University of Chicago Press.

Bangstad, Sindre. 2014. *Anders Breivik and the Rise of Islamophobia*. London, UK: Zed Books.

Benesch, Susan. 2012. "Dangerous Speech: A Proposal to Prevent Group Violence." World Policy Institute, New York, NY. Retrieved from 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." Working paper, United States Holocaust Memorial Museum, Washington, DC. Retrieved from https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf.

Benkler, Yochai, Robert Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford: Oxford University Press.

Bennett, Lance, and Steven Livingston. 2018. "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions." *European Journal of Communication* 33 (2): 122–39. https://doi.org/10.1177/0267323118760317.

Bleich, Erik. 2013. "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.

Bounegru, L., Gray, J., Venturini, T., & Mauri, M. (2018). "A Field Guide to 'Fake News' and Other Information Disorders." A Field Guide to "Fake News" and Other Information Disorders: A Collection of Recipes for Those Who Love to Cook with Digital Methods, Public Data Lab, Amsterdam.

Brisset-Foucault, Florence. 2016. "Serial Callers: Communication Technologies as a Canvas for Political Personhood in Contemporary Uganda." *Ethnos* 83 (2): 255–273.

Brown, Alexander. 2018. What Is So Special about Online (as Compared to Offline) Hate Speech?" *Ethnicities* 18 (3): 297–326. https://doi.org/10.1177/1468796817709846.

Butler, Judith. 1997. *Excitable Speech: A Politics of the Performative*. New York: Routledge.

Buyse, Antoine. 2014. "Words of Violence: 'Fear Speech,' or How Violent Conflict Escalation Relates to the Freedom of Expression." *Human Rights Quarterly* 36 (4): 779–797.

Charland, Maurice. 2009. "Constitutive Rhetoric: The Case of the Peuple Québécois." *Quarterly Journal of Speech* 73 (2): 133–150.

Citron, Danielle Keats, and Helen Norton. 2011. "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age." *Boston University Law Review* 91: 1435–1484.

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): 141–160. https://doi.org/10.1080/1057610X.2018.1513984.

Daniels, Jessie. 2009. "Cloaked Websites: Propaganda, Cyber-Racism, and Epistemology in the Digital Era. *New Media & Society* 11 (5): 659–83. https://doi.org/10.1177/1461444809105345.

Daniels, Jessie. 2018. "The Algorithmic Rise of the Alt-Right." *Contexts* 17 (1): 60–65. https://doi.org/10.1177/1536504218766547.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. *Automated Hate Speech Detection and the Problem of Offensive Language* (arXiv:1703.04009v1 [cs. CL]).

Deem, Alexandra. 2019. "The Digital Traces of #whitegenocide and Alt-Right Affective
Economies of Transgression." *International Journal of Communication* 13: 3183–3202.

De Genova, Nicholas. 2010. "Migration and Race in Europe: The Trans-Atlantic Metastases of a Postcolonial Cancer." *European Journal of Social Theory* 13: 405–419.

De Seta, Gabriele. 2018. "Wenming Bu Wenming: The Socialization of Incivility in Postdigital China." *International Journal of Communication* 12: 2010–2030.

Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. "Hate Speech Detection with Comment Embeddings." *Proceedings of the 24th International Conference on World Wide Web*, 29–30.

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." *University of New South Wales Law Journal* 41 (3): 1–38.

Edge, Peter W. 2018. "Oppositional Religious Speech: Understanding Hate Preaching." *Ecclesiastical Law Journal* 20: 278–289. https://doi.org/10.1017/S0956618X18000467

EUFRA [European Union Agency for Fundamental Rights]. 2018. *Hate Crime Recording and Data Collection Practices across the EU*.

Faris, Robert, Ashar Amar, Urs Gasser, and Daisy Joo. 2016. "Understanding Harmful Speech Online." Networked Policy Series, Berkman Klein Center Publication Series No. 2016–18. http://dx.doi.org/10.2139/ssrn.2882824.

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

Forelle, Michelle, Phil Howard, Andres Monroy-Hernandez, and Saiph Savage. 2015. "Political Bots and the Manipulation of Public Opinion in Venezuela." Available at https://arxiv.org/abs/1507.07109.

Foxman, Abraham H., and Christopher Wolf. 2013. *Viral Hate: Containing Its Spread on the Internet*. New York: Palgrave MacMillan.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35 (4): 665–668. https://doi.org/10.1080/10584609.2018.1477506.

Ganesh, B. (2018). "The Ungovernability of Digital Hate Culture." *Journal of International Affairs* 71 (2): 30–49.

Gagliardone, Iginio. 2016. "'Can You Hear Me?' Mobile–Radio Interactions and Governance in Africa." *New Media & Society* 18 (9): 2080–2095.

Gagliardone, Iginio, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright, Mulatu Moges, Michael Seifu, Nicole Stremlau, Patricia Taflan, Tewodros Makonnen Gebrewolde, and Zelalem Teferra. 2016. *Mechachal: Online Debates and Elections in Ethiopia-From Hate Speech to Engagement in Social Media*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2831369.

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.

Goh, Shawn, and Carol Soon. 2019. "Governing the Information Ecosystem: Southeast Asia's Fight against Political Deceit." *Public Integrity* 21: 523–536.

Goldstein, R. J. 2001. *Political Repression in Modern America*. Urbana & Chicago: University of Illinois Press.

Golovchenko, Yevgeniy, Mareike Hartmann, and Rebecca Adler-Nissen. 2018. "State, Media and Civil Society in the Information Warfare over Ukraine: Citizen Curators of Digital Disinformation." *International Affairs* 94 (5): 975–994. https://doi.org/10.1093/ia/iiy148.

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

Harsin, Jayson. 2015. Regimes of posttruth, postpolitics, and attention economies. *Communication, Culture & Critique* 8 (2): 327–33. https://doi.org/10.1111/cccr.12097.

Haynes, Nell. 2019. "Writing on the Walls: Discourses on Bolivian Immigrants in Chilean Meme Humor." *International Journal of Communication* 13: 3122–3142.

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): 1028–1041. https://doi.org/10.1037/0022-3514.81.6.1028.

Hervik, Peter. 2019. "Ritualized Opposition in Danish Online Practices of Extremist Language and Thought." *International Journal of Communication* 13: 3104–3121.

Hodapp, Christa. 2017. *Men's Rights, Gender, and Social Media*. London: Lexington Books.

Howard, Philip N., and Bence Kolanyi. 2016. "Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum." Working Paper 2016.1, Project on Computational Propaganda, Oxford, UK. http://dx.doi.org/10.2139/ssrn.2798311.

Kajimoto, M., and Stanley, S. 2019. "Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3134581.

Kaur, Kanchan, and Shyam Nair, Yenni Kwok, Masato Kajimoto, Yvonne T. Chua, Ma. Diosa Labiste, Carol Soon, Hailey Jo, Lihyun Lin, Trieu Thanh Le, and Anne Kruger. "Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam." 2018. Available at SSRN: https://ssrn.com/abstract=3134581 or http://dx.doi.org/10.2139/ssrn.3134581.

Keen, Sam. 1986. *Faces of the Enemy: Reflections of the Hostile Imagination*. San Francisco: Harper and Row.

King, Gary, Jennifer Pan, Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111 (3): 484–501. https://doi.org/10.1017/S0003055417000144.

Lee, Rohan. 2019. "Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis." *International Journal of Communication* 13: 3203–3224.

Lewandowsky, Stephan, Werner G. K. Stritzke, Alexandra M. Freund, Klaus Oberauer, and Joachim 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, NY: Data & Society Research Institute. https://datasociety.net/output/alternative-influence/.

