COUNTERING DANGEROUS SPEECH:
NEW IDEAS FOR GENOCIDE PREVENTION

Working Paper

Susan Benesch
Director, Dangerous Speech Project
Faculty Associate, Berkman Center for Internet and Society

United States Holocaust Memorial Museum

Table of Contents

Acknowledgments

This paper was produced thanks to the United States Holocaust Memorial Museum's Edith Everett Fellowship in Genocide Prevention. I am very grateful.

INTRODUCTION

Inflammatory hate speech catalyzes mass killings including genocide, according to scholars, survivors and, notably, some former perpetrators.[1] By teaching people to view other human beings as less than human, and as mortal threats, thought leaders can make atrocities seem acceptable – and even necessary, as a form of collective self-defense. Such speech famously preceded the Holocaust,[2] the 1994 genocide in Rwanda,[3] and other intergroup mass killings,[4] and unfortunately it is still rife in many countries at risk of collective violence, such as Nigeria, Myanmar, Egypt, and Greece.

In fact this speech may be proliferating. It can be disseminated further and faster, thanks to the Internet and other digital communication including SMS messaging. Hateful and divisive speech is also a feature of two familiar contemporary scenarios: 1) tension between immigrants and majority populations, as in most of Western Europe, notably Greece, France, Italy, and the Netherlands 2) ethnic or religious leaders jockeying for power by inciting their followers against one another, especially after the fall of repressive central governments. Such incitement has happened and is still underway in

---

[1] Ervin Staub, *Overcoming Evil; Genocide, Violent Conflict, and Terrorism*, Oxford University Press, 2011; Susan Benesch, Vile Crime or Inalienable Right: Defining Incitement to Genocide, *Virginia Journal of International Law*, *48*(3), 2008; David Livingstone Smith, *Less Than Evil*, St. Martin's Griffin, 2012. Some perpetrators have indicated that inflammatory speech motivated them; others identified themselves as perpetrators because of the effect that their speech evidently had on others who went on to commit mass murder: for example, the International Criminal Tribunal for Rwanda (ICTR) has accepted several guilty pleas for incitement to genocide.

[2] Jeffrey Herf, *The Jewish Enemy; Nazi Propaganda During World War II and the Holocaust*, Harvard University Press, 2006; Victor Klemperer, *The Language of the Third Reich: LTI – Lingua Tertii Imperii: A Philologist's Notebook,* Athlone Press, 2000; *The Trial of German Major War Criminals,: Proceedings of the International Military Tribunal Sitting at Nuremberg, Germany*, Part 21, Aug. 9 1946 to Aug. 21 1946, volume 22, p. 501.

[3] Alison Des Forges, *Leave None to Tell the Story: Genocide in Rwanda*, Human Rights Watch, 1999)., p. 66; William A. Schabas, Hate Speech in Rwanda: The Road to Genocide, *McGill Law Journal 46*, 2001, 141-144.

[4] Examples are rife. *See, e.g.*, Ashutosh Varshney, *Ethnic Conflict; Hindus and Muslims in India*, Yale University Press, 2002; Daniel Chirot and Clark McCauley, *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*, Princeton University Press, 2006, at 191 and generally; Suketu Mehta, *Maximum City; Bombay Lost and Found*, Random House, 2004, especially 39-130 "Powertoni."

the former Yugoslavia, Iraq, and Egypt, among other countries. In Myanmar, for example, influential Buddhist monks are teaching their followers to despise and fear Muslims (especially but not only members of the Rohingya group) and to regard them as animals who pose a threat to the survival of Myanmar as a Buddhist nation. Not surprisingly, such language has already been closely followed by ethnic cleansing and massacres.[5]

Alarming though this phenomenon is, inflammatory speech presents opportunities for preventing mass violence, since it commonly precedes such violence.[6] At a minimum, it can serve as a new early warning indicator. Also, early evidence suggests that violence might be forestalled or at least diminished by limiting inflammatory hate speech or, without limiting it, blunting its impact.

Most policies to counter inflammatory speech are punitive or censorious, such as prosecuting, imprisoning, or even killing inflammatory speakers,[7] bombing a television station,[8] jamming a radio signal,[9] or blocking access to SMS or the Internet. These techniques often fail at suppressing hateful speech however: a bombed television station resumes broadcasting within an hour, or an extremist cleric's speeches continue to proliferate on the Internet long after he has been killed. Moreover, these methods may curb freedom of

---

[5] Human Rights Watch, *"All You Can Do is Pray": Crimes Against Humanity and Ethnic Cleansing of Rohingya Muslims in Burma's Arakan State*, April 22, 2013. hrw.org/reports/2013/04/22/all-you-can-do-pray-0

[6] For an expanded account of this, see Benesch, supra note 1.

[7] For example, Anwar al-Awlaki, an American Islamic militant known for his influential sermons, was killed by an American drone in Yemen in September 2011, in part because of his inflammatory speech.

[8] In April 1999, for example, NATO bombed the headquarters of Radio Television of Serbia (RTS), because it "was making an important contribution to the propaganda war which orchestrated the campaign against the population of Kosovo," according to NATO headquarters.

