International Journal of Communication 13(2019), 3068–3087 1932–8036/20190005

Defining Online Hate and Its "Public Lives": What is the Place for "Extreme Speech"?

IGINIO GAGLIARDONE
University of the Witwatersrand, South Africa

Following Sahana Udupa and Matti Pohjonen's (2019) invitation to move the debate beyond a normative understanding of hate speech, this article seeks to build a foundation for conceptual and empirical inquiry of speech commonly considered deviant and disturbing. It develops in three stages. It first maps the public lives of terms that refer to online vitriol and how they have been used by different communities of researchers, politicians, advocacy groups, and national organizations. Second, it shows how different types of "haters" have been interpreted as parts of "swarms" or "armies," depending on whether their violent potential emerges around critical incidents or whether they respond to longer-term strategies through which communities and their leaders tie their speech acts to explicit narratives. The article concludes by locating "extreme speech" within this broader conceptual tapestry, arguing that the paternalistic gaze that characterizes a lot of research on online hate speech is tied to what Chantal Mouffe has referred to as the "moralization of politics," a phenomenon that cannot be matched by responses that are themselves moral.

Keywords: extreme speech, hate speech, violent extremism, moralization of politics, narratives, and counter-narratives.

The concept of "extreme speech" has emerged to both signal and surpass the normative bias toward recognizing, isolating, and combating vitriolic online speech practices. Aimed at promoting theoretical and methodological innovation, stressing the importance of comprehension over classification and the need to connect online practices with specific cultures of communication, it has become part of a dense constellation of concepts seeking to describe practices often labeled as "deviant" or "disturbing," but that have become an increasingly defining trait of contemporary communication.

This article examines this constellation and the concepts that belong to it, focusing on their "public lives" and on how they connect with and are used by specific communities of researchers, politicians, advocacy groups, and national organizations. It is divided into three parts. Resonating with other efforts to develop a genealogical approach to hate speech (Shepherd, Harvey, Jordan, Srauy, & Miltner, 2015), the first part seeks to offer a typology of terms that have appeared to qualify specific aspects of online vitriol—from "fear speech" to "violent extremism"—and to chart their relationships.

Iginio Gagliardone: iginio.gagliardone@wits.ac.za
Date submitted: 2018–03–12

Copyright © 2019 (Iginio Gagliardone). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at http://ijoc.org.

The second part takes stock of this survey to illustrate how different terms have mapped onto different social phenomena. Hate speech and the concepts gravitating around it have tended to be associated with *swarming* behaviors, usually involving ordinary users coalescing around events and incidents, simultaneously responding to and reinforcing discriminatory practices that have become progressively accepted in the societies they belong to. Violent extremism has more often been used to refer to recruits, rather than ordinary citizens; planned activities, rather than incidents; and all-encompassing narratives, rather than individual speech acts—all elements that evoke the image of an *army* acting to achieve a common goal, rather than a response to less discernible, lingering transformations occurring within a specific social formation.

The third part illustrates how extreme speech, as a concept and construct, can help to turn the normative bias toward recognizing, isolating, and combating specific forms of speech into more analytical efforts to understand them, together with the actors behind expressions of hate and intolerance.

The progression of the article is also meant to highlight a fundamental tension that has characterized terms such as hate speech and violent extremism and has been poignantly captured by Robert Post (2009). As he has observed:

> Hate speech regulation imagines itself as simply enforcing the given and natural norms of a decent society […]; but from a sociological or anthropological point of view we know that law is always actually enforcing the mores of the dominant group that controls the content of law. (Post, 2009, p. 130)

Hate speech and extremism are terms with strong political connotations, and analytical efforts to define what may or may not be labeled as such have tended to go hand in hand with more practical impulses to act, contain, and punish. When some academics —anthropologists in particular—have sought to go deeper, observing, for example, the practices of communities engaging in flame wars or trolling (Lee, 2005; Walker, 2008) and seeking to uncover the values attached by users to their behaviors, they have been accused of condoning deviant practices and siding with the haters (Jane, 2014).

The concept of "extreme speech" seeks to reaffirm the legitimacy—even the necessity—of this latter type of explorative and analytical inquiries, considering even the most extreme forms of expression as objects in need of an explanation, rather than something that can be fixed or wished away.

Defining Online Vitriol: A Survey of Terms and Their Communities

This article builds on two expert surveys carried out as part of separate projects developed with the support of the Holocaust Museum in Washington, D.C., and of UNESCO. Both surveys involved interviews with key national and international organizations that were variously engaged in addressing hate speech and violent extremism as part of their core mandates (e.g., No hate speech movement, the Sentinel Project for Genocide Prevention, Sisi Ni Amani, Umati), as elements of their public engagement with critical social issues (e.g., Ford Foundation, MacArthur Foundation), as components of broader efforts

in shaping media systems (e.g., Internews, BBC Media action, Albany Associates), or as increasingly critical elements emerging as a by-product of their commercial activities (e.g., Facebook, Google).[1]

These surveys offered the opportunity to understand how different terms have tended to be privileged by different types of organizations. For example, dangerous speech has offered a framework especially to local NGOs seeking to map trends and potentially violent outbursts around critical events (e.g., elections); and violent extremism has been a particular concern of governments in the Global North, but has also slowly been adopted by other political actors—especially in authoritarian regimes—to justify the closing of online spaces and the persecution of political opponents. Tech companies such as Facebook, Google, and Twitter have often adopted different terminologies—including both hate speech and terrorism—to define the limits of what is and is not allowed on the platforms they own, but each term has led to developing distinctive strategies in how these forms of expression are dealt with (e.g., efforts to combat terrorism and radicalization have often led to closer collaborations with national governments). The focus on organizations has led to privilege terms that have not only been at the center of academic debates, but have also been used to inform practical efforts to respond to the proliferation of online vitriol, thus excluding other related concepts such as e-bile, incivility, or trolling, which have been discussed in academic literature (Buckels, Trapnell, & Paulhus, 2014; Jane, 2014; Shin, 2008) but have been of lesser practical concern to these organizations.