Livingston, Steven. 2011. "Africa's Evolving Infosystems: A Pathway to Security and Stability." Africa Center Research Paper no. 2, Africa Center for Strategic Studies, Washington, DC. https://africacenter.org/publication/africas-evolving-infosystems-a-pathway-to-security-and-stability/.

Lueders, Adrian, Mike Prentice, and Eva Jonas. N.d. "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): 1471–1479. https://doi.org/10.1002/ejsp.2580.

Lyons, Matthew. 2017. *Crtl-Alt-Delete: The Origins and Ideology of the Alternative Right*. Political Research Associates. http://www.politicalresearch.org/2017/01/20/ctrl-alt-delete-report-on-the-alternative-right/?print=print#Part2.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yan, 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): 230–248. https://doi.org/10.1080/1369183X.2013.851473.

Marwick, Alice, and Rebecca Lewis. 2017. *Media Manipulation and Disinformation Online*. New York, NY: Data & Society Research Institute. https://datasociety.net/output/media-manipulation-and-disinfo-online/.

Mejias, Ulises, and Nikolai Vokuev. 2017. "Disinformation and the Media: The Case of Russia and Ukraine." *Media, Culture & Society* 39 (7): 1027–1042. http://doi.org/10.1177/0163443716686672.

Milan, Stefania, and Vidushi Marda. 2018. *Wisdom of the Crowd: Multistakeholder Perspectives on the Fake News Debate*. Philadelphia: Internet Policy Observatory, Annenberg School, University of Pennsylvania.

Milan, Stefania, Emiliano Trere. 2019. "Big Data from the South(s): Beyond Data Universalism." *Television & New Media* 20 (4): 319–335.

Mirchandani, M. 2018. *Digital Hatred, Real Violence: Majoritarian Radicalisation and Social Media in Indi*a (No. 167; ISBN : 978-93-88262-27-9). New Delhi.

Mondal, Mainack, Leandro Araujo Silva, and Fabricio Benevenuto. 2017. "A Measurement Study of Hate Speech in Social Media." *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94.

Nagle, Angela. 2017. *Kill All Normies*. Winchester, UK: Zero Books.

Nobata, Chikasi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. "Abusive Language Detection in Online User Content." *Proceedings of the 25th International Conference on World Wide Web*, 145–153.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O'Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Padraig Cunningham. 2015. "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems." *Social Science Computer Review* 33 (4): 459–478. https://doi.org/10.1177/0894439314555329.

Onbaşi, Funda Gencoglu. 2015. "Social Media and the Kurdish Issue in Turkey: Hate Speech, Free Speech and Human Security." *Turkish Studies* 16 (1): 115–130. https://doi.org/10.1080/14683849.2015.1021248.

Ong, Jonathan Corpus, and Jason Vincent A. Cabanes. 2018. "Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines." Newton Tech4Dev Network: University of Leicester and De La Salle University. https://doi.org/10.7275/2cq4-5396.

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

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

Petersen, Michael Bang, Mathias Osmundsen, Kevin Arceneaux. 2018. "A 'Need for Chaos' and the Sharing of Hostile Political Rumors in Advanced Democracies." PsyArXiv. September 1. https://doi.org/10.31234/osf.io/6m4ts.

Pohjonen, Matti. 2018. *Horizons of Hate: A Comparative Approach to Social Media Hate Speech*. VOX-Pol Network of Excellence. https://www.voxpol.eu/download/vox-pol_publication/Horizons-of-Hate.pdf.

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): 8511–8516.

Richey, Mason. (2017). "Contemporary Russian Revisionism: Understanding the Kremlin's Hybrid Warfare and Strategic and Tactical Deployment of Disinformation." *Asia Europe Journal* 16 (1): 101–113. https://doi.org/10.1007/s10308-017-0482-5.

Rosenfeld, Michel. 2001. "Hate Speech in Constitutional Jurisprudence: A Comparative Analysis." *Cardozo Law Review* 24: 1523–1567.

Roshani, Niousha. 2016. "Grassroots Perspectives on Hate Speech, Race, and Inequality in Brazil and Colombia." Berkman Klein Center Research Publication No. 2016-18, Harvard University, Cambridge, MA. http://dx.doi.org/10.2139/ssrn.2882234.

Rowbottom, Jacob. 2012. "To Rant, Vent and Converse: Protecting Low Level Digital Speech." *Cambridge Law Journal* 71 (2): 355–383.

Saleem, Haji Mohammad, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. "A Web of Hate: Tackling Hateful Speech in Online Social Spaces." ArXiv Preprint ArXiv:1709.10159.

Schaflechner, J. (In review). "Blasphemy Accusations as Extreme Speech Acts in Pakistan." *In Digital Hate: The Global Conjuncture of Online Extreme Speech*, edited by S. Udupa, I. Gagliardone, and P. Hervik. Indiana University Press.

Siapera, Eugenia, Elena Moreo, and Jiang Zhou. 2018. *Hate Track: Tracking and Monitoring Racist Speech Online.* Dublin: Irish Human Rights and Equality Commission. Retrieved from http://hdl.handle.net/10197/9916.

Soral, Wiktor, Michal Bilewicz, and Mikolaj Winiewski. 2018. "Exposure to Hate Speech Increases Prejudice through Desensitization." *Aggressive Behavior* 44 (2): 136–146. 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): 502–522. https://doi.org/10.1080/13602004.2016.1259054.

Stengel, Richard. 2019. *Information Wars: How We Lost the Global Battle Against Disinformation*. New York: Atlantic Monthly Press.

Stewart, James. 2019. Anti-Muslim Hate Speech and Displacement Narratives: Case Studies from Sri Lanka and Australia. *Australian Journal of Social Issues* 54: 418–435. 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 (2–3): 159–174.

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): 1510–1526. https://doi.org/10.1080/01436597.2015.1048797.

Tactical Tech. 2019. *Personal Data: Political Persuasion Inside the Influence Industry. How it Works*. Available at https://cdn.ttc.io/s/tacticaltech.org/Personal-Data-Political-Persuasion-How-it-works.pdf.

Tan, N. Forthcoming. "Electoral Management of Digital Campaigns and Disinformation in East and Southeast Asia."

Tanner, Samuel, and Aurelie Campana. 2019. "'Watchful Citizens' and Digital Vigilantism: A Case Study of the Far-Right in Quebec." *Global Crime*, 1–21. https://doi.org/10.1080/17440572.2019.1609177.

Taylor, Philip M. 2013. *Munitions of the Mind: A History of Propaganda from the Ancient World to the Present Era*. Manchester and New York: Manchester University Press.

Tell MAMA. 2014. *Facebook Report: Rotherham, Hate, and the Far-Right Online*. Retrieved from 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: An International Journal* 3 (3): 353–374. http://dx.doi.org/10.3138/gsp.3.3.353.

Topinka, Robert J. 2018. "Politically Incorrect Participatory Media: Racist Nationalism on R/ImGoingToHellForThis." *New Media & Society* 20 (5): 2050–69.

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), 2–22. https://doi.org/10.5130/portal.v11i1.3287.

Udupa, Sahana. 2019. "Nationalism in the Digital Age: Fun as a Metapractice of Extreme Speech." *International Journal of Communication* 13: 3143–3163.

Udupa, Sahana, and Matti Pohjonen. 2019. "Introduction: Extreme Speech and Global Digital Media Cultures." *International Journal of Communication* 13: 3019–3067.

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

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

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

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

Warner, William, and Julia Hirschberg. 2012. "Detecting Hate Speech on the World Wide Web." *Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics*, 19–26.

Wei, Xiaocong, Hongfei Lin, Liang Yang, and Yihai Yu. 2017. "A Convolution-LSTM-Based Deep Neural Network for Cross-Domain MOOC Forum Post Classification." *Information* 8(3): 92. https://doi.org/10.3390/info8030092.

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