Genocide Studies and Prevention: An International Journal

Volume 18
Issue 1 *Evidence-Based Approaches to Preventing Genocide and Mass Atrocities*

Article 9

10-2-2024

Dangerous Speech as an Atrocity Early Warning Indicator: Measuring Changing Conflict Dynamics

Catherine Buerger
Dangerous Speech Project

Susan Benesch
Dangerous Speech Project

Recommended Citation

Buerger, Catherine and Benesch, Susan (2024) "Dangerous Speech as an Atrocity Early Warning Indicator: Measuring Changing Conflict Dynamics," *Genocide Studies and Prevention: An International Journal*: Vol. 18: Iss. 1: 84–95.
DOI:
https://doi.org/10.5038/1911-9933.18.1.1955

Available at: https://digitalcommons.usf.edu/gsp/vol18/iss1/9

Dangerous Speech as an Atrocity Early Warning Indicator: Measuring Changing Conflict Dynamics

Catherine Buerger
Dangerous Speech Project
Washington, DC, U.S.A.

Susan Benesch
Dangerous Speech Project
Washington, DC, U.S.A.

Introduction

Identifying precursors to mass violence is a vital component of genocide prevention since it can reveal opportunities for preventing, or at least diminishing, atrocities. In the months or years before an outbreak of atrocities, in-group norms shift toward perceiving mass violence as acceptable or even necessary. Public discourse that encourages this view of violence is an instrument of such normative change, and a useful indicator that it is underway. Detecting such shifts, especially in their early stages, would offer opportunities for atrocity prevention. A new, sensitive early warning system for impending violence could be developed and, in turn, used to design efforts to forestall the normative shift in favor of violence. The first step for either early warning or intervention is to build methods for detecting normative shifts in favor of mass violence. Detection should occur as early as possible, as interventions are more likely to succeed in these cases. This paper describes such a method.

Human communication is a pathway for changing human attitudes, including expectations about the beliefs and behavior of other people. We have observed striking similarities in the public language used by political and social leaders—and then echoed and spread by their followers—in extended periods before mass violence in a wide variety of cases. Such language is meant to convince members of an in-group that outsiders (or insiders depicted as sympathetic to the out-group) pose an imminent threat to the power, purity, or the very existence of the in-group. This makes violence and atrocities seem defensive, justified, and even virtuous, so atrocities can be perpetrated without significant resistance—and sometimes with encouragement—from in-group civilians.

It is very likely that such language drives norms toward condoning or even committing violence, and it is indisputably at least an indicator of normative change. Such changes happen gradually, but they are dramatic, since prevailing social norms all over the world strongly oppose atrocities: by definition, atrocities are transgressive of norms. After observing striking similarities in leaders' language preceding atrocities, we defined and named a category called dangerous speech: any human expression (e.g. speech, text, or images, online or offline) that can increase the likelihood that someone will commit or condone violence against members of another group.[1] An increase in dangerous speech, or in its severity, is a signal for a normative shift toward intergroup violence.

Powerful though it is, dangerous speech is not always easy to spot. It cannot be reliably identified with a list of words, since the effect—in this case, the dangerousness—of a message depends not only on its content, but also how it is communicated: by whom, to whom, and under what circumstances. What is benign in one context may be highly inflammatory

[1] We study dangerous speech in the hope of finding ways to prevent intergroup violence. For context, violence against people is not always immoral or illegitimate, in our view. It can be justified in rare cases, for example if it were the only way of preventing worse or greater violence against people. Mass violence against people simply because they belong to an identity group cannot be morally justified, however, so dangerous speech never promotes legitimate violence.

Catherine Buerger and Susan Benesch. "Lessons Dangerous Speech as an Atrocity Early Warning Indicator: Measuring Changing Conflict Dynamics." In "Evidence-Based Approaches to Preventing Genocide and Mass Atrocities." Edited by Jamie D. Wise and Kristina Hook. Special Issue, *Genocide Studies and Prevention* 18, no. 1, 84–95. https://doi.org/10.5038/1911-9933.18.1.1955.

in another, often due to historical or social factors. People who know those factors can—and should—be trained to recognize dangerous speech. As we discuss in more detail later in the paper, local expertise is essential here, as is finding ways to connect those who can identify dangerous speech with individuals and organizations who can respond effectively.

Dangerous speech is different from the term "hate speech" which, though widely used, is hard to define clearly and consistently. Speech can be both hateful and dangerous, but there are also hateful messages that do not lower social barriers to violence, and messages that can be very effective dangerous speech without being hateful. For example, an influential leader reporting that members of another community are coming to attack would likely convince the audience that a violent response was warranted. This speech is dangerous, but not hateful.

The category of dangerous speech also overlaps with that of disinformation/misinformation,[2] which has recently come into the spotlight as a global threat to democracy and peace.[3] Dangerous speech is usually false—not surprising, since it denigrates and often dehumanizes whole groups of human beings. Unfortunately, people can be quite easily persuaded of mis- and disinformation, especially in crises due to armed conflict, disease outbreaks, and natural disasters. In such times, reliable information is often scarce, and rumors blaming another group for the crisis may be particularly attractive. In circumstances such as these, mis- and disinformation can quickly become dangerous speech. In other words, they can persuade people to endorse or commit violent attacks on members of other groups.

This paper will review the literature on the connection between speech and violence and explain how dangerous speech—and responses to it—can serve as a signal of changing conflict dynamics. One of the best ways to monitor for such signals is to track dangerous speech online, including on social media, since it circulates widely there, and many platforms register and display users' reactions to content, which serves as a rough, if flawed, proxy for the popularity of messages.

We go on to make the case that embassy staff, as well as civil society practitioners, when trained to identify dangerous speech, can serve as a much-needed bridge, bringing local knowledge to government officials and NGOs who can marshal resources for effective interventions.

Speech and Violence—The Connection

Atrocities are socially abnormal by definition. So, when large-scale atrocities occur, it is generally the result of a shift in attitudes that has lowered prevailing normative barriers against such acts.[4] Individuals come to see intergroup violence as justified—or even necessary. There is general agreement among scholars that speech plays a critical role in this shift,[5] although there is no consensus about the specific causal relationship.

Some argue that there is a direct and measurable relationship between speech and violence. In his well-known study, David Yanagizawa-Drott investigated the extent to which radio station Radio Télévision Libre des Mille Collines (RTLM) inspired violence during the

---

[2] Disinformation is usually understood to mean false assertions made or spread by people who know they are false, and misinformation is falsehoods spread by people who believe them to be true.

[3] Carme Colomina et al., "The Impact of Disinformation on Democratic Processes and Human Rights in the World," (Brussels: European Parliament, 2021), accessed July 13, 2024, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf.

[4] Paul Morrow, "The Thesis of Norm Transformation in the Theory of Mass Atrocity," *Genocide Studies and Prevention* 9, no. 1 (2015), 66–82, accessed July 13, 2024, http://dx.doi.org/10.5038/1911-9933.9.1.1303; Jonathan Leader Maynard, *Ideology and Mass Killing: The Radicalized Security Politics of Genocides and Deadly Atrocities* (Oxford: Oxford University Press, 2022).

[5] Jonathan Leader Maynard and Susan Benesch, "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention," *Genocide Studies and Prevention* 9, no. 3 (2016), 70–95, accessed July 13, 2024, http://dx.doi.org/10.5038/1911-9933.9.3.1317; Rhiannon S. Neilsen, "'Toxification' as a More Precise Early Warning Sign for Genocide Than Dehumanization? An Emerging Research Agenda," *Genocide Studies and Prevention* 9, no. 1 (2015), 83–95, accessed July 13, 2024, http://dx.doi.org/10.5038/1911-9933.9.1.1277.

100-day genocidal campaign in Rwanda in 1994 in which 800,000 Tutsis were murdered.[6] Before and during this period, RTLM was known for spreading hateful and even genocidal speech against Tutsis, and it was widely and popularly described as a catalyst of the genocide. In 1997, the United Nations International Criminal Tribunal for Rwanda (ICTR) indicted three Rwandans for incitement to genocide: Ferdinand Nahimana and Jean-Bosco Barayagwiza, co-founders of RTLM, and Hassan Ngeze, the founder and editor of a pro-Hutu and violently anti-Tutsi newspaper called *Kangura*. All three were convicted in 2003 of charges including "direct and public incitement to genocide." They appealed, and in 2007 many of their convictions were reversed, but Nahimana and Ngeze were still found guilty of incitement to commit genocide.

Yanagizawa-Drott attempted to isolate the impact of RTLM's radio broadcasts by making use of data derived from Rwanda's extraordinarily hilly terrain. (The country is nicknamed "the land of a thousand hills," hence the name of the RTLM radio station, "*Mille collines*," which means one thousand hills). Positing that Rwandan villages at or near the top of hills had clear reception of RTLM radio broadcasts, while those in the valleys could not get the signal, Yanagizawa-Drott compared matched pairs of villages that were demographically similar and near to each other, except for altitude. He found a significantly higher level of genocidal violence in hilltop villages that presumably received RTLM broadcasts, using the number of people later tried for genocide in the corresponding villages as a proxy for numbers of people who participated in the genocide. He thus concludes that RTLM broadcasts inspired more killing. Maja Adena et al. used a similar mathematical test to study the impact of Nazi propaganda, finding that radio exposure led to increased support of Nazi policies, at least among those who did not "disagree with the propaganda message a priori."[7]

Several human rights reports also provide some evidence that exposure to inflammatory speech catalyzed mass violence. For example, a Human Rights Watch report detailing post-election violence in Côte d'Ivoire in 2010 and 2011 includes a speech delivered by Charles Blé Goudé, Youth Minister under then-President Koudou Laurent Gbagbo, telling Gbagbo supporters to secure their neighborhoods against "*allogènes*" or "foreigners" (other West African nationals and ethnic groups from the northern part of the country).[8] Multiple victims later said they were attacked by people who spoke of Blé Goudé's "order."

Such examples, in which evidence suggests that a particular speech act incited a specific attack, are relatively rare, since speech affects beliefs and behavior over time, gradually moving people toward condoning or committing violence against members of another group. Even in cases such as the one described above, Blé Goudé's "order" did not come across in an information vacuum. It was interpreted by people who had already been exposed to dangerous political rhetoric that preceded the election.[9]

Many scholars indeed describe a complex causal relationship between speech and violence. Charles Mironko, for example, interviewed confessed genocide perpetrators in Rwanda and found that some felt RTLM was primarily geared toward urban audiences, and many rural Rwandans described hearing about influential anti-Tutsi speeches from their friends instead of directly on the radio.[10] Darryl Li made similar findings, suggesting that although RTLM was influential, its messages were amplified by individuals through their social

---

[6] David Yanagizawa-Drott, "Propaganda and Conflict: Evidence from the Rwandan Genocide," *Quarterly Journal of Economics* 129, no. 4 (2014), 1947–1994, accessed July 13, 2024, https://doi.org/10.1093/qje/qju020.

[7] Maja Adena et al., "Radio and the Rise of the Nazis in Prewar Germany," *Quarterly Journal of Economics* 130, no. 4 (2015), 1890, accessed July 13, 2024, https://doi.org/10.1093/qje/qjv030.

[8] "Côte d'Ivoire: Crimes Against Humanity by Gbagbo Forces," *Human Rights Watch*, March 15, 2011, accessed July 13, 2024, https://www.hrw.org/news/2011/03/15/cote-divoire-crimes-against-humanity-gbagbo-forces.

[9] United Nations Human Rights Council, *Report of the High Commissioner for Human Rights on the Situation of Human Rights in Côte d'Ivoire*, February 25, 2011 (UNHRC Doc. A/HRC/16/79), accessed July 13, 2024, https://www.refworld.org/pdfid/4d8b3e162.pdf.

[10] Charles Mironko, "The Effect of RTLM's Rhetoric of Ethnic Hatred in Rural Rwanda," in *The Media and the Rwanda Genocide*, ed. by Allan Thompson (London: Pluto Press, 2007), 125–135, accessed July 13, 2024, https://doi.org/10.2307/j.ctt18fs550.15.

networks.[11] Both of these studies suggest that the presence of dangerous speech on the radio was connected to the violence, but that it reached audiences via both direct and indirect pathways. Some heard the messages on the radio, others heard them from friends, neighbors, and other influential people.

In the same vein, Scott Straus surveyed confessed perpetrators and analyzed how they had been exposed to hateful messages on RTLM, considering timing, content, and the source of messages. He concluded that the radio station engendered violence primarily by reinforcing narratives that people heard from other sources and normalizing violence.[12] The work of Straus, Mironko, and Li confirm the importance of speech in changing beliefs and motivating violent behavior, even if they do not conclude that RTLM was a primary cause of the genocide. Instead, they point to the important role played by other influential voices in spreading dangerous narratives throughout communities. Speech can be more persuasive when those in the audience hear a dangerous message multiple times from different speakers whom they trust.

Jonathan Leader Maynard has described the point at which a message "becomes something 'everybody says'," as "discursive saturation."[13] When individuals who have not yet made up their mind about something perceive that "everyone" is stating the same opinion about it, it is more likely that they will fall into agreement. In these cases, they are guided by what scholars call "descriptive" norms.[14] Human behavior is significantly shaped by social norms; that is, their beliefs about what behavior other members of their group condone.

This is why dangerous speech on social media serves as such a useful warning sign of changing conflict dynamics. Increasing abundance and/or severity of dangerous speech signals a lowering of the social barriers against violent rhetoric and, potentially, violence itself. Social media users craft posts with the intention of getting approval from the "right" users (those with whom they wish to associate themselves), signaled through one-click responses (e.g., "likes" and "hearts").[15] Research has shown that not getting this approval can threaten a feeling of belonging[16] and deter users from posting in the future.[17] Algorithms that select which content users see, such as Facebook's newsfeed algorithm, are designed to increase "engagement," or the time people spend on a platform. Therefore, they amplify posts and comments that capture attention, usually by heightening human emotions, and inspire them to react, for example with "likes" and replies. An abundance of dangerous speech online therefore signals that such speech has become acceptable with at least a segment of a population (and that social media platforms have neither removed the content nor hidden it with downranking).

Such normative shifts have been described with reference to the Overton Window, a theory of the way the acceptable range of political discourse, or policies, changes over time within groups. The theory's originator, Joseph Overton, imagined a window containing views or policies that are acceptable to influential members of the public at a particular time. As once-

---

[11] Darryl Li, "Echoes of Violence: Considerations on Radio and Genocide in Rwanda," *Journal of Genocide Research* 6, no. 1 (2004), 9–27, accessed July 13, 2024, https://doi.org/10.1080/1462352042000194683.

[12] Scott Straus, "What is the Relationship Between Hate Radio and Violence? Rethinking Rwanda's 'Radio Machete,'" *Politics & Society* 35, no. 4 (2007), 609–637, accessed July 13, 2024, https://doi.org/10.1177/0032329207308181.

[13] Jonathan Leader Maynard, "Rethinking the Role of Ideology in Mass Atrocities," *Terrorism and Political Violence* 26, no. 5, (2014), 88, accessed July 13, 2024, https://doi.org/10.1080/09546553.2013.796934.

[14] Robert B. Cialdini, "Basic Social Influence is Underestimated," *Psychological Inquiry* 16, no. 4 (2005), 158–161, accessed July 13, 2024, https://doi.org/10.1207/s15327965pli1604_03.

[15] Lauren Scissors et al., "What's in a Like? Attitudes and Behaviors Around Receiving Likes on Facebook," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (New York: Association for Computing Machinery, 2016), 1501–1510, accessed July 13, 2024, https://doi.org/10.1145/2818048.282006.

[16] Sabine Reich et al., "Zero Likes – Symbolic Interactions and Need Satisfaction Online," *Computers in Human Behavior* 80, (2018), 97–102, accessed July 13, 2024, https://doi.org/10.1016/j.chb.2017.10.043.

[17] Chandan Sarkar et al., "Predicting Length of Membership in Online Community 'everything2' Using Feedback," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion* (New York: Association for Computing Machinery, 2012), 207–210, accessed July 13, 2024, https://doi.org/10.1145/2141512.2141579.

radical positions or ideas become more acceptable, the imaginary window's frame gradually moves in their direction.[18] An increased presence of dangerous speech online could signal that language that was once considered unacceptable or radical (such as dehumanizing speech used to describe another group), is now inside the Overton Window.

When one sees the same dangerous speech narrative or harmful disinformation being shared by many different people, it signals that a context is moving in the direction of "discursive saturation."[19] It also can reveal a collective sense of "mounting panic," something that anthropologist Veena Das identified in her study of the massacre of Sikhs that followed the murder of India's former Prime Minister Indira Gandhi in 1984. Rumors spread and become more powerful as people begin to hear them from many sources. Das notes that this repetition and reinforcement of frightening messages increases feelings of panic: "My *fear* of the other is transformed into the notion that the other is *fearsome*."[20] Shifts in beliefs like this are strongly connected to social and historical context, which have already laid the groundwork for feelings of difference, and potentially fear, between groups.

What Constitutes Dangerous Speech and How Can it Best be Monitored?

Messages (verbal or nonverbal) that can convince an audience to endorse or even commit mass violence, which we call "dangerous speech," are highly context dependent. The way audience members interpret messages, and how convincing they find them, depends on the speaker's authority and charisma, how the message is communicated, and numerous other social and historical factors. At the Dangerous Speech Project, a research team, we use a systematic method for analyzing speech in context—a five-part framework which includes the message itself, the audience, the historical and social context of the message, the speaker, and the medium by which a speaker delivers a message. Analyzing each of these five elements is not only essential for identifying how dangerous speech operates, but also useful for designing interventions to diminish the dangerousness of that speech.

Because the social, historical, and cultural context in which speech was made or disseminated is essential for understanding its possible impact, those with extensive knowledge of the relevant language, culture, and social conditions are best able to identify dangerous speech. In fact, dangerous speech is often expressed in language so coded, and so particular to an in-group or location, that someone without localized knowledge would miss it entirely. For example, MTN is the name of a South African telecommunications company well known in sub-Saharan Africa, where it sells mobile phones and service in over 20 countries. In nearly all of those, the company's name and its slogan "everywhere you go" (meant to refer to ubiquitous cell phone reception) have always been innocuous.

In South Sudan however, during the early days of the 2016 conflict there, the slogan was given a new meaning—that members of the Dinka tribe were encroaching on the lands of other ethnic groups—and the letters MTN became a dangerous slur, used to spread fear about the Dinka. Used on social media and also in person, for example to identify people to be pulled off a bus and murdered, the term MTN stirred "fear by exaggerating the number and location of Dinkas within South Sudan, suggesting an increasing presence and pervasive (negative) influence throughout the country, specifically in competition for land, access to water, government services, and jobs."[21] Someone who was trained in dangerous speech ideas would have recognized MTN as an example; others might well have missed it.

---

[18] Mackinac Center for Public Policy, "A Brief Explanation of the Overton Window," *Mackinac Center for Public Policy* (webpage), accessed July 13, 2024, https://www.mackinac.org/overtonwindow#top.

[19] Leader Maynard, *Rethinking the Role of Ideology in Mass Atrocities*.

[20] Veena Das, "Specificities: Official Narratives, Rumour, and the Social Production of Hate," *Social Identities* 4, no.1 (1998), 125, accessed July 13, 2024, https://doi.org/10.1080/13504639851915.

[21] PeaceTech Lab, *Social Media and Conflict in South Sudan: A Lexicon of Hate Speech Terms* (Washington, DC: PeaceTech Lab, December 2016), 7, accessed July 13, 2024, https://static1.squarespace.com/static/54257189e4b0ac0d5fca1566/t/5b0f0c321ae6cf107119712e/1563308852571/South+Sudan+Lexicon+-+PeaceTech+Lab.

For dangerous speech to serve as an early warning signal of conflict, at least two things must happen: individuals must be trained to recognize it, and those with the power and resources to intervene must be made aware of its presence. Embassy staff are one group ideally positioned to document and report on dangerous speech, as they could bridge the gap between local understanding and state-level resources. This is especially true for staff at American embassies as they already receive wide-ranging training, and the U.S. State Department has committed to training its staff specifically in atrocity prevention, recognizing the important role that "on the ground" staff can play in early warning.[22] The United States has diplomatic missions in 169 countries and "interest sections" or de facto missions under another name, in three more (Afghanistan, Iran, and Syria). Each of these missions employs both American citizens and local staff.

A challenge, however, is that staff at embassies and consulates are often over stretched, and workload capacity is even more limited for those in countries already at risk of conflict and atrocities. Because of this, staff at international organizations and civil society groups (such as Human Rights Watch or Mercy Corps) that already have strong ties with policy makers in governments or at intergovernmental organizations, should also be trained to identify dangerous speech. Their preexisting ties, built on established and trusted relationships, will provide clear channels for communication, allowing for better responses in crisis. Tech companies already maintain lists of "trusted flaggers," unpaid volunteers who understand which content violates the companies' terms of service, and who have access to a streamlined process for reporting or "flagging" it to them. Several years ago the European Commission endorsed this model as a response to hateful content—that it is illegal under the national laws of member states—in 2016 passing a "Code of conduct on countering illegal hate speech online," which, among other things, enjoined tech companies to enlarge their networks of trusted flaggers and train them to identify and counter hateful rhetoric and prejudice.[23] Facebook reported that it brought on 66 new EU NGOs as trusted flaggers and Twitter added 40 new NGOs in 21 EU countries. The companies also trained the groups and worked with them to design campaigns to promote tolerance and pluralism.[24]

To use dangerous speech as an early warning signal of atrocities, local NGO staff must learn to recognize the "hallmarks," or recurring rhetorical patterns, in dangerous speech. As Kjell Anderson notes, the committing of atrocities, "does not require true believers; acquiescence and rationalization of wrongful acts are enough. This process is facilitated by the individual's need to frame their action in such a way that it remains consistent with their notions of moral selfhood."[25] Dangerous speech enables such framing by asserting a threat, suggesting that members of another group (or disloyal members of the in-group) pose a danger so grave that violence against them is justified. Such speech makes people perceive an existential threat: they believe they must commit (or permit) violence in order to protect their people. In Anderson's words, when one's support of violence is justified, the sense of moral deviation is "neutralized." Hatred need not be part of this process. One can assert that another group is planning to attack one's own group without expressing or fomenting hatred, yet that message might easily convince people to condone or commit violence, to fend off ostensible attacks. Such frightening messages may spread even more widely and quickly than purely

---

[22] Institute for Genocide & Mass Atrocity Prevention, "Jeff Sizemore: Atrocity Prevention in Practice: All Hands On Deck," *YouTube* (video), November 9, 2022, accessed July 13, 2024, https://www.youtube.com/watch?v=AaAOquKgplU.

[23] European Commission, "Code of Conduct on Countering Illegal Hate Speech Online," June 30, 2016, accessed July13, 2024, http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf.

[24] Věra Jourová, "How the Code of Conduct Helped Countering Illegal Hate Speech Online," *European Commission*, 2019, accessed September 27, 2024, https://commission.europa.eu/document/17662c1b-bcbc-42fe-b5a6-9d0ca970874d_en.

[25] Kjell Anderson, "'Who Was I to Stop the Killing?' Moral Neutralization Among Rwandan Genocide Perpetrators," *Journal of Perpetrator Research* 1, no. 1 (2017), 39–63, accessed July 13, 2024, https://doi.org/10.21039/jpr.v1i1.49.

hateful ones, since many people share them without malevolent intentions, or even the desire to incite violence. They feel genuine fear.

By analyzing examples of public speech that preceded incidents of intergroup violence in many different contexts, we have identified five hallmarks of dangerous speech: dehumanization, accusation in a mirror, threats to group integrity or purity, assertions of attacks against women and children, and questioning in-group loyalty.

Dehumanization is the most familiar among them, in common parlance and literature on violence prevention.[26] Dehumanization may come in the form of describing or depicting humans as nonhuman living things (such as insects or animals), diseases, or even objects. By describing other groups of people as something other than human, or less than human, speakers can prepare audiences to condone or commit violence by making their targets' death and suffering seem less significant, or even by making it seem useful or necessary.

The most powerful way to foment intergroup conflict is to frame violence as the only way to protect an in-group against greater harm, even annihilation. To that end, dangerous speech often includes a special kind of justification of violence that has become known as *accusation in a mirror*. The concept comes from a manual for propaganda and recruitment found in Butare, Rwanda, after the 1994 genocide, written by an unnamed author or authors. The document advises attributing to one's enemies the very acts of violence the speaker hopes to commit against them. When "the party which is using terror will accuse the enemy of using terror" this will so frighten "honest people" that they will agree that violence is necessary "for legitimate [self-] defense."[27] Dehumanization can make violence against other human beings seem acceptable. Accusation in a mirror can be even more powerful since it makes such violence seem necessary.

Dangerous speech may also claim that members of another group (by their presence or values) contaminate the purity or challenge the integrity of the in-group. Members of the out-group may be described as "stains," or "rotten," or a message may claim that the values of the out-group may destroy or poison the in-group's culture or religious values. A classic example of this appeared in a 1931 German cartoon from Julius Streicher's Nazi newspaper *Der Stürmer* that shows an apple sliced open with a knife marked with a swastika. Inside the apple is a worm that has a stereotypically Jewish face. The caption reads "*Wo etwas faul ist, ist der Jude die Ursache*" ("Where something is rotten, the Jew is the cause").[28]

Related to the previous hallmark is the suggestion that women or children of the in-group have been or will be harmed by members of an out-group. Throughout history, accusations of threats against women or children have been used repeatedly against minority groups, many times leading to violence. In the United States, false claims of attacks against white women often led to lynchings and other violence against black people, especially in parts of the country where Africans had been enslaved. One of the most well-known examples of this hallmark as it pertains to children is the false allegation known as "blood libel," that Jews murder Christian children to use their blood for religious rituals. Since women and children are seen as vulnerable, precious, and needing protection, in virtually all human societies, it is

---

[26] Beyond Conflict, *Decoding Dehumanization: Policy Brief for Policymakers and Practitioners* (Boston: Beyond Conflict, 2019), accessed July 15, 2024, https://beyondconflictint.org/wp-content/uploads/2020/06/Decoding-Dehumanization-Policy-Brief-2019.pdf; Roger Giner-Sorolla et al., "Dehumanization, Demonization, and Morality Shifting: Paths to Moral Certainty in Extremist Violence" in *Extremism and the Psychology of Uncertainty*, ed. Michael A. Hogg and Danielle L. Blaylock (Hoboken: Wiley-Blackwell, 2012), 165–182; Nick Haslam, "Dehumanization: An Integrative Review," *Personality and Social Psychology Review* 10, no. 3 (2006), 252–264, accessed July 13, 2024, https://doi.org/10.1207/s15327957pspr1003_4; David Livingstone Smith, "Paradoxes of Dehumanization," *Social Theory and Practice* 42, no. 4 (2016), 416-443, accessed July 13, 2024, https://doi.org/10.5840/soctheorpract201642222.

[27] Anonymous, cited in Alison Des Forges, "Leave None to Tell the Story" (New York: Human Rights Watch, March 1999). Accessed July 13, 2024, https://www.hrw.org/reports/1999/rwanda/.

[28] "Caricatures from *Der Stürmer*: 1927–1932," *German Propaganda Archive* (webpage), accessed October 9, 2018, http://research.calvin.edu/german-propaganda-archive/sturm28.htm.

honorable to defend them, even with violence. Therefore, claiming that women or children are in danger, or already being harmed, is a quick and effective way to lower psychosocial barriers against violence.

Though dangerous speech usually describes members of the out-group, some is focused on members of the in-group, describing them as insufficiently loyal, or even traitorous, for being sympathetic to an out-group. Over time, speech such as this may redefine the parameters of the group, separating those with more moderate beliefs and sympathies and marking them as "others." Speech containing this hallmark not only poses a danger to those who have exhibited moderate beliefs or behavior in the past, but it silences future dissent, making it more likely that extreme ideology will take hold.

A description of these hallmarks and an overview of how to use the dangerous speech five-part framework to identify and analyze dangerous speech should be added to the existing atrocity prevention training received by staff at embassies and international peacebuilding NGOS. Then, if discourse norms begin to shift online in the country toward more dangerous rhetoric, staff would be more prepared to quickly spot it and elevate that knowledge to others who may be working on conflict and stabilization efforts.

Social media and the internet have immeasurably changed the way people reach each other to communicate ideas: messages spread rapidly and broadly, and threats, real or fictitious, can be easily distorted to make them more frightening. Well-meaning people may share these messages to warn their loved ones of a threat that they believe to be real (whether this is an accurate appraisal or not). Ideas and narratives once confined to the fringes of popular discourse —including extremist ideas—are now widely accessible online. People can also communicate anonymously online. On social media platforms like Twitter or Reddit, or messaging platforms like WhatsApp or Discord, they can spread ideas that they might not dare to express offline, in circumstances in which their identities are evident. All these factors can increase the harmful impacts of dangerous speech, including exacerbating already tense conflict dynamics.

In her article, "Atrocity Prevention in the New Media Landscape," Rebecca Hamilton argues that reporting on evolving atrocity risks and dynamics is key to early warning—and it is very hard to do considering the dominance of social media and the declining number of journalists based in overseas bureaus. "Living in a country for a prolonged period enables journalists to notice things that those who 'parachute in,' even for weeks at a time, are unlikely to identify," she notes.[29] Moreover, social media "only exacerbates a longstanding problem with early warning information—its weak signal/noise ratio. At any point, many situations display risk factors of atrocity, making it hard for policymakers to prioritize." The presence of so-called deep fakes also makes it harder to detect authentic signals of emerging violence.

But social media also provides new windows into shifting conflict dynamics. First, focusing on the relative changes in the abundance and severity of dangerous speech online in a context removes the need to rely on (and thus substantiate) any one report. When those familiar with the context begin to see a new form of dangerous speech spreading, an influential speaker endorsing dangerous speech, or a proliferation of such speech coalescing around one narrative or target group, it is an indicator that conflict dynamics may be changing, requiring closer attention from policy makers.

Second, monitoring the spread of dangerous rhetoric online allows someone to see how the speech is being received. In her 2021 essay, "The Insidious Creep of Violent Rhetoric," Susan Benesch argues that speech regulators at social media companies have focused too much on the (unknowable) intent of those who post inflammatory content.[30] For violence prevention, the intent of a speaker is much less important than their effect on other people: how the content was understood by the audience. On social media, this is often easy to see—reactions are visible as

---

[29] Rebecca Hamilton, "Atrocity Prevention in the New Media Landscape," *American Journal of International Law* 113 (2019), 265, accessed July 13, 2024, https://doi.org/10.1017/aju.2019.45.

[30] Susan Benesch, "The Insidious Creep of Violent Rhetoric," *NOĒMA*, March 4, 2021, accessed July 13, 2024, https://www.noemamag.com/the-insidious-creep-of-violent-rhetoric/.

comments. Others may agree with the content or even expound on it in a way that makes it more explicit and more dangerous, or they may denounce it as false or unacceptable. This information can reveal a lot about how close a community is to accepting violence.

To illustrate this point, Benesch uses the example of a tweet posted by former U.S. President Donald Trump on December 19, 2021: "Big protest in D.C. on January 6th. Be there, will be wild!" Although this was not an explicit call for violence, many of Trump's followers interpreted it as exactly that. On the online forum "TheDonald," for example, users immediately reacted with statements like: "Well, shit. We've got marching orders, bois" and "We have been waiting for Trump to say the word." Another replied, "Then bring the guns we shall," and they went on to describe plans for attacking the Capitol and arresting or even killing legislators. As Benesch notes, "It was abundantly clear, more than two weeks before they went to Washington D.C.—that they had been incited to violence."

A third advantage of using dangerous speech online as an early warning signal for conflict is that even as conflicts intensify, and accessing communities in-person becomes more difficult, online speech can still be viewed. Assessments of whether speech is dangerous must be made by those familiar with local history and social systems, but they need not necessarily be physically present in the place from which the speech is emerging. This allows for continuous access to valuable information despite shifting conflict dynamics.

There are caveats, however. First, access to social media is not evenly distributed through most societies. A focus on what is being said online inevitably means getting only a partial view. As scholars such as Li and Mironko noted in the case of Rwanda, the same ideas that are circulating through media (such as the radio, television, or Facebook), may get amplified in communities through in-person social networks.[31] Again, this is a place where locally-based staff can assist as they are more likely to know whether the speech they are seeing online is representative of what is being said offline.

The second caveat is that internet shutdowns are an increasingly common reaction from authoritarian governments to quell protest and silence dissent.[32] Internet shutdowns may be partial (where a government blocks access to a specific set of websites) or total (where access the internet or all telecommunication services is cut). In 2021, Access Now's #KeepItOn campaign documented 182 internet shutdowns in 34 countries and, in many cases, governments instituted these in times of political instability under the guise of protecting "national security." If public dissent and conflicts escalate to the point where authoritarian governments remove access to social media, then tracking dangerous speech becomes more difficult. In these cases, it would be important to pay attention to what is being said in diaspora communities, as they often remain closely involved in the politics of their home countries, even after moving abroad.[33]

Conclusion: The Future of Using Dangerous Speech as an Early Warning Signal

Learning to recognize dangerous speech can be a useful tool in the quest to prevent atrocities. The spread of dangerous speech online allows atrocity prevention practitioners access to that content. When analyzed by people who can understand it in context, it provides a window into normative changes in places at risk for atrocities—even early in the process of societal degradation that permits atrocities. An increase in dangerous speech can be detected well before the risk of atrocities is easily observable by other means. This is vital for preventing atrocities,

---

[31] Li, *Echos of Violence*; Mironko, *Effect of RTLM's Rhetoric*.

[32] Marianne Díaz Hernández et al., "#KeepItOn Update: Who is Shutting Down the Internet in 2021?" *Access Now* (webpage), last updated January 13, 2023, accessed July 13, 2024, https://www.accessnow.org/who-is-shutting-down-the-internet-in-2021/.

[33] William J. Lahneman, "Impact of Diaspora Communities on National and Global Politics: Report on Survey of the Literature," *Strategic Assessments Group: Impact of Diaspora Communities on National and Global Politics* (College Park: University of Maryland Center for International and Security Studies, July 2005), accessed August 20, 2024, https://commons.erau.edu/db-security-studies/3; Sarah Wayland, "Ethnonationalist Networks and Transnational Opportunities: The Sri Lankan Tamil Diaspora," *Review of International Studies* 30, no. 3 (2004), 405–426, accessed July 13, 2024, https://doi.org/10.1017/S0260210504006138.

since once the risk is plain, it is generally too late to intervene effectively. When credible, early indications of risk are supplied to people who have the will and power to intervene, they allow for more tailored, sophisticated interventions to prevent violence.