The following sections discuss these terms and the communities that have adopted them to justify their actions and inform their strategies, seeking to highlight similarities and differences among various approaches and how they seek to address distinctive phenomena connected to online vitriol.

Hate Speech, Dangerous Speech, and Fear Speech: Too Broad or Too Narrow?

Hate speech is a broad and emotive concept. It lies in a complex nexus with freedom of expression; individual, group, and minority rights; and concepts of dignity, liberty, and equality. As a phenomenon, it calls into question some of the most fundamental principles on which societies are built. The answers each society has developed to balance between the freedom of expression and the respect for equality and dignity have created unique rifts and alliances. Much comparative research on hate speech has focused on the divide that exists between the American and European approaches to regulating hate speech (Bleich, 2014; Hare & Weinstein, 2010; Rosenfeld, 2002). The United States has protection for the freedom of expression that stretches well beyond the boundaries of speech that is tolerated in Europe. Its emphasis on the clear and present danger that is necessary to be identified to ban or punish certain forms of speech has emerged as the defining characteristic of this approach. Numerous European countries, including Germany and France, have instead adopted an approach that not only bans

[1] The full list of organizations interviewed included: No hate speech movement, Media Smarts, Justice Base, Panzagar, ADL, the Sentinel Project for Genocide Prevention, the Online Hate Prevention Institute, Facebook, MacArthur Foundation, Ford Foundation, Microsoft, Google, Internews, Digital Rights Foundation, Konrad Adeauer Foundation, CITAD Nigeria, Institute for Social Accountability Kenya, Sisi Ni Amani, Umati, International media support, Institute of Media Law, Ofcom, Search for Common Ground, La Benevolencjia, BBC Media Action, OSCE, Article 19.

forms of speech because of their likelihood to lead to harm, but also for their intrinsic content (Heinze, 2009; Mbongo, 2009).

Other societies have developed unique mechanisms to identify and counter hate speech, which may variously combine customary law and formal law. In Somalia, for example, where poetry constitutes a popular vehicle for the dissemination of information and ideas, poets who repeatedly compose poems that community elders consider to be derogatory of individuals or groups can be banned from composing new work (Stremlau, 2012). Important research has emerged from the study of the role hate speech plays in atrocities or major outbursts of violence (Kellow & Steeves, 1998; Thompson, 2007). But systematic research examining the phenomenon of hate speech and its regulation beyond the United States and Europe is still marginal.

At a global level, multilateral treaties, such as the International Covenant on Civil and Political Rights (ICCPR), have sought to address the challenge of defining hate speech by reverting to the narrower concept of "incitement to violence." Multistakeholder processes (e.g., the Rabat Plan of Action) have been initiated to bring greater clarity and to suggest mechanisms to identify these types of messages. Under Article 20 (2) of the ICCPR, and also in different conditions under Article 4 (a) of the International Convention on Elimination of All Forms of Racial Discrimination, states are obliged to prohibit expression that amounts to incitement to discrimination, hostility, or violence. At the same time, under international and regional standards, states are also obliged to protect and promote—both in legislation and practice—the rights of equality and nondiscrimination.

Internet intermediaries—organizations that mediate online communication, such as Facebook, Twitter, and Google—have also become increasingly powerful actors in defining the limits of freedom of expression, advancing their own definitions of hate speech and binding users to rules that set the boundaries of what is and is not permissible. These definitions have come under scrutiny, especially because of the nature of these platforms as private spaces for public expression, hosting an increasingly significant amount of national and global conversations. Responding to mounting criticism, companies such as Facebook have introduced measures to offer greater transparency in the process leading to the removal of specific types of content and have sharpened their tools to communicate with users who are flagging messages as inappropriate or disturbing.[2]

Despite multiple efforts to define the contours of hate speech, in everyday discourse the term has tended to be used well beyond the boundaries set by international bodies, national bodies, and Internet companies. Incidents (e.g., hate crimes against immigrants or members of a minority group) and critical events (e.g., elections) have played an important role in raising attention around the power of speech to lead to violence. They have also promoted debates that largely ignore the long-term definitional efforts described above and have been open to manipulation by political entrepreneurs. Empirical studies that have explored popular perceptions of hate speech and compared them to those used among scholars and

[2] Germany, for example, has forced Facebook to take a more aggressive stance on hate speech and fake news to continue freely operating in the country (Deutsche Welle, 2017).

in policy circles (iHub Research, 2013) have highlighted how personal insults, propaganda, and negative commentary about politicians tend to be referred to as hate speech.

Building on these expansive definitions, accusations of "hate speech" have also been employed by governments and those in positions of power to shut down legitimate debates on matters of public interest. Before the Ethiopian elections of 2015, for example, a group of bloggers seeking to pressure the government to respect fundamental freedoms included in Ethiopia's own constitution were arrested with the accusation of producing speech that could destabilize the country (Gagliardone & Pohjonen, 2016). This has extended to judgments that assume speakers intentionally advocated harm when their intent may have been more frivolous—an ill-judged or flippant comment on social media, for example; a nuanced satire intended to provoke a debate on a challenging issue; or through art. As Rowbottom (2012) reported, in the United Kingdom, individuals have been sentenced for posting hateful content on Facebook while drunk or for joking about blowing up an airport after their flights were canceled. In instances where the speech is offensive or provocative, the labeling of the expression as hate speech could overstate the connection between the speech and the alleged harm—either by misjudging the influence of a speaker or the likelihood of harm occurring—or overlooking the propensity of individuals to engage in counter speech that has a strong and positive impact.