[9] In 1994, Western forces considered jamming the signal of Rwanda's famous Radio Television Libre des Milles Collines (RTLM) because of its highly inflammatory broadcasts, but did not do so. *See* Jamie Frederic Metzl, Rwandan Genocide and the International Law of Radio Jamming, *The American Journal of International Law 91*(4), 628-651. jstor.org/stable/2998097

expression, which must be protected, not only as a fundamental human right but also because denying it can increase the risk of mass violence, by closing off nonviolent avenues for the resolution of grievances.

Other methods to diminish hateful inflammatory speech – or reduce its impact – without infringing on freedom of expression are emerging. Activists, journalists, clergy, lawyers, and other have begun experimenting with such methods in a variety of countries. Technology plays a role in many of these efforts: just as new communications technologies are being used to amplify inflammatory hate speech, they can also be marshaled to prevent and counter it. New technologies are also being employed to detect where hate speech may signal an increased risk of mass violence.

Hateful speech can cause diverse forms of harm, of which the most familiar is the pain it causes *directly* to members of the group (usually a minority) that it purports to describe. They hear themselves described as vermin, for example, and are terrified. They can also be harmed indirectly (but no less viciously) when another audience hears the same speech and becomes more likely to hate them, discriminate against them, or condone or even participate in violence against them.

Genocide prevention efforts should focus on this indirect – but powerful – harm caused by speech, and on speech that has a special capacity to catalyze mass violence, which I call Dangerous Speech. In order to measure, counter, or diminish that speech, it must be reliably identified, i.e. distinguished from the much larger (and variously-defined)[10] category of hate speech.

[10] Hate speech is defined differently in bodies of law and in common parlance. In general it refers to speech that denigrates a person or people based on their membership in a group, usually an immutable group defined by race, ethnicity, sexual orientation, or disability, for example, and sometimes also religion or political affiliation or views.

Inflammatory speech preceding outbreaks of mass violence exhibits certain rhetorical hallmarks, even across historical periods and in diverse languages and cultures. Drawing on the work of other scholars and my own research, I have described these hallmarks and identified five contextual factors with which to estimate the capacity of speech to inspire mass violence. This work may be useful in developing and testing new preventive strategies for responding to inflammatory speech, especially in societies at risk of mass violence.

This paper outlines those strategies, and describes how they are being employed in the field. They are multiplying amid growing interest in atrocity prevention, so it is an ideal moment to examine them as a group, evaluate their effectiveness, and plan new experiments.

THE DANGEROUS SPEECH FRAMEWORK

Genocide and other forms of mass violence occur neither spontaneously nor abruptly. They follow a process of social conditioning to build up hatred and fear until those emotions become reflexive, and to place other human beings outside the "universe of moral obligation."[11] As the genocide scholar Helen Fein has explained, "[t]he conscience is then limited to one's own kind, members of one's class, excluding the other class from the universe of obligation-the range of persons and groups toward whom basic rules or "oughts" are binding."[12]

Speech is an essential tool for this conditioning, of course, as it is for any collective human effort. Periods preceding genocide, massacres, or ethnic cleansing typically see inflammatory public speech from an array of influential sources – politicians and comedians, athletes and bartenders. All those utterances are as different as snowflakes, and while it would be useful to find a way to classify them, especially in

[11] Helen Fein, *Imperial Crime and Punishment; The Massacre at Jallianwalla Bagh and British Judgment, 1919-1920*, University Press of Hawaii, 1977.
[12] Id, at 28.

terms of their capacity to inspire violence, such a task would be impossible. Even a trivial human action is the consequence of many factors, in proportions that cannot be measured.

It is possible, however, to make an educated and systematic guess about the capacity of a particular "speech act" (any form of expression, including an image) to increase the likelihood of violence, in the circumstances in which the speech act was made or disseminated. Not by coincidence, U.S. First Amendment jurisprudence criminalizes incitement not for its actual capacity to lead to violence (which can't be measured), but for the probability that it will do so: hateful incitement isn't generally a crime in the United States except when it is probable that it will lead to imminent lawless action.[13] One can estimate the dangerousness of speech in context, where dangerousness is defined as capacity to increase the chance of collective[14] violence.

A set of guidelines for estimating dangerousness[15] – especially in their current early form – may not be precise enough to define a crime. In any case dangerous speech is not criminalized, as such, in any body of law. The guidelines may be useful, however, for efforts to prevent violence by finding ways to limit the force, or impact, of inflammatory speech.

The Dangerous Speech Guidelines are summarized briefly here.[16] They are based on two key insights. First, the dangerousness of speech can be gauged with reference to five factors: the speaker, the audience, the speech act itself, the historical and social context, and

[13] Brandenburg v Ohio, 395 U.S. 444 (1969).
[14] Speech may also catalyze or inspire individual acts of violence. That process is outside the scope of this paper.
[15] My research to develop this idea has built on the work of social psychologists, historical sociologists and genocide scholars such as Ervin Staub (1989, 2003), Helen Fein (1979), Frank Chalk and Kurt Jonassohn (1990), Philip Zimbardo (2007), and James Waller (2007), and on speech act theory (Austin, 1962; Searle, 1975), as well as discourse analysis of many historical cases of inflammatory speech that preceded episodes of mass violence.
[16] For a more detailed account of the guidelines, see voicesthatpoison.org

the means of dissemination of the speech.[17] Second, some rhetorical patterns that arise in dangerous speech can serve as hallmarks or telltale signs of it.

Dangerousness can be estimated with reference to the five factors, some or all of which may contribute to dangerousness in a given case (all five are not required). In most cases, one or more factors weigh more heavily than others.

The most dangerous speech act would be one for which all five factors or variables are maximized:

- a powerful speaker with a high degree of influence over the audience most likely to react
- an audience with grievances and/or fears that the speaker can cultivate
- a speech act understood by the audience as a call to violence
- a social or historical context propitious for violence – for any of a variety of reasons, including longstanding competition between groups for resources, lack of social or political mechanisms for solving grievances, or previous episodes of violence, especially if they followed inflammatory speech
- an influential means of dissemination, such as a radio station that is the sole or primary source of news for the relevant audience

HALLMARKS OR TELLTALE SIGNS

- References to the target group as pests, vermin, insects, or animals, since such dehumanization tends to make killing and atrocities seem acceptable.

[17] These are described in Susan Benesch, Dangerous Speech: A Proposal to Prevent Group Violence, and in Ervin Staub, The Roots of Evil: The Origins of Genocide and Other Group Violence (Cambridge University Press, 1992).

- Claims that members of the target group pose a mortal or existential threat to the audience, aptly dubbed "accusation in a mirror" in a Rwandan Hutu propaganda manual.[18] The speaker accuses the target group of plotting the same harm to the audience that the speaker hopes to incite, thus providing the audience with the collective analogue of the only ironclad defense to homicide: self-defense. One of the most famous examples is the Nazi assertion, before the Holocaust began, that Jews were planning to wipe out the German people.

- Assertions that the members of the target group are besmirching the audience group, or damaging its purity or integrity

- Identifying the target group as foreign or alien, as if to expel them from the audience's group

In sum, the Guidelines are intended to allow simultaneous progress toward two essential goals which sometimes seem at odds: preventing violence and protecting freedom of expression. By shifting the focus from "hate speech" to the narrower category of Dangerous Speech and the specific harm of group violence, we hope to maintain – or even expand – protection of freedom of expression in societies at risk, while also diminishing the risk of mass violence including genocide.

Traditional Options For Preventing or Countering Dangerous Speech: Punishment and Censorship

Only two types of strategies have traditionally been used by governing authorities to suppress dangerous speech, or, for that matter, any kind of speech that they find objectionable, in diverse legal systems and societies: punish the speaker or disseminator of the speech, sometimes by means of criminal law and sometimes without law; or suppress the

[18] Alison des Forges, Leave None to Tell the Story, supra note 3; Kenneth L. Marcus, "Accusation in a Mirror," *Loyola University Chicago Law Journal 89*(3), 2012. ssrn.com/abstract=2020327

speech itself by censoring it or shutting down the medium by which it was disseminated, such as a newspaper, radio station, a social media platform, or access to all of the Internet.

These options are inadequate for many reasons. First, they aren't available when the state itself produces or endorses inflammatory, pre-genocidal speech, as in many cases including the Holocaust and the Rwandan genocide. Second, in practice states often misuse laws against hateful or inflammatory speech, interpreting them unevenly or overbroadly, or deploying them to punish or silence political opponents. Rwanda has been a case in point since the 1994 genocide: new statues criminalizing "genocidal ideology" and "ethnic divisionism" have been interpreted so broadly as to prohibit relatively mild expressions of opinion, and have been used disproportionately to punish opposition figures.

A third reason why censorship and punishment are inadequate is that they are often unsuccessful in limiting the impact of inflammatory speech on the people it is intended to inflame. In some cases such methods even backfire.[19] Prosecution and punishment can expand the audience for inflammatory speech, by publicizing it, or inadvertently help to radicalize a speaker's existing followers. Moreover, punishment by means of criminal law is slow, and in the case of international speech crimes, it has always come *after* mass violence has already occurred – sometimes many years later. Censorship often fails at its ostensible purpose of suppressing speech, especially now that speech can be disseminated so quickly and easily online, where it is difficult to suppress speech without shutting down access to the Internet entirely. To take just one illustrative example, Google and YouTube staff agonized whether to block the "Innocence of Muslims"

[19] Among myriad examples: the suspects in the April 15, 2013 bombing near the Boston marathon were reportedly inspired by the speech of the extremist cleric Anwar al-Awlaki, months after al-Awlaki was killed by a U.S. drone. His speeches were (and still are) widely available on the Internet. In another case, moments after South African political leader Julius Malema was convicted for singing a song meaning "shoot the farmer, kill the Boer," his supporters began belting out the song on the steps of the courthouse. See Susan Benesch, Words as Weapons, *World Policy Journal 29*(1), 2012, available at worldpolicy.org/journal/spring2012/words-weapons.

video clip on YouTube when it was reportedly being used to catalyze protests and violence in 2013. In the end they resolved to block it temporarily from YouTube in Egypt and Libya.[20] However the video was never successfully suppressed in either country, since it had been posted to other websites.