Going forward, practitioners must work toward creating tools to detect dangerous speech at scale. An efficient, effective monitoring system would use automated tools called classifiers to find content that may be dangerous and submit it for review by humans with knowledge of the local setting and of structural factors that influence how groups in a society relate to one another. Review by humans is vital also to test whether classifiers are reasonably successful at surfacing dangerous content.

Building classifiers for dangerous speech will be difficult, but recent improvements in large language modeling have made the process easier and faster, likely without significantly sacrificing quality. If implemented in several countries, such a two-part system using classifiers and knowledgeable people could be used not only to design better interventions against violence, but it would also permit invaluable and pathbreaking comparative study. Until that is possible, however, staff at organizations with large international networks and individuals working at diplomatic missions around the world can provide essential knowledge by being taught to recognize dangerous speech and the normative shifts that are precursors to mass violence.

Bibliography

Access Now. "Internet Shutdowns and Elections Handbook." Last updated April 2021. Accessed July 13, 2024. https://www.accessnow.org/internet-shutdowns-and-elections-handbook/.

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. "Radio and the Rise of the Nazis in Prewar Germany." *Quarterly Journal of Economics* 130, no. 4 (2015), 1885–1939. Accessed July 13, 2024. https://doi.org/10.1093/qje/qjv030.

Anderson, Kjell. "'Who Was I to Stop the Killing?' Moral Neutralization Among Rwandan Genocide Perpetrators." *Journal of Perpetrator Research* 1, no. 1 (2017), 39–63. Accessed July 13, 2024. https://doi.org/10.21039/jpr.v1i1.49.

Benesch, Susan. "The Insidious Creep of Violent Rhetoric." *NOĒMA*, March 4, 2021. Accessed July 13, 2024. https://www.noemamag.com/the-insidious-creep-of-violent-rhetoric/.

Cialdini, Robert B. "Basic Social Influence is Underestimated." *Psychological Inquiry* 16, no. 4 (2005), 158–161. Accessed July 13, 2024. https://doi.org/10.1207/s15327965pli1604_03.

Colomina, Carme, Héctor Sánchez Margalef, Richard Youngs, and Kate Jones. "The Impact of Disinformation on Democratic Processes and Human Rights in the World." Brussels: European Parliament. April 2021. Accessed July 13, 2024. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf.

Dangerous Speech Project. "Dangerous Speech: A Practical Guide." Last updated April 19, 2021. Accessed July 13, 2024. https://dangerousspeech.org/guide/.

Das, Veena. "Specificities: Official Narratives, Rumour, and the Social Production of Hate." *Social Identities* 4, no. 1 (1998), 109–130. Accessed July 13, 2024. https://doi.org/10.1080/13504639851915.

Datareportal. "Global Social Media Statistics." *Datareportal* (webpage). Accessed July 13, 2024. https://datareportal.com/social-media-users.

Des Forges, Alison. "Leave None to Tell the Story." New York: Human Rights Watch. March 1999. Accessed July 13, 2024. https://www.hrw.org/reports/1999/rwanda/.

Díaz Hernández, Marianne, Rafael Nunes, Felicia Anthonio, and Sage Cheng. "#KeepItOn Update: Who is Shutting Down the Internet in 2021?" *Access Now* (webpage), last updated January 13, 2023. Accessed July 13, 2024. https://www.accessnow.org/who-is-shutting-down-the-internet-in-2021/.

European Commission. "Code of Conduct on Countering Illegal Hate Speech Online." June 30, 2016. Accessed July 13, 2024. http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf.

Facing History & Ourselves. "The Power of a Lie: The History of the Blood Libel." *Facing History & Ourselves* (video), last updated February 11, 2014. Accessed July 13, 2024. https://www.facinghistory.org/resource-library/power-lie-history-blood-libel.

Hamilton, Rebecca. "Atrocity Prevention in the New Media Landscape." *American Journal of International Law* 113, (2019), 262–266. Accessed July 13, 2024. https://doi.org/10.1017/aju.2019.45.

Haslam, Nick. "Dehumanization: An Integrative Review." *Personality and Social Psychology Review* 10, no. 3 (2006), 252–264. Accessed July 13, 2024. https://doi.org/10.1207/s15327957pspr1003_4.

Human Rights Watch. "Côte d'Ivoire: Crimes Against Humanity by Gbagbo Forces." March 15, 2011. Accessed July 13, 2024. https://www.hrw.org/news/2011/03/15/cote-divoire-crimes-against-humanity-gbagbo-forces.

Institute for Genocide & Mass Atrocity Prevention. "Jeff Sizemore—Atrocity Prevention in Practice: All Hands On Deck." *YouTube* (video), November 29, 2022. Accessed July 13, 2024. https://www.youtube.com/watch?v=AaAOquKgplU.

Jourová, Věra. "How the Code of Conduct Helped Countering Illegal Hate Speech Online." European Commission, 2019. Accessed September 27, 2024. https://commission.europa.eu/document/17662c1b-bcbc-42fe-b5a6-9d0ca970874d_en.

Kühl, Stefan. *Ordinary Organisations: Why Normal Men Carried Out the Holocaust*. Cambridge, UK: Polity, 2016.

Lahneman, William J. "Impact of Diaspora Communities on National and Global Politics: Report on Survey of the Literature." *Strategic Assessments Group: Impact of Diaspora Communities on National and Global Politics*. College Park: University of Maryland Center for International and Security Studies. July 2005. Accessed August 20, 2024. https://commons.erau.edu/db-security-studies/3.

Leader Maynard, Jonathan. *Ideology and Mass Killing: The Radicalized Security Politics of Genocides and Deadly Atrocities*. Oxford: Oxford University Press, 2022.

----------. "Rethinking the Role of Ideology in Mass Atrocities." *Terrorism and Political Violence* 26, no. 5 (2014), 821–841. Accessed July 13, 2024. https://doi.org/10.1080/09546553.2013.796934.

Leader Maynard, Jonathan, and Susan Benesch. "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention." *Genocide Studies and Prevention* 9, no. 3 (2016), 70–95. Accessed July 13, 2024. http://dx.doi.org/10.5038/1911-9933.9.3.1317.

Li, Darryl. "Echoes of Violence: Considerations on Radio and Genocide in Rwanda." *Journal of Genocide Research* 6, no. 1 (2004), 9–27. Accessed July 13, 2024. https://doi.org/10.1080/1462352042000194683.

Mackinac Center for Public Policy. "The Overton Window." *Mackinac Center for Public Policy* (webpage). Accessed July 13, 2024. https://www.mackinac.org/overtonwindow#top.

Masnick, Mike. "Dubious Studies and Easy Headlines: No, a New Report Does Not Clearly Show Facebook Leads to Hate Crimes." *Techdirt*, August 23, 2018. Accessed July 13, 2024. https://www.techdirt.com/2018/08/23/dubious-studies-easy-headlines-no-new-report-does-not-clearly-show-facebook-leads-to-hate-crimes/.

Mironko, Charles. "The Effect of RTLM's Rhetoric of Ethnic Hatred in Rural Rwanda." In *The Media and the Rwanda Genocide*, edited by Allan Thompson, 125–135. London: Pluto Press, 2007. Accessed July 13, 2024. https://doi.org/10.2307/j.ctt18fs550.15.

Morrow, Paul. "The Thesis of Norm Transformation in the Theory of Mass Atrocity." *Genocide Studies and Prevention* 9, no. 1 (2015), 66–82. Accessed July 13, 2024. http://dx.doi.org/10.5038/1911-9933.9.1.1303.

Neilsen, Rhiannon S. "'Toxification' as a More Precise Early Warning Sign for Genocide Than Dehumanization? An Emerging Research Agenda." *Genocide Studies and Prevention* 9, no. 1 (2015), 83–95. Accessed July 13, 2024. http://dx.doi.org/10.5038/1911-9933.9.1.1277.

PeaceTech Lab. *Social Media and Conflict in South Sudan: A Lexicon of Hate Speech Terms*. Washington, DC: PeaceTech Lab, December 2016. Accessed July 13, 2024. https://static1.squarespace.com/static/54257189e4b0ac0d5fca1566/t/5b0f0c321ae6cf107119712e/1563308852571/South+Sudan+Lexicon+-+PeaceTech+Lab.

Reich, Sabine, Frank M. Schneider, and Leonie Heling. "Zero Likes: Symbolic Interactions and Need Satisfaction Online." *Computers in Human Behavior* 80, (2018), 97–102. Accessed July 13, 2024. https://doi.org/10.1016/j.chb.2017.10.043.

Relia, Kunal, Zhengyi Li, Stephanie H. Cook, and Rumi Chunara. "Race, Ethnicity and National Origin-Based Discrimination in Social Media and Hate Crimes Across 100 U.S. Cities." *Proceedings of the International AAAI Conference on Web and Social Media* 13, no. 1 (2019), 417–427. Accessed July 13, 2024. https://doi.org/10.1609/icwsm.v13i01.3354.

Sarkar, Chandan, Donghee Yvette Wohn, and Cliff Lampe. "Predicting Length of Membership in Online Community 'everything2' Using Feedback." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, 207–210. New York: Association for Computing Machinery, 2012. Accessed July 13, 2024. https://doi.org/10.1145/2141512.2141579.

Scissors, Lauren, Moira Burke, and Steven Wengrovitz. "What's in a Like?: Attitudes and Behaviors Around Receiving Likes on Facebook." In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1501–1510. New York: Association for Computing Machinery, 2016. Accessed July 13, 2024. https://doi.org/10.1145/2818048.282006.

Smith, David Livingstone. "Paradoxes of Dehumanization." *Social Theory and Practice* 42, no. 2 (2016), 416–443. Accessed July 13, 2024. https://doi.org/10.5840/soctheorpract201642222.

Straus, Scott. "How Many Perpetrators Were There in the Rwandan Genocide? An Estimate." *Journal of Genocide Research* 6, no. 1 (2004), 85–98. Accessed July 13, 2024. https://doi.org/10.1080/1462352042000194728.

----------. "What is the Relationship Between Hate Radio and Violence? Rethinking Rwanda's 'Radio Machete.'" *Politics & Society* 35, no. 4 (2007), 609–637. Accessed July 13, 2024. https://doi.org/10.1177/0032329207308181.

United Nations Human Rights Council. *Report of the High Commissioner for Human Rights on the Situation of Human Rights in Côte d'Ivoire*. February 25, 2011. UNHRC Doc. A/HRC/16/79. Accessed July 13, 2024. https://www.refworld.org/pdfid/4d8b3e162.pdf.

Wayland, Sarah. "Ethnonationalist Networks and Transnational Opportunities: The Sri Lankan Tamil Diaspora." *Review of International Studies* 30, no. 3 (2004). 405–426. Accessed July 13, 2024. https://doi.org/10.1017/S0260210504006138.

Williams, Matthew L., Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *British Journal of Criminology* 60, no. 1 (2020), 93–117. Accessed July 13, 2024. https://doi.org/10.1093/bjc/azz049.

Yanagizawa-Drott, David. "Propaganda and Conflict: Evidence from the Rwandan Genocide." *Quarterly Journal of Economics* 129, no. 4 (2014), 1947–1994. Accessed July 13, 2024. https://doi.org/10.1093/qje/qju020.

Zeitzoff, Thomas. "How Social Media is Changing Conflict." *Journal of Conflict Resolution* 61, no. 9 (2017), 1970–1991. Accessed July 13, 2024. https://journals.sagepub.com/doi/abs/10.1177/0022002717721392.