Indeed, such practices have led to the implication that all hate speech is unlawful, calling for criminal or other sanctions that might be inappropriate or ineffective and might indeed hinder democratic outcomes. For instance, this interpretation could lead to increased policing and state or private surveillance of discourse, including online, and encourage overreliance on censorship instead of addressing systemic discrimination. In cases where there are open hostilities among groups, the use of the term "hate speech" could have consequences that go in the opposite direction of what is intended. Defining an expression as a hate-speech act may increase the audience of its speakers, especially if they are able to position themselves as "martyrs" of censorship or to frame unsuccessful attempts at censorship as a vindication of their views. Arguably, this could have more severe consequences than a lack of response (Mouffe, 2018). For these reasons, alternative, more narrowly defined concepts—such as "dangerous speech" or "fear speech"—have been proposed, focusing more on the propensity of expression to cause widespread violence.

Dangerous Speech

"Dangerous speech," as a term, has been used in the context of inciting mass violence, or as defined by Benesch (2012), expressions that have a significant probability of "catalyzing or amplifying violence by one group against another, given the circumstances in which [they were] made or disseminated" (p. 1).

More specifically, dangerous speech increases the risk of mass violence targeting certain people because of their membership in a group on an ethnic, a religious, or a racial basis. Dangerous speech can capture any form of expression that constitutes incitement to violence or conditions its audience to accept, condone, and commit violent acts against people who belong to a targeted group. For example, Hutu

extremists were able to incite genocide in Rwanda in part because years of propaganda had influenced Hutus to view Tutsis as less than human and so dangerous that they must be eliminated from the country.

Dangerous speech may include various expressions, such as verbal or written speech, traditional print and broadcast media pieces, blogs and social media posts, images or symbols, or even music and poetry. Its message may be interpreted as a call to violence against a specific group or may portray this group in a way that legitimizes or justifies violence against it. To this end, dangerous speech often uses dehumanizing imagery or symbols when referring to a group (e.g., by analogizing its members to vermin), accusing the group of planning to harm the audience, and generally presenting the group's existence as a threat to the audience.

The dangerous-speech framework is sensitive to the climate in which a dangerous-speech act takes place. It invites practitioners, researchers, and policymakers to consider speakers' influence (e.g., speakers with a high degree of influence over the audience—such as a political, cultural, or religious leader or media personality—will likely have influence over a crowd); audience receptiveness (e.g., the audience may have grievances and fears that the speaker can cultivate); speech content (i.e., content that may be taken as inflammatory by the audience and understood as a call to violence and can use coded language to do this); medium of dissemination, including the language used and the medium for dissemination, which may be the sole or primary source of news for the relevant audience, replacing the "marketplace of ideas"; social or historical speech context (e.g., longstanding competition among groups for resources, lack of efforts to solve grievances, or previous episodes of violence may provide a climate propitious for violence; Benesch, 2012, pp. 3–5).

In this way, dangerous speech is seen as a potential early warning signal for violence since it is often a precursor—if not also a prerequisite—for mass violence. It may therefore also be possible to limit such speech or its dangerousness. Dangerous speech also provides a framework through which speakers may be held accountable for speech that constitutes crime.

Nevertheless, it is extremely difficult to accurately and systematically predict the likely effect of speech on an audience. A change in the political climate or the audience's responsiveness to a particular speaker can significantly alter the propensity of a violent corollary of a speech act. For example, as further described below, since the rise of the League, a far-right political party, in Italy in 2018, episodes of violence against migrants have significantly increased (Tondo & Giuffrida, 2018). Identifying dangerous-speech acts therefore requires an appreciation of the historical, political, and socioeconomic context in which the expression is communicated to gauge the likely consequences or impact of speech.

Fear Speech

The concept of "fear speech" (Buyse, 2014; George, 2015) has been recently advanced to emphasize language that can incite the fear in one group that "the other group" plans to use violence or even destroy them in a near future. Also, based on an interdisciplinary study of mass atrocities, the idea of fear speech offers a pathway to understand whether the preconditions for violence may gradually

emerge. Ultimately, the sowing of extreme anxiety is a better predictor for the escalation of violence than adopting more generic analytical frameworks such as hate speech.

Buyse (2014) identifies several factors that can accelerate words eliciting violence. Both the content of the message and its context are relevant. Similar to other frameworks, not only does it matter who makes a certain statement (i.e., how influential someone is), but also which media are used and what their scope is. More specifically, Buyse highlights that an important predictor is to look at the frames through which possible violence between two groups has been interpreted or justified in the past. For example, is a violent act explained as an incidental crime or as part of a large-scale threat of violence from one group toward the other? Is a group or a person blamed, and does this involve the use of stereotypes? The more inhuman the stereotypes used, the lower the threshold to using violence. The threshold is also lower if, in a certain statement, nonviolent solutions are rejected as an option. Such language is able to progressively create a siege mentality and can ultimately lead to legitimizing violent acts as defensive of a group's safety or integrity.

While jurisprudence is one of the instruments for addressing violent escalations among groups, education and promoting awareness about the mechanisms that increase or decrease violence are equally considered important instruments for preventing the escalation in violence within this framework. At the same time, the idea that keeping the public debate alive and propose alternative frameworks can reduce the risk of an escalation in violence should be tested against other paradigms that have stressed how deep rooted grievances cannot simply be fought through more speech or counterspeech (Coleman, 2004).

Violent Extremist Speech

"Violent extremist speech" and its counterpart—what is increasingly being referred to as Countering Violent Extremism (CVE) or Preventing Violent Extremism (PVE)—have rapidly become a common way of describing speech that can lead to violent outcomes. CVE appears to be more commonly used in the United States and by the U.S. government and security organizations, while PVE is more commonly used by international organizations such as the United Nations.

Governmental agencies such as the U.S. 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 U.S. and European governments to address extremist speech that incites violence online (Andrews & Seetharaman, 2016). While prohibitions to the incitement of terrorism are often mentioned alongside hate speech in Internet intermediaries' terms of reference, the practical removal of different typologies of speech may vary significantly. While hate speech cases commonly involve ordinary users flagging content and companies responding to such requests on an individual basis, when it comes to violent extremism, government agencies have been given a "trusted" flagger status to prioritize their reporting of dangerous or illegal material (Brown & Cowls, 2015; Malik, Laville, Cresci, & Gani, 2014).

CVE speech generally refers to measures aimed at preventing individuals from radicalizing and reversing the process of those who have already been radicalized. A distinctive trait of these initiatives has been their tendency to deal with speech acts not in isolation (as it may happen in cases aimed at assessing whether an individual has produced messages that may be equated to hate speech), but as part of larger "narratives." Narratives, in this context, refer to particular world views or ways of framing events. Radicalized individuals seeking to incite violence advance a particular narrative that strategic communications companies, NGOs, or funding organizations may try to counter with another competing narrative of events (Archetti, 2012; Ferguson, 2016).

Militant groups such as ISIS have engaged in large-scale propaganda campaigns, producing copious amounts of audiovisual materials to desensitize individuals to violence and promote their cause. With increasingly high production values, technical resources, and vast video archives, propaganda has been continuously distributed to local affiliates and adapted to local audiences in different cultural and geographic contexts (Mahlouly & Winter, 2018). Indeed, their centralized media organization acts as an information repository, operating to supply content for others to disseminate. In addition to the vast collection of audiovisual content collated by these terrorist organizations' media wings, there has also been a shift by extremist organizations away from using static websites to using social media platforms. This shift has been made in part to eliminate the risk of hosting such material and in part to allow them to reach wider audiences in a more interactive fashion. Social networking platforms, however, have increasingly sought to develop strategies to respond to extremist content, forcing some organizations—especially violent jihadi groups—to move to encrypted channels such as Telegram or file-sharing sites such as Pastebin or, in the case of the extreme right, to migrate toward other platforms such as VKontake, where content is less often removed.

Although there is not a definitive formulation, it appears that CVE counternarratives have four key objectives: preventing violent extremism (changing behavior, namely violence and incitement); preventing extremism (changing minds); protecting one's country or region from violent extremist influence; and preventing the violent extremist narrative from spreading (Hedayah, 2014). It means challenging the prevailing narrative that is being 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).

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) is a network of nearly 40 NGOs formed to provide a coordinated CVE approach to Boko Haram, the insurgency group in the north that has been responsible for widespread violence. PAVE has received

international funding for media engagement and ongoing CVE initiatives to develop and promote positive and counternarratives in conjunction with its network.

CVE and efforts to change narratives, either through propaganda or restricting voices, have been criticized by some rights groups. The emphasis on counternarratives has also come under criticism (Archetti, 2015). They are, by definition, reactive; they are countering an already established narrative that has been given the space to take root and grow. As some strategic communication organizations have pointed out, the term "counternarrative" is an acceptance of failure; it implies having lost ground and stresses the need to get it back.

In March 2016, the NGO Article 19, along with 58 human rights organizations and NGOs, wrote an open letter to the UN Human Rights Council, urging them to consider serious concerns about certain initiatives around countering and preventing violent extremism. The letter argued that "the lack of an agreed definition for 'violent extremism' opens the door to human rights and other abuses, compounded by the danger of conflating itself with 'terrorism' and thereby leading to the overbroad application of 'counterterrorism' measures." In doing so, they state that there is a risk that CVE initiatives have the potential to undermine human rights to equality and freedom from discrimination, as well as the right to privacy and the freedom of expression, association, and religion or belief. As with hate speech and dangerous speech, labeling expressions as violent extremist speech can be used as a tool of government power and can undermine and censor the legitimate concerns raised by political opponents, journalists, and human rights defenders. In other words, CVE can be added to a growing political lexicon that provides governments with further grounds to stifle freedom of expression and crush dissent.

Other Terms: From Cyberhate to Microaggression

"Cyberhate" is another broad term closely tied to hate speech that is also increasingly used to refer to online speech with the potential of inciting violence. It has been referred to as the "globalization of hate" or having the ability to use the virtual environment to stretch across borders and encourage transnational groups and communities that are able and willing to incite violence (Bakalis, 2018; Burnap & Williams, 2015; Glassman, 2000; Perry & Olsson, 2009). The Anti-Defamation League (ADL) defines "cyberhate" as any use of electronic communications technology to spread anti-Semitic, racist, bigoted, extremist, or terrorist messages or information. These electronic communications technologies include the Internet (i.e., websites, social networking sites, "Web 2.0" user-generated content, dating sites, blogs, online games, instant messages, and e-mail) as well as other computer– and cell phone–based information technologies (Anti-Defamation League, 2016).

Reflections on "cyberhate" have built on earlier debates about hate crimes and have questioned whether specific measures are needed to persecute acts committed online, as compared with those perpetrated in the physical world (Bakalis, 2018). They have asked, for example, whether the permanence and publicity of messages appearing on social media may have particular psychological and practical consequences on the targets of hateful speech and have highlighted how constant exposure to public shaming may, in the long term, increase feelings of vulnerability and fear of physical harm among vulnerable populations (Awan & Zempi, 2017).

Similarly, in recent years, the term "microaggression" has gained increasing salience and has been applied—particularly in university college environments—to identify behaviors and unconscious biases that threaten a particular group. The term "microaggression" has been conceptualized as "the brief and commonplace daily verbal, behavioral and environmental indignities, whether intentional, that communicate hostile, derogatory or negative racial, gender, and sexual orientation, and religious slights and insults to the target person or group" (Sue et al., 2007, p. 271). However—and crucially—these indignities do not necessarily have to be the outcome of intentional behavior. Instead, "perpetrators of microaggressions are often unaware" of the indignities that they inflict on others (Sue et al., 2007, p. 271).

This focus on the unconscious or unwitting dimension of microaggression is important—and, in part, distinguishes microaggression from more intentional forms of hate speech or dangerous speech. "Microaggressions are often unconsciously delivered in the form of subtle snubs or dismissive looks, gestures, and tones" (Sue et al., 2007, p. 273); they are encounters underpinned by a pervasive unconscious bias against those of different ethnicities and racial groups, women, LGBT individuals, and disability groups.

Categories of statements that could be conceived as microaggressions are those that refer to an individual as an alien in his or her own land; the ascription of intelligence; criminality or the assumption of criminal status; pathologizing cultural values or communication styles; ascription as a second-class citizen; the use of sexist or heterosexist language; and the use of traditional gender role prejudicing and stereotyping. In this context, microagressions can be an important component of laying the groundwork for violence; the term reflects an ongoing process of subtle "othering."

Microaggressions can be humiliating, and in their most pernicious forms can compromise a person's dignity and give an individual a sense that important people think he or she doesn't really belong, imposing real psychological damage (Sunstein, 1995). However, microaggressions can also have detrimental effects on the groups that produce them—typically white, male, heterosexual, and middle class—contributing to a sense of prejudice and superiority. While the issue of microaggression has risen in the popular consciousness and associated policies are adopted to counter them, it is important to recognize that if charges of microaggression are taken too far, they can impose a stifling orthodoxy, undermining freedom of thought and expression. In other words, as speech acts can inflict verbal violence and trauma, this can also lead to excessive scrutiny of dialogue and conversation.

Of Swarms and Armies

The comparative analysis of the broad variety of concepts "extreme speech" enters in conversation with offers the opportunity to appreciate significant differences in how related phenomena have been categorized and how distinctive measures have been developed in response to them. The distinction between hate speech and its related concepts—dangerous speech and fear speech—and violent extremism in particular provides insight to understand how different potentials of speech to lead to harm, and associated aspects of human behavior have been recognized and addressed by different actors.

Most initiatives centered around the concepts of hate speech and dangerous speech have focused on critical moments in the political life of a national community—elections above all—or have emerged in response to incidents that have acted as warnings of potentially violent trends emerging in society (e.g., attacks by ordinary citizens targeting minority groups or immigrants).

The Ford Foundation, for example, has sponsored numerous initiatives in West Africa aimed at preventing media from exacerbating violence. Through the organization Media Development for West Africa, during the 2012 and 2016 Ghanaian elections, it allowed monitoring extremist, hateful, and violent speech on 70 radio stations across Ghana. Biweekly media monitoring reports were then released, identifying politicians and media outlets spreading hate messages. The UMATI project and the Institute for Social Accountability (TISA) in Kenya and the Centre for Information Technology and Development (CITAD) in Nigeria have operationalized the Dangerous Speech framework, incorporating local context to map trends in online hate around elections. Ahead of the 2018 Italian elections, Amnesty International developed a "Barometer of Hate," following key politicians on social media and seeking to map how the language they used could be tied to the growing xenophobic sentiments in the country.[3]

When it comes to incidents, mounting anti-immigrant feelings in Europe and the rise in hate crimes (Weaver, 2016) have created the need for mechanisms to map and understand new trends and their significance. The eMORE project, for example, has connected different national research institutes across Europe to "test and transfer a knowledge model on online hate speech and offline hate crime" (para. 1). In Asia, the proliferation of online messages inciting hatred against the Rohingya minority in Myanmar has led to the emergence of initiatives to promote counter-speech and alert toward gravest cases of online vitriol potentially leading to violence.[4]

This emphasis on critical moments and incidents suggests how hate speech, especially in its online form, has tended to be treated as a lingering phenomenon, quietly and progressively permeating societies, but whose significance can be fully understood only around specific events. An image that can capture this quality, both of hate speech as a phenomenon and of the measures developed in response to it, can be that of a swarm, of a connected network of individuals responding to similar stimuli and showing emerging behaviors in ways that often appear unpredictable.[5] Unpredictability is one of the elements that has contributed to generating a feeling of discomfort around hate speech and its consequences and of how it can manifest itself through acts carried out by ordinary citizens. Although a climate of intolerance is the product

[3] The is a long list of projects that have sought to address hate speech in critical circumstances. For an updated overview, see for example the Electoral Foundation for Electoral Systems (www.ifes.org).

[4] This has been the case of the NGO Panzagar, which was created to offer a middle ground for debates on minorities in Myanmar and put pressure on big players such as Facebook to act and limit the spread on online vitriol against the Muslim minorities.

[5] The connection between hate speech and swarming behaviors has also interestingly been captured by popular culture. One of the most highly rated episodes of the dystopic TV series *Black Mirror*, "Hated in the nation," describes how words of hate published in social media against specific targets could lead swarms of robotic bees developed to compensate the global disappearance of their biological counterparts to actually attack the targets of hate.

of complex interactions among agendas promoted by political entrepreneurs, media reporting, and popular opinion, it is often the case that ordinary users, without a significant track record of violence or continued political engagement, are those responsible for posting messages amounting to hate speech.

Shifting the focus on violent extremism, both the analysis of the phenomenon and the measures developed in response to it point to a different imagery. Emphasis has been on recruits, rather than ordinary citizens; on planned activities, rather than incidents; and on all-encompassing narratives, rather than individual speech acts. The image of an army, rather than of a swarm, seems a better fit to succinctly capture this combination of traits. A reflection on the leaders creating a political space for different typologies of individuals to act, and on the motivations to act they evoke, points to some significant differences. In the case of violent extremism, leaders have tended to position themselves outside the boundaries of what is considered civil; they have similarly appeared unapologetic about their agendas and refer to a future yet to come when some of the values and behaviors considered deviant or dangerous can become the norm. In the case of hate speech, especially of the kind occurring in regimes considered democratic, leaders seem to have chosen instead to play within the boundaries of what is considered civil; when accused of fomenting hate speech, they have sought to justify their actions with the need to protect values that are under threat or have ducked down, waiting for attention to move on. The imagery they have evoked has tended to be reactive and nostalgic, pointing at something that is under attack and needs protection.

The distinction between swarms and armies resonates with similar analyses that have emerged, for example, in the field of networked politics, seeking to unpack and understand the behavior of different types of political organizations (Kahler, 2009; Mueller, 2010). Milton Mueller has suggested distinguishing between "associative clusters," as "unbounded and decentred clusters of actors [emerging] around repeated patterns of exchange or contact" (Mueller, 2010, p. 41) and "networked organizations," as "consciously constructed organizations [whose members] design their relationship among a bounded set of individuals or organizations to pursue a common objective" (Mueller, 2010, p. 42). Margaret Keck and Kathryn Sikkink (1998), studying activists and campaigners, have illustrated how even loosely connected networks can reach their goals by imposing a new language and creating a new imagery that can progressively become the norm.

Swarms and armies should not be considered mutually exclusive categories. Swarming behaviors can progressively turn into more coordinated efforts to act against specific targets. In contemporary Europe, political entrepreneurs on the far right have capitalized on lingering sentiments of fear and disquiet toward immigrants and consolidated an imagery leading to more frequent and coordinated violent attacks toward minorities (Tondo & Giuffrida, 2018; Weaver, 2016). However, the distinction between the two can offer an additional analytical lens to understand peculiar aspects of related phenomena, their possible evolution, and the responses they are likely to trigger.

How is Extreme Speech Different?

This brief survey of terms used to refer to online vitriol and of the communities coalescing to address it as a phenomenon can provide a useful vantage point to uniquely locate the concept of "extreme speech" and its contributions to on-going debates.

As Pohjonen and Udupa (2017) have explained, and this Special Section reaffirms, the idea of extreme speech should be interpreted as a program of research, connecting different scholars and sensitivities rather than as a defined and rigid concept. There are some distinctive traits, however, that the comparison with other terms brings into relief.

The first is the fact that existing terms have been closely linked to pressures "to do something about it." Both scholarly and policy work around hate speech and violent extremism may have recognized the importance of a better understanding of the phenomena, but these have tended to prioritize the need to recognize, isolate, and combat what falls under one or the other rubric. This tendency has been further exacerbated by the increasing availability of tools able to collect and analyze large quantities of text. The growing interest in hate speech and its salience as a phenomenon have led researchers to embark on ambitious programs of research seeking to chart the emergence of hate speech in different national contexts or targeting specific groups (Faris, Ashar, Gasser, & Joo, 2016; Fortuna & Nunes, 2018; Olteanu, Castillo, Boy, & Varshney, 2018). Correlations between online hate speech and offline cases of violence have started to be more systematically mapped (Müller & Schwarz, 2017). Many of these efforts, however, have been informed by a need to identify and catalogue speech in ways that can be algorithmically processed and automated, but they have contributed little toward understanding why hate speech emerges and spreads and what value is attached to it by those who engage in it.

The second-related but more nuanced trait characterizing the debate on extreme speech is the recognition of the double bind in which other terms, the communities adopting these terms, and the answers they have produced have been caught.

Hate speech in its contemporary form is strictly interconnected to what Chantal Mouffe has referred to as the moralization of politics (Mouffe, 2005). The distinction between right and left has become less relevant, replaced instead by a distinction between right and wrong, good and evil. As she pointed out, politics has not disappeared; it is increasingly played on the moral register.

Different communities, at different points in their histories, have coined terms that serve to affirm this distinction, disconnecting one group from another and severing the need to know more about a specific issue because the source from which it is being discussed has been judged a priori as one from which nothing worth listening can be coming. In Silvio Berlusconi's Italy, this disconnecting, antagonizing word was "communist." In Trump's America, "liberal" has increasingly served this function. In contemporary Ethiopia, the term "Woyane" has been used when targeting the ruling class, and "chauvinist," "anti-peace," or "anti-constitution" have been used when targeting the opposition (Gagliardone, 2016). This disconnection clearly does not amount to hate speech. It is instead the beginning of the relationship of antagonism, the creation of an enemy, the start of a process that can lead to more or less aggressive forms of expression. Much depends on what the group being targeted is, what power it holds in relation to other groups in a given society, and how much its presence threatens existing balances.

The second element of this double bind depends on the fact that a lot of hate speech—as observed, as read, as witnessed—is hideous and that the response that is likely to trigger in the

researcher or in the outsider peeking in, not in the intended audience, is of judgment and disconnect. Also this reaction plays on the moral register.

As Stanley Fish (1997) noted, this moral response often amounts to a kind of "Oh Dearism." It expresses discomfort and disdain, and it is likely to emerge from a supposedly higher moral ground from which the observer claims to be speaking. It ties with what Fish has defined as "boutique multiculturalism," flagging liberal responses to hate speech as hypocritical and self-righteous, something that can read more or less like "you hater are not allowed to sit at our dinner table, of rational thinkers" (p. 393). This liberal, moral gaze is what, as is described in research by Polletta and Callahan (2017), for example, is likely to infuriate extremists even more.

The still-evolving debate on "extreme speech" points out how responses to phenomena that closely depend on—and express—the moralization of political life cannot play on the moral register. A change of perspective is needed. Carol McGranahan's (2017) anthropology of lying, for example, has stressed the need not to ask how to correct lies, but how to understand lies and liars in their cultural, historical, and political contexts. Francesca Polletta's take on fake news has similarly stressed the importance of interrogating not why people believe such stories, but rather examining the pleasure that comes from reading surprising news (Polletta & Callahan, 2017).

Recognizing some forms of expression as "extreme speech" means suppressing the urge to catalog and judge and accept—with Chantal Mouffe—conflict as the natural site of democracy. Hate speech, against this background, should not be considered the exception, but rather the extreme form of a relation of contestation. As Barbara Perry has illustrated in her research on hate crimes, they, like hate speech, are better understood not as an aberration, but rather as a by-product of a society struggling to face changes in power relations in the context of shrinking opportunities (Perry, 2002).

This recognition does not come without risks. Researchers adopting the lenses of "extreme speech" may be accused of condoning practices considered abhorrent and potentially dangerous. This potential criticism cannot be solved a priori. It is only the rigor and sensitivity of researchers accepting the challenge of engaging in these types of enquiries and the balance achieved in each individual case that can offer an answer and practically illustrate the value of this approach. Similar research agendas have been pursued in other fields of study, including the analysis of criminal organizations (Gambetta, 2011; Varese, 2011) and of perpetrators of hate crimes (Chakraborti & Garland, 2015; Dunbar, 2003; Valeri & Borgeson, 2017), and lessons can be learned from these types of engagements.

This change of perspective also opens the possibility of investigating processes and relationships that are often ignored and overwhelmed by the need to classify, contain, and combat online vitriol. What is, for example, the relationship between hate speech and inequality? In research conducted in Ethiopia, it was found that the majority of antagonistic speech, as well as the most virulent forms of hate speech, were coming from individuals with little or no influence, not by ringleaders or political entrepreneurs (Gagliardone et al., 2016). Vitriolic comments appeared as a bitter response to power and to those who represent it. In Trump's America, emerging research has begun to pinpoint links between hate crimes and rising inequalities (Majumder, 2017).

These findings must be substantiated by additional research, but they suggest that the links between power inequalities and the more aggressive tones used by individuals with little influence should be further examined before dismissing certain types of speech as inappropriate. Especially in contexts where individuals enjoy few opportunities to affect change in a concrete way or have access to outlets in which they can voice discontent or criticism, social media offers an opportunity to voice frustration, not just toward the speakers, but at the processes through which power is exercised more broadly.

Extreme speech, as a research agenda, is not meant to surpass other efforts to map online vitriol, but to unlock paths that have been so far been rarely pursued and to contribute to developing a fuller understanding of the reasons why certain behaviors are on the rise, rather than flagging them as unacceptable.

References

Andrews, N., & Seetharaman, D. (2016, February 12). Facebook steps up efforts against terrorism. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/facebook-steps-up-efforts-against-terrorism-1455237595

Anti-Defamation League. (2016, March). *Responding to cyberhate. Progress and trends*. Retrieved from https://www.adl.org/sites/default/files/documents/assets/pdf/combating-hate/2016-ADL-Responding-to-Cyberhate-Progress-and-Trends-Report.pdf

Archetti, C. (2012). *Understanding terrorism in the age of global media: A communication approach*. London, UK: Springer.

Archetti, C. (2015). Terrorism, communication and new media: Explaining radicalization in the digital age. *Perspectives on Terrorism*, *9*, 49–59.

Awan, I., & Zempi, I. (2017). "I will blow your face OFF"—VIRTUAL and physical world anti-muslim hate crime. *The British Journal of Criminology*, *57*, 362–380.

Bakalis, C. (2018). Rethinking cyberhate laws. *Information & Communications Technology Law*, *27*, 86–110.

Benesch, S. (2012). Dangerous speech: A proposal to prevent group violence. *World Policy Institute*. Retrieved from http://www.worldpolicy.org/sites/default/files/Dangerous%20Speech%20Guidelines%20Benesch%20January%202012.pdf

Bleich, E. (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*, 283–300.

Brown, I., & Cowls, J. (2015, November 23). *Check the web: assessing the ethics and politics of policing the Internet for extremist material*. Retrieved from https://apo.org.au/node/58979

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

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*, 223–242.

Buyse, A. (2014). Words of violence: Fear speech or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, *36*, 779–797.

Chakraborti, N., & Garland, J. (2015). *Responding to hate crime: The case for connecting policy and research*. Bristol, UK: Policy Press.

Coleman, P. T. (2004). Paradigmatic framing of protracted, intractable conflict: Toward the development of a meta-framework-II. *Peace and Conflict*, *10*, 197–235.

Deutsche Welle. (2017, March 14). *Germany to force Facebook, Twitter to delete hate speech*. Retrieved from https://www.dw.com/en/germany-to-force-facebook-twitter-to-delete-hate-speech/a-37927085

Dunbar, E. (2003). Symbolic, relational, and ideological signifiers of bias-motivated offenders: Toward a strategy of assessment. *American Journal of Orthopsychiatry*, *73*, 203–211.

Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). Understanding harmful speech online. *Berkman Klein Center Research Publication*, *21*, 1–19.

Ferguson, K. (2016). Countering violent extremism through media and communication strategies. *Reflections*, *28*, 1–40.

Fish, S. (1997). Boutique multiculturalism, or why liberals are incapable of thinking about hate speech. *Critical Inquiry*, *23*, 378–395.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*, 1–30.

Gagliardone, I. (2016). *The politics of technology in Africa*. Cambridge, UK: Cambridge University Press.

Gagliardone, I., & Pohjonen, M. (2016). Engaging in polarized society: Social media and political discourse in Ethiopia. In B. Mutsvairo (Ed.), *Digital activism in the social media era* (pp. 22–44). London, UK: Springer International Publishing.

Gagliardone, I., Pohjonen, M., Beyene, Z., Zerai, A., Aynekulu, G., Bekalu, M., . . . Teferra, Z. (2016). *Mechachal: Online debates and elections in Ethiopia—from hate speech to engagement in social media*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2831369

Gambetta, D. (2011). *Codes of the underworld: How criminals communicate*. Princeton, NJ: Princeton University Press.

George, C. (2015). Managing the dangers of online hate speech in South Asia. *Media Asia*, *42*, 144–156.

Glassman, E. (2000). Cyber hate: The discourse of intolerance in the new Europe. In L. Hengel (Ed.), *Culture and technology in the new Europe: Civic discourse in transformation in post-communist nations* (pp. 145–164) . Stamford, CT: Ablex Publishing.

Hare, I., & Weinstein, J. (2009). *Extreme speech and democracy*. Oxford, UK: Oxford University Press.

Hedayah (2014). *Developing effective counter-narrative frameworks for countering violent extremism.* Retrieved from http://www.hedayahcenter.org/Admin/Content/File-3032016135442.pdf

Heinze, E. (2009). Wild-west cowboys versus cheese-eating surrender monkeys: Some problems in comparative approaches to hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 182–203). Oxford, UK: Oxford University Press.

iHub Research. (2013). *Umati final report*. Nairobi, Kenya. Retrieved from https://preventviolentextremism.info/sites/default/files/Umati%20Final%20Report.pdf

Jane, E. A. (2014). "Your a ugly, whorish, slut." Understanding e-bile. *Feminist Media Studies*, *14*, 531–546.

Kahler, M. (2009). *Networked politics: Agency, power, and governance*. Ithaca, NY: Cornell University Press.

Keck, M., & Sikkink, K. (1998). *Activists beyond borders: Advocacy networks in international politics*. Ithaca, NY: Cornell University Press.

Kellow, C. L., & Steeves, H. L. (1998). The role of radio in the Rwandan genocide. *Journal of Communication*, *48*, 107–128.

Lee, H. (2005). Behavioral strategies for dealing with flaming in an online forum. *The Sociological Quarterly*, *46*, 385–403.

Mahlouly, D., & Winter, C. (2018). *A tale of two Caliphates. Comparing the Islamic State's internal and external messaging priorities*. Retrieved from https://icsr.info/wp-content/uploads/2018/07/ICSR-Report-A-Tale-of-Two-Caliphates-Comparing-the-Islamic-State's-Internal-and-External-Messaging-Priorities.pdf

Majumder, M. (2017, January 23). Higher rates of hate crimes are tied to income inequality. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/

Malik, S., Laville, S., Cresci, E., & Gani, A. (2014, September 24). Isis in duel with Twitter and YouTube to spread extremist propaganda. *The Guardian*. Retrieved from https://www.theguardian.com/world/2014/sep/24/isis-twitter-youtube-message-social-media-jihadi

Mbongo, P. (2009). Hate speech, extreme speech, and collective defamation in French law. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 221–236). Oxford, UK: Oxford University Press.

McGranahan, C. (2017). An anthropology of lying: Trump and the political sociality of moral outrage. *American Ethnologist*, *44*, 243–248.

Mouffe, C. (2005). *On the political*. London, UK: Psychology Press.

Mouffe, C. (2018, September 10). Populists are on the rise but this can be a moment for progressives too. *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2018/sep/10/populists-rise-progressives-radical-right

Mueller, M. (2010). *Network and states. The global politics of Internet governance*. Cambridge, MA: MIT Press.

Müller, K., & Schwarz, C. (2017). *Fanning the flames of hate: Social media and hate crime*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *ArXiv Preprint ArXiv.* Retrieved from https://arxiv.org/abs/1804.05704

Perry, B. (2002). *In the name of hate: Understanding hate crimes*. London, UK: Routledge.

Perry, B., & Olsson, P. (2009). Cyberhate: the globalization of hate. *Information & Communications Technology Law*, *18*, 185–199.

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

Polletta, F., & Callahan, J. (2017). Deep stories, nostalgia narratives, and fake news: Storytelling in the Trump era. *American Journal of Cultural Sociology*, *5*, 392–408.

Post, R. (2009). Hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 123–139). Oxford, UK: Oxford University Press.

Rosenfeld, M. (2002). Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo L. Rev.*, *24*, 1523–1567.

Rowbottom, J. (2012). To rant, vent and converse: Protecting low level digital speech. *The Cambridge Law Journal*, *71*, 355–383.

Shepherd, T., Harvey, A., Jordan, T., Srauy, S., & Miltner, K. (2015). Histories of hating. *Social Media + Society*, *1*, 1–10.

Shin, J. (2008). Morality and Internet behavior: A study of the Internet troll and its relation with morality on the Internet. In *Society for Information Technology & Teacher Education International Conference*, *1*, 2834–2840.

Stremlau, N. (2012). Somalia: Media law in the absence of a state. *International Journal of Media & Cultural Politics*, *8*, 159–174.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, *62*, 271–286.

Sunstein, C. (1995). Democracy and the problem of free speech. *Publishing Research Quarterly*, *11*, 58–72.

Thompson, A. (2007). *The media and the Rwanda genocide*. Ottawa, Canada: IDRC.

Tondo, L., & Giuffrida, A. (2018, August 3). Warning of "dangerous acceleration" in attacks on immigrants in Italy. *The Guardian*. Retrieved from https://www.theguardian.com/global/2018/aug/03/warning-of-dangerous-acceleration-in-attacks-on-immigrants-in-italy

Udupa, S. & Pohjonen, M. (2019). Extreme speech and global digital cultures: Introduction. *International Journal of Communication 13* (this Special Section).

Valeri, R. M., & Borgeson, K. (2017). *Skinhead history, identity, and culture*. London, UK: Routledge.

Varese, F. (2011). *Mafias on the move: How organized crime conquers new territories*. Princeton, NJ: Princeton University Press.

Walker, B. (2008). Deliberative democracy in online lesbian communities. *Feminist Media Studies*, *8*, 197–223.

Weaver, M. (2016, September 28). "Horrible spike" in hate crime linked to Brexit vote, Met police say. *The Guardian*. Retrieved from https://www.theguardian.com/society/2016/sep/28/hate-crime-horrible-spike-brexit-vote-metropolitan-police