This is not to say that punishment and censorship are to be abandoned, only that their limited and specific utility should be taken into account. A decision to prosecute an inflammatory speaker or to suppress content may send an important symbolic message, for example, even if it does not prevent the speech from circulating. This may diminish the dignitary and psychological harm suffered by the targets of inflammatory hateful speech, and may diminish the persuasive force of the speech on others.[21]

ALTERNATIVE METHODS FOR PREVENTING OR COUNTERING DANGEROUS SPEECH

In considering methods for preventing or countering dangerous speech, it is helpful to keep in mind the three essential ingredients for communication: a speaker, a 'speech act,' and an audience that is receptive to the message. Punishment and censorship focus on the first two ingredients, as do some non-restrictive approaches – for example, persuading speakers and the media to voluntarily limit the dangerousness of the speech they produce. Other methods focus instead on the third ingredient – the audience – by working to make an audience less susceptible or receptive to dangerous speech.

[20] Susan Benesch and Rebecca MacKinnon, "The Innocence of YouTube," *Foreign Policy*, October 5, 2012. foreignpolicy.com/articles/2012/10/05/the_innocence_of_youtube

[21] The preeminent human rights organization Article 19 has produced a useful set of recommendations for implementing Article 20 of the International Covenant on Civil and Political Rights, whose provisions direct states parties to prohibit incitement to discrimination, hostility, and violence. www.article19.org/resources.php/resource/3572/en/prohibiting-incitement-to-discrimination,-hostility-or-violence

DEVELOPING AUDIENCE RESISTANCE TO DANGEROUS SPEECH

Since the goal of incitement to collective violence is to condition a group to condone or participate in attacks against members of another group, that purpose can be frustrated if the relevant audience becomes less receptive to such speech. To borrow language from public health (a field with long experience in improving human life in part by changing norms of belief and behavior), an audience must develop resistance to incitement to violence.

Resistance to dangerous speech seems to increase with the development of habits such as critical and skeptical thinking, empathy with members of other groups, and willingness to express dissent from the views expressed by a leader.[22] Members of the relevant audience must discern that (incitement to) mass violence is an instrument of political power for the inciter. The Genocide Prevention Task Force described this cogently in its report: "mass atrocities are generally perpetrated when underlying risk factors…are exploited by opportunistic elites seeking to amass power and eliminate competitors."[23] Or as Philip Gourevich put it, "The specter of an absolute menace that requires absolute eradication binds leader and people in a hermetic utopian embrace…."[24] When audience members understand that they are being manipulated to believe in a specter, they can better resist the temptation of that disastrous embrace.

---

[22] Ervin Staub, Overcoming Evil: Genocide, Violent Conflict, and Terrorism, generally and p. 21, discussing factors that "create resistance to the influences that lead to violence." See also David A. Hamburg, *Preventing Genocide: Practical Steps Toward Early Detection and Effective Action*, Paradigm Publishers, 2008, and James Waller, *Becoming Evil; How Ordinary People Commit Genocide and Mass Killing*, Oxford University Press, 2002.

[23] *Preventing Genocide, a Blueprint for U.S. Policymakers*, p. 3636, ushmm.org/m/pdfs/20081124-genocide-prevention-report.pdf. This observation applies also to mass atrocities other than genocide. See, e.g. Daniel Chirot and Clark MacCauley, *Why Not Kill Them All?: The Logic and Prevention of Mass Political Murder*, p.Princeton University Press, 2010, p. 60 (describing mass killings after British India was partitioned, in 1946 and 1947: "What motivated the local political leaders, however, was primarily their belief that it would be easier to maintain control over their communities and territories if other groups were disposed of").

[24] Philip Gourevitch, *We Wish to Inform You That Tomorrow We Will Be Killed With Our Families: Stories of Rwanda*, Picador, 1999, p. 95.

In a similar vein, influential leaders or community members can guide an audience to resist incitement by speaking out against it, warning against its effects, or warning against violence itself. To borrow another medical term, 'injections' of counterspeech can diminish the risk of violence.

Finally, informal (uncodified) speech regulation can be extremely effective, and it can shift quickly, in frightening but also salutary directions. This was true even before digital communications and the Internet provided acceleration. Consider, for example, the likelihood that an American political figure will use the 'n-word' in public in 2014 and remain in office. It is close to zero. Only a few decades ago, the use of that word was anything but prohibited: it would have boosted one's political fortunes in some parts of the country, where savage violence against African-Americans was also tolerated. American society still harbors racism in many forms, but there has been undeniable progress, and discourse norms related to race have changed dramatically.

None of these options impinges on freedom of expression, since they impose no state punishment for inflammatory speech – or any speech. The field is young, but useful experiments have recently been performed, including 'inoculating' audiences against inflammatory speech by explaining that such speech is a tool used by leaders to manipulate groups, as well as related interventions to render such speech less influential. Some of the experiments have also been independently evaluated to gauge their capacity to make audiences resistant to dangerous speech, and this has produced some specific findings, or lessons learned. Those include:

- Increasing empathy with members of other groups counteracts incitement, since it makes it difficult to see other people as subhuman – an essential element of the process described by sociologists and historians as "social death," and which,

according to the philosopher Claudia Card, distinguishes genocide from other forms of mass murder.

- Counterspeech by influential members of a community can lead group members to respond more positively to members of other groups, and diminish their susceptibility to incitement to mass violence. These speakers may be political, cultural, or religious leaders, or simply "active bystanders" with the courage to speak up. [25]
- Modeling resistance to incitement, even in fictional accounts, can increase such behavior among members of a group.[26]

"INOCULATING" THE AUDIENCE AGAINST INFLAMMATORY SPEECH

To prevent mass violence, especially in societies at high risk for it, advocates have begun to experiment with media programming to render audiences less likely to become convinced by inflammatory speech, or to act on it. We call this 'inoculating' an audience, following the example of the nongovernmental organization Radio la Benevolencija, which uses the term to refer to its own pathbreaking work. The term is apt in several ways. First, hateful and inflammatory speech would be exceedingly difficult to eradicate, like pathogens. Second, even inoculation of a significant part of an audience would be useful – one need not reach the entire population. Just as with campaigns of inoculation, it would be ideal to reach all members of a population, but that is not necessary to prevent an epidemic. In any society, even a relatively democratic and peaceful one, some people advocate violence against minority groups. These haters cannot bring about genocide or large-scale atrocities, however, unless they convince a critical mass of non-extremists to agree with them.

---

[25] Paul M. Sniderman and Louk Hagendoorn, *When Ways of Life Collide: Multiculturalism and its Discontents in the Netherlands*, Princeton University Press, 2007, 115-8118 (reporting results of the 'Segregation,' 'End of Interview,' and 'Political Leader' experiments, all of which showed considerable shifts toward accepting cultural pluralism, under political and social influence. The change was largest, paradoxically, among subjects who value conformity).

[26] Staub, Overcoming Evil, supra note 1, at 369-403.

The main question, of course, is how to inoculate successfully. In the past decade, and especially in the past two years, several attempts have been made to inoculate populations against dangerous speech and, in some cases, have been independently evaluated. These results, described below, give cause for some optimism and, at least, for further experiments.

Radio la Benevolencija (RLB), based in Amsterdam and working in several central African countries, has collaborated with the social psychologist and genocide scholar Ervin Staub to develop what they call "knowledge tools" – guides on how to deal with an array of manipulative pressures that move individuals and whole societies to physical and mental harm, and how to resist such pressures. RLB delivers these tools in entertaining programs such as a radio soap opera called Musekeweya ("New Dawn"),[27] which has become a hit in Rwanda since its launch there in 2004. What makes Musekeweya distinct from other soap operas in Rwanda (or indeed, elsewhere), according to RLB, is that it "explicitly deals with the psychology of incitement to hate and violence that leads to mass conflict." Musekeweya is set in two fictional but familiar Rwandan villages, Bumanzi and Muhumura, situated on top of two of Rwanda's innumerable hills. The two groups of villagers are polarized by land disputes, and by their different (although unnamed) ethnic identities. Many conflicts have arisen and have been resolved in the course of the show's nearly 10 years of weekly episodes.

Musekeweya's impact on its listeners was independently studied by a scholar who described her effort[28] as "the first experimental evaluation of a radio program's impact on intergroup prejudice and conflict in a real world setting."[29] In her year-long study, Elizabeth Levy Paluck found "a pattern of norm and behavior change" and an

[27] Another forthcoming RLB program has this memorable, brilliant title "Hate: a Course in Ten Easy Lessons."
[28] Elizabeth Levy Paluck, Reducing intergroup prejudice and conflict with the mass media: A field experiment in Rwanda, PhD dissertation, Yale University, 2007, gradworks.umi.com/326730.pdf
[29] Id., at 2.

increase in empathy, on the part of Musekeweya listeners, for other Rwandans. With specific respect to "some of the most critical issues for Rwanda's post conflict society, such as intermarriage, open dissent, trust, and talking about personal trauma," she found change, not only in perceptions but also in behavior.

Listeners to Musekeweya were more likely to think for themselves, and to express their own dissenting views, i.e. not only to think but also to behave differently. They were "more likely than members of the comparison group to believe in speaking their minds and to actually do so, to express controversial views, and to show independence from authority."[30] Similarly, Musekeweya viewers were more likely than members of the control group to reject the statement, "if I disagree with something that someone else is saying or doing, I keep quiet."[31] This is encouraging because it has been widely reported, in pre-genocidal periods, that extremist views gain purchase within a society when dissenters remain quiet.

In light of these findings, further experiments and further study would be useful. To follow up and expand on Paluck's work, RLB is now taking part in a more exhaustive academic study of the impact of its programming on audiences.

I conducted another effort in Kenya in 2012, in collaboration with the NGO Media Focus on Africa. We produced four episodes of a highly popular, longstanding Kenyan television comedy/drama called Vioja Mahakamani ("Events in the Courtroom"), on the topic of inflammatory speech. In each episode of the show, which has aired weekly over the airwaves of the public broadcaster KBC since 1974, one or more characters is accused of a crime, and the case is adjudicated in the course of the 30-minute show. In each of our four episodes, a Kenyan (or group of Kenyans) stands accused either of

[30] Ervin Staub, Overcoming Evil, supra note 1, at 373.
[31] Id at 374.

making inflammatory speech (at a rally or on a printed flyer, for example) or of acting upon it. The episodes define inflammatory speech and illustrate: 1) that it is a political tool typically intended to aggregate the power of the speaker, and 2) that it can lead to mass violence. The episodes' impact on Kenyan audiences has been evaluated by Scholars at the Center for Global Communication Studies at the University of Pennsylvania evaluated the episodes' impact on Kenyan audiences, and indicates that Kenyans who watched the episodes felt better able to identify and to resist incitement.

It should be noted that many other efforts intended to promote the rule of law, or build democratic institutions, may also help populations to become less receptive to incitement – as one of their favorable consequences. Therefore any project to build democracy can be seen as an anti-genocide effort as well. Here we focus specifically on efforts to diminish the power of inflammatory speech by helping audiences to become more resistant to it.

"Injecting" Counterspeech

Inoculation takes some time, and therefore should be conducted in advance (just like the more familiar kind of inoculation, against disease) before the risk of violence becomes acute. Especially (but not only, of course) during that acute stage, there is a second type of alternative method for diminishing the force or effectiveness of dangerous speech that I describe with the general term counterspeech – or speech to refute dangerous speech.

Several distinct forms of counterspeech may be effective at forestalling mass violence, and there are more to be identified. Three with some promise are: counterspeech by influential leaders, counterspeech from a variety of sources in unison, and speech that effectively refutes false rumors.

COUNTERSPEECH BY INFLUENTIAL LEADERS

Not surprisingly, some of the indicators of successful counterspeech are the same as the indicators of dangerousness for inflammatory speech. For example, there is evidence of success when a speaker with influence over the relevant audience gives a strong signal of disapproval of inflammatory speech – or of violence itself. Since mass violence is often supported (or even carried out) by state authorities, they cannot be relied upon to oppose it, however there are some notable examples that have reportedly made all the difference. For example, according to research on why Hindu-Muslim riots and massacres have happened at some volatile times and places – but not in others – in India, influential figures have been able to thwart violence by publicly withdrawing their support for it, even where extremist parties controlled relevant state governments.[32] Likewise the King of Denmark and Danish political leaders have been credited with helping to save the lives of nearly all of Denmark's Jews during the Holocaust, in part by keeping them firmly within the universe of moral obligation in the minds of the Danish people, referring to Jews as part of the same national community as other, non-Jewish Danes. "I considered our own Jews to be Danish citizens," King Christian wrote, "and the Germans could not touch them. The prime minister shared my view and added that there could be no question about that."[33]

Also in societies not at risk of genocide, there are indications that killings have been forestalled with counterspeech. After the killing of filmmaker Theo van Gogh in 2004, for example, Amsterdam's mayor

---

[32] Amrita Basu, "When Local Riots Are Not Merely Local," *Economic and Political Weekly 29*(40), 1994, pp. 2619-20, jstor.org/stable/4401857; Jayati Chaturvedi and Gyaneshwar Chaturvedi, "Dharma Yudh; Communal Violence, Riots and Public Space in Ayodhya and Agra City: 1990 and 1992," in Paul Brass, ed., *Riots and Pogroms*, New York University Press, 1996, pp. 187-90, as cited in Donald Horowitz, *The Deadly Ethnic Riot*, University of California Press, 2003, at 517.

[33] Bo Lidegaard. *Countrymen*, Alfred A. Knopf, 2013, at 20 and *generally*, describing Danish leaders' speech refusing to separate Jews from the rest of the nation, rhetorically as well as literally.

Job Cohen spoke out firmly against the angry anti-Muslim rhetoric and sentiment that followed. He "intiated the peace script," as the New York Times later put it,[34] for example by telling the people of his city, "An Amsterdamer is murdered. You fight with the pen and, if necessary, to the court. But never take the law in your own hands."[35] In the days after van Gogh was killed, revenge attacks against Muslims happened in some Dutch cities, but not in Amsterdam.[36]

When and where dangerous speech is proliferating, influential leaders – political, religious, and cultural – must be made aware of their capacity and indeed their responsibility to attempt to prevent violence with counterspeech. This can be framed in terms of the Responsibility to Protect (R2P), which emphasizes incitement to grave crimes, and asserts that states have an affirmative obligation to counter incitement.[37] In a 2009 report on implementing the Responsibility to Protect, U.N. Secretary General Ban ki-Moon emphasized incitement, pointedly listing a number of pre-genocidal situations in which the international community failed to react to incitement that doubtless constituted dangerous speech:

"The world body failed to take notice when the Khmer Rouge called for a socially and ethnically homogenous Cambodia with a "clean social system" and its radio urged listeners to "purify" the "masses of the people" of Cambodia. Nor did it respond vigorously to ethnically inflammatory broadcasts and rhetoric in the Balkans in the early 1990s or in Rwanda in 1993 and 1994 in the months preceding the genocide. Despite several reports during those critical months by the United Nations Assistance Mission in Rwanda and the Special

[34] Russell Shorto, The Integrationist, *The New York Times*, May 28, 2010, available at nytimes.com/2010/05/30/magazine/30Mayor-t.html

[35] paleisamsterdam.nl/en/the-palace/explore/the-palace-and-dam-square/theo-van-gogh

[36] worldmayor.com/results06/interview_amsterdam.html

37 The World Summit Outcome Document of 2005, which set forth the Responsibility to Protect, asserts that, "Each individual State has the responsibility to protect its populations from genocide, war crimes, ethnic cleansing and crimes against humanity. This responsibility entails the prevention of such crimes, including their **incitement**, through appropriate and necessary means." (emphasis added) *See* 2005 World Summit Outcome, 24 October 2005, A/RES/60/1, paragraph 138, un.org/en/preventgenocide/adviser/pdf/World%20Summit%20Outcome%20Document.pdf

Rapporteur on extrajudicial, arbitrary or summary executions on the incendiary programming of Radio Mille Collines, there was no attempt by the international community to jam those hateful and fateful broadcasts."[38]

To uphold the responsibility to protect populations against mass violence in future, the Secretary-General declared, "When a State manifestly fails to prevent such incitement, the international community should remind the authorities of this obligation and that such acts could be referred to the International Criminal Court, under the Rome Statute." This form of state responsibility is not difficult to discharge, he asserted. "Because of the typically public and explicit character of such incitement, it should be relatively easy to identify it and to rally international support for efforts to discourage it."

COUNTERSPEECH IN UNISON

Counterspeech can be effective also when it comes from a wide variety of sources, speaking in unison. Kenya produced a striking example of this in the weeks and months before its presidential election in March 2013, the country's first since inflammatory speech and severe violence accompanied the attempted election of 2007.

Thought leaders of all kinds called on Kenyans to forsake violence. Ecumenical groups of clerics appeared on billboards and on the radio, calling for peace. Popular football stars recorded brief public service announcements, appealing directly to young men like themselves to remain calm. Even television journalists, who anchored nonstop coverage of the voting and vote-counting, stepped somewhat out of their role as news reporters to appeal directly to viewers to maintain the peace. This unprecedented volume of "peace propaganda" was effective, according to anecdotal evidence from Kenyans who said it helped them to remain calm and patient even as the vote-counting

[38] Implementing the Responsibility to Protect, Report of the Secretary-General, Jan 12, 2009, A/63/677, at 23, responsibilitytoprotect.org/SGRtoPEng%20(4).pdfpdf

dragged on. Moreover the election was completed with only one episode of serious violence, by a local extremist group which tried to disrupt the election by attacking polling places in the city of Mombasa. Although several people were killed, the violence did not spread, as it did in 2007-8.

It must also be noted that many Kenyans, especially but not only supporters of the losing presidential candidate Raila Odinga, felt that the overweening emphasis on keeping peace had the pernicious effect of suppressing dissent, political debate, and even hard-hitting news reporting. The commentator and columnist Patrick Gathara complained, for example, that Kenya had fallen into a "peace coma."[39] It will be interesting and important to see how Kenya strikes the balance during its next presidential election.

COUNTERSPEECH TO REFUTE FALSEHOODS AND SUPPLY RELIABLE INFORMATION

In most cases, it is difficult to prove a causal link between specific examples of inflammatory speech and violence, but there are exceptions. A notable one is false rumors. In 2007-8, for example, Michele Osborn of Oxford University traced rumors as they moved through the Nairobi slum of Kibera, and eventually inspired violence.[40] In response to the special power of false rumor to ignite mass violence, several projects to counter this speech have emerged, independently and in different parts of the world.

In Ambon, Indonesia, where Muslim-Christian violence is all too common, it has often been catalyzed by false rumors that a member of one of those groups had been attacked by members of the other. In recent years, the rumors spread faster and further, via SMS messaging. In response, a group of self-described "Peace

[39] Patrick Gathara, "Coming Home to Roost," April 26, 2013, gathara.blogspot.com/2013/04/home-to-roost.html

[40] Michelle Osborn, Fuelling the Flames: Rumour and Politics in Kibera, *Journal of Eastern African Studies*, *2*:(2,), 2008, 315-327, dx.doi.org/10.1080/17531050802094836

Provocateurs" began countering false rumors by SMS in September 2011.[41] Their counterspeech provides tangible proof that the rumors are false – where a girl was said to have been seriously injured, for example, they send a photograph showing that she is healthy. Already, the Provocateurs (who are Christians and Muslims, perhaps adding to their credibility) have had success.

In Kenya the *Nipe Ukweli* (Kiswahili for "Give me Truth") campaign was born in January 2013, two months before the election, because false rumors had been shown to give rise to specific violence in 2008, when they were widely distributed by text messages.[42] The name Nipe Ukweli was intended to emphasize the fact that lies are often used by inflammatory speakers to manipulate a population or to 'play' them in Kenyan parlance. The project sought to stimulate (gentle) indignation against this practice, and to encourage Kenyans to resist and, where possible, to refute false rumors. Nipe Ukweli was also inspired by the example of one Kenyan Twitter user who, in August 2012, had countered a rumor that Muslims in the Coast Province were burning churches en masse, by tweeting a photo of one of the churches that according to rumor had been burned down.

Influencing the Speaker

There is also new evidence that counterspeech can be effective online, in some cases. More than ever before, hateful expression easily crosses the boundaries between normative groups. For example, if an American man posts a rape joke on a Facebook page instead of telling it to a group of his friends in a locker room or a men's club (more common in the past), women are more likely to see the joke and be hurt by it. This causes new pain. It also presents new opportunities for attempting counterspeech – and for measuring its effectiveness, since

[41] Andrew Stroehlein, How 'peace provocateurs' are defusing religious tensions in Indonesia, The Independent, March 12, 2012, independent.co.uk/news/world/asia/how-peaceprovocateurs-are-defusing-religious-tensions-in-indonesia-7562725.html
[42] Osborn, supra note 43.

the trajectories and effects of speech can be measured much more easily online than offline.

For example the Kenyan monitoring project Umati ('crowd' in Kiswahili) collected more than 5,000 examples of hateful and dangerous speech from Kenyan online spaces, including blogs, forums, newspaper sites including comments, Facebook pages, and Twitter, during the course of 2013. A strikingly small proportion of the examples (fewer than three percent) were found on Twitter, although Kenyans were producing hundreds of thousands of tweets. (During a debate among the Kenyan presidential candidates in early 2013, the hashtag #KEdebate13 was the number one trending topic on Twitter worldwide.)[43] In the crucial days before and after the March 4, 2013 election, KOTs (Kenyans on Twitter) produced abundant counterspeech which, in at least some cases, convinced the producers of hateful Tweets to stop, or even to apologize.[44]

Working together, Twitter, the Umati team, and I have collected other examples in which hateful speakers recanted or apologized, in response to counterspeech from other Twitter users. For now, we have found this effect in the United States, in France, and in Kenya: a small but diverse list. During 2014 and 2015, we will continue this research on a larger scale, and in more normative environments, including countries at risk of genocide.

CONCLUSION

I hope to have primed the pump for further brainstorming, research, and genocide prevention in countries at risk, by describing a new set of approaches for attempting to prevent or counter speech that has the capacity to catalyze mass violence. They cannot work in all situations,

[43] nairobiwire.com/2013/02/presidential-debate-trend-worldwide-on.html
[44] Umati Final Report, Sept 2012-May 2013, iHub Research and Ushahidi, dx.doi.org/10.1080/17531050802094836

of course, but may constitute some useful tools to add to a still-sparse box.

Two useful efforts would be to develop a taxonomy of counterspeech and to create a guide for choosing which tools to use in a particular situation. Like the Dangerous Speech Guidelines, a taxonomy of counterspeech would identify factors that make counterspeech more effective and more powerful. This can be done, as more experiments are conducted, by collecting examples of counterspeech and studying them for patterns and for effectiveness.

A guide or framework for choosing which tools hold the most promise for preventing or countering dangerous speech in a particular situation should consider many of the same contextual factors that make speech dangerous. Some of the factors such a framework should examine include:

1) The speakers: what are their motivations and intentions and what is the basis of their influence with the audience? Possible intervention approaches might include:

- *Persuading some speakers to voluntarily limit the dangerousness of their speech*, such as guidelines for discourse agreed to by political parties before an election;
- *Discrediting a speaker*, for example by refuting falsehoods
- *Enlisting influential persons* to provide countering messages

2) The target audience: what are the factors that make them receptive to the message of dangerous speech? What are the factors motivating their animus toward the target group?

3) The means of dissemination: how is dangerous speech most commonly disseminated and what are the most common sources of information and means of communication used by the target audience? Can the same channels be used to inject counterspeech or provide alternate sources of information?

4) The message: counterspeech must be carefully tailored to the relevant audiences. The preceding factors will be useful for a detailed understanding of the context in which a message will be received and for creating content that will resonate with the audience. Some possible messaging approaches include:

- *Humanizing the out group.*
- *Appealing to self interest.* In this vein, another program in Kenya reminded message recipients of how much they themselves had suffered as a result of the violence.[45] Often when one group targets a second one for violence, both groups end up suffering, even while damaging the entire society and country that they share.
- *Appealing to common interests and identity*, including the targets and the audience in the same group of 'us,' inside the same universe of moral obligation.

Experiments to diminish the impact of inflammatory speech are most likely to succeed when conducted by 'insiders,' – members of the groups that the work seeks to influence – because of their familiarity with the social and cultural context and because they usually have the greatest capacity to influence fellow members of the group and to understand what messages will appeal to them. Where insiders are not available or able to take the lead, they should at least play a major role in such projects.

This paper outlines some methods that have been shown through research and practice to have some success in countering dangerous speech. Many other potential methods – such as non-repressive interventions to prevent media dissemination of dangerous speech – are being studied and tested. In fact, an increasing number of projects in an expanding range of countries are producing valuable data for understanding what role countering dangerous speech can play in

[45] This project is called PeaceTXT, and is described in my paper *Countering Dangerous Speech to Prevent Mass Violence During Kenya's 2013 Election*, also available at the Sudikoff seminar.

preventing mass atrocities in at-risk societies, as well as in assessing the potential of dangerous speech to serve as an indicator for the imminence of mass violence. Providing opportunities for those involved in such projects to share their experiences will significantly advance the number and effectiveness of available tools for countering dangerous speech.