DANGEROUS SPEECH PROJECT

Speech as a Driver of Intergroup Violence: A Literature Review

Cathy Buerger

INTRODUCTION

People have been trying to understand the catalysts of human behavior, especially violent behavior, for thousands of years. In this review, we focus on a topic that has been largely overlooked in the literature so far: how speech, especially public, often online speech, can inspire civilians of one group to attack civilians of another or create an atmosphere in which such violence is encouraged. We refer to this as intergroup violence, and distinguish it from other forms of collective violence such as war. The relevant groups are often defined by identity markers including immutable ones, such as race, ethnicity, religion, or sexual orientation, but intergroup violence can also target groups defined by other characteristics, such as occupation. Researchers have explored other possible drivers of violence against individuals, such as violent movies and video games, producing bodies of literature that other scholars have already synthesized.[1] What follows seems to be the first standalone review on our topic.

Investigators who work on factors that have led to mass atrocities often assert a causal relationship between speech and violence. Some have tried to use statistics to establish this. David Yanagizawa-Drott (2014), for example, used Rwandan topographic data to argue that interference from the country's hills caused substantial variation in which villages received transmissions from a notorious radio station that spread incitement to kill in the months before the 1994 genocide. He argues that this variation corresponds to the relative numbers of people later tried for genocide in the corresponding villages (a proxy for numbers of people who participated in the genocide). Adena et al (2015) also used radio exposure data to study the impact of Nazi propaganda, finding that it led to increased support of Nazi policies, at least among those who did not "disagree with the propaganda message a priori" (1890).

Other studies offer anecdotal evidence of causation. For example, a Human Rights Watch (2011) report detailing violence in Côte d'Ivoire following its presidential election in 2011 and 2012 includes a speech delivered by Charles Blé Goudé, Youth Minister under then President Koudou Laurent Gbagbo, telling Gbagbo supporters to secure their

[1] For a review of the literature on video games and violence, see Prescott, Anna T., James D. Sargent, and Jay G. Hull. 2018. "Metaanalysis of the relationship between violent video game play and physical aggression over time." *Proceedings of the National Academy of Sciences* 115, no. 40: 9882-9888. Robert Sapolsky also offers an excellent summary of the literature on the relationship between violent media consumption and aggression in his book, *Behave* starting on page 198. See Sapolsky, Robert. 2017. *Behave: The biology of humans at our best and worst*. New York: Penguin Press.

neighborhoods against 'foreigners' (other West African nationals and ethnic groups from the northern part of the country). Multiple victims later said they were attacked by people who spoke of Blé Goudé's 'order.'

Such examples, in which evidence points to one speech that incited a particular attack, are relatively rare, since dangerous speech affects beliefs and behavior over time, gradually moving people toward condoning or committing violence against members of another group. Also, lack of control over extraneous variables often makes it difficult for researchers to identify a direct causal relationship between a single speech act, and action..

Experimental studies on the relationship between speech and violence have been limited by serious ethical obstacles. It is difficult to design experiments on what factors drive people to commit violence without risking harm. Before academic researchers were required to follow ethical research guidelines, several studies tried to provoke people to commit violence (or at least to believe they were committing violence) in experimental settings. One of the most famous was Stanley Milgram's (1963) experiment on obedience, in which participants were told to follow a 'teacher's' instructions to deliver painful electric shocks to a 'learner' (an actor pretending to suffer from the shocks)[2]. Another is the Stanford Prison Experiment, led by Philip Zimbardo (1972), in which subjects were assigned to play the role of either prisoner or guard.[3] These studies focused on group dynamics and obedience, and fall outside of the scope of this literature review. We mention them since they are landmark related research that also led to consensus against testing a person's willingness to commit violence in a lab setting, since that may traumatize the experimental subjects (*The Belmont Report,* 1979).

In this literature review, our scope extends beyond studies that attempt to measure whether speech directly caused violence. Each article in this review adds to our understanding regarding the manner in which speech may move someone to commit or condone violence against members of another group, and the factors that play a role.

The review has five sections:

1. Theories on the nexus between speech and violence
2. Rumor
3. Dehumanization

[2] J.M. Berger (2009) partially replicated this experiment, following the same initial protocol, but introducing several safeguards to protect participants; he pre-screened participants, informed them three times that they could stop participating at any point, had the experimenter administer a 'very mild' sample shock to participants, and informed participants immediately after the study that the learner had not actually been shocked. The experiment was also run by a clinical psychologist "who was instructed to end the study immediately if he saw any signs of excessive stress" (2). His study produced comparable results to Milgram's. See Burger J. M. 2009. "Replicating Milgram: Would people still obey today?" *American Psychologist.* 64:1–11.

[3] Both of these studies have been criticized because they may have inflicted severe psychological harm on the experimental subjects who weren't informed in advance of the risks they were taking, nor given a chance to opt out. In addition to ethical concerns, other researchers have wondered whether the findings of Zimbardo's study can be generalized to 'real-world' settings (Banuazizi & Movahedi, 1975).

4. Online speech and hate crimes
5. Context-specific reports

The first three sections address how, and in what conditions, speech seems to move people to condone or commit intergroup violence. Sections four and five largely contain case studies from particular times and places, such as leading up to and during the Rwandan genocide.

1. THEORIES ON THE NEXUS BETWEEN SPEECH AND VIOLENCE

As Susan Benesch (2011, 254) notes, "the idea that inflammatory speech is a catalyst for genocide is widely believed [...] but the impact of speech on the ground is complex, and difficult to measure or prove." This has not stopped scholars and international courts from asserting causation, even without evidence. Benesch and other scholars, such as Jonathan Leader Maynard, have developed theories to understand the relationship between speech and violence, offering guidance for how factors like the authority of the speaker and the social and historical context affect the impact of speech. As they, and the other scholars included in this section, note, there is rarely strong evidence that speech alone is directly causal to violence, but there is evidence that speech may be jointly causal when combined with other factors.

- **Benesch, Susan. 2011. "The Ghost of Causation in International Speech Crime Cases."** ***Propaganda, War Crimes Trials and International Law: From Speakers' Corner to War Crimes*****, Edited by Predrag Dojcinovic. New York: Routledge. 254-268.**
  https://dangerousspeech.org/the-ghost-of-causation-in-international-speech-crime-cases

  In this book chapter, Benesch notes that in incitement to genocide cases, international courts have declared that speech caused genocide without evidence, though there is no legal need for such evidence, since incitement is a crime whether it is successful or not. As she points out, causation is difficult to identify "since the effect of speech on large groups of people is hard to measure, poorly understood, and [speech] is only one of a constellation of forces that affect why people act as they do" (257). Instead Benesch introduces a five-part framework for estimating the 'dangerousness' (the capacity to inspire intergroup violence) of speech. The five parts are the speaker, the audience, the content of the speech act, the socio-historical context, and mode of transmission (262-264). Rather than attempting to demonstrate a causal relationship between a particular speech act and specific violence, this framework is intended for gauging the *likelihood* that certain speech acts have led, or can lead, to mass violence.

- **Dangerous Speech Project. 2018. "Dangerous Speech: A Practical Guide." Dangerous Speech Project. https://dangerousspeech.org/guide/**

  In this paper, the authors describe many features of dangerous speech, which they define as "any form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or commit violence against members of another group." The guide offers a five-part framework (the same one in the "Ghost of Causation" chapter above, but described in more detail here) for estimating whether, and to what extent, speech is dangerous. The five elements to consider are the message itself, the speaker, the audience or people exposed to the speech, their social and historical context, and the medium by which the message spread. As part of the 'message' section, the guide also offers a list of 'hallmarks,' or rhetorical patterns, that are often found in dangerous speech, and illustrates them using specific examples from around the world.

- **Leader Maynard, Jonathan. 2014. "Rethinking the Role of Ideology in Mass Atrocities." *Terrorism and Political Violence*, *26*(5), 821-841. https://ora.ox.ac.uk/objects/uuid:a42b946e-fde8-4f68-8e3d-48eec742bf0b**

  Leader Maynard posited that ideology may encourage people to commit, or permit, mass violence by shaping motives to commit atrocities, creating a perception that violence is permissible, or providing a narrative to justify violence after the fact (11). The author discussed who might be affected by ideology and how it spreads through society. He noted, for example, that effective ideological dissemination often occurs on several levels at once; it may be communicated through social interactions and also through more institutionalized channels, such as through public education (11). The author also offered six 'justificatory mechanisms'' to describe the ways that ideology increases a person's willingness to kill: dehumanization, guilt-attribution, threat-construction, deagentification (suggesting perpetrators lack responsibility for killing, by, for example, saying that atrocities were 'inevitable'), virtuetalk, and future-bias (13).

- **Leader Maynard, Jonathan, and Susan Benesch. 2016. "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention." *Genocide Studies and Prevention: an International Journal.* 9(3), 70-95. http://doi.org/10.5038/1911-9933.9.3.1317**

  In this paper, the authors knit together their respective work on dangerous speech and the ideological dynamics of mass atrocities and offer an integrated model to identify the types of speech and ideology that raise the risk of atrocities and genocide. Their model suggests that "identifying when speech is in danger of causing violence" requires inquiry into two aspects: the content of the speech and its context (including the speaker, audience, social and historical context, and mode of

dissemination) (78). The authors note that although dangerous content often describes members of the out-group (e.g. characterizing them as a threat to the in-group), the out-group need not be mentioned at all. For example, sometimes dangerous speech relies on 'virtuetalk,' a term coined by Leader Maynard (2014) to describe speech that suggests that those who do not participate in violence are 'weak' or worthy of social ridicule (84). The authors conclude with a brief description of how the model is useful in developing strategies to counter dangerous speech and ideology.

- **Waller, James. 2007. *Becoming Evil: How Ordinary People Commit Genocide and Mass Killing*. Oxford: Oxford University Press.**

  In this book, Waller advances a model to explain "how ordinary people commit genocide and mass killing." He suggests that the process of change from ordinary person to perpetrator involves changes in worldview (including collective values, authority orientation, and social dominance), the psychological construction of the 'other,' and the influence of group membership on notions of cruelty. He used examples from places such as the Balkans, Guatemala, and Rwanda to illustrate how speech, myths, and symbols can move people toward condoning and committing violence. Waller considers how language dehumanizes members of other groups and creates an 'us vs. them' dynamic, and how 'euphemistic labeling' of violence allows perpetrators to morally disengage from those acts (211).

- **Wilson, Richard Ashby. 2017. *Incitement on Trial: Prosecuting International Speech Crimes*. Cambridge: Cambridge University Press.**

  In this book, Wilson describes the efforts of international criminal tribunals to prosecute public figures for incitement. Specifically, he explores the causal connection between speech and subsequent violence: how tribunals have handled it and how social science illuminates it. Based on examples from international courts such as the International Criminal Tribunal for the Former Yugoslavia (ICTY) and the International Criminal Tribunal for Rwanda (ICTR), Wilson concludes that courts have "relied on outdated models of propaganda" (9), "demanded proof of causation even when it is not warranted" (9), and have failed to form a consensus about the evidentiary threshold for proving the direct effects of speech (8). He also describes how social science research on the subject of persuasion has refuted notions of a 'mechanistic' relationship between propaganda and perpetrators, instead demonstrating the agency of the audience and the complex ways that people receive and interact with messages. 'Revenge speech,' for example, may serve to lower empathy for the outgroup and increase the willingness of listeners to morally justify violence.

2. RUMOR

The articles in this section all seek to explain the role of rumors in inciting acts of violence. An important common theme is that rumors are not only reports of fictional incidents, but also reframe real incidents to fit within historical narratives of intergroup tension. Das (1998), for example, notes how rumors described former Indian Prime Minister Indira Gandhi's assassination as revenge for Operation Blue Star, a 1984 military operation during which the Indian military attacked a Sikh leader and his armed followers who had encamped in a Sikh temple. This rumor contributed to the anti-Sikh violence that followed the assassination. The reframing that happens through rumors is important — as Espeland (2011) notes, rumors not only respond to a conflict, but are constitutive of it (18).

- **Arun, Chinmayi. 2019. "On WhatsApp, Rumours, and Lynchings."** ***Economic & Political Weekly*** **54(6), 30-35. https://doi.org/10.5325/jinfopoli.10.2020.0276**

  In this paper Arun describes a wave of lynchings in India in 2019, following rumors spread through WhatsApp, falsely claiming that people were kidnapping children. Arun argued that the Indian government has incorrectly treated these rumors as primarily a 'fake news' problem, when they are actually a form of incitement to violence. She also discusses the steps that WhatsApp took in the wake of the lynchings to prevent other content that can incite violence from circulating on its platform, such as limiting the number of accounts to which one can forward a message, at one time — and offered suggestions for improvement..

- **Bhavnani, Ravi, Michael G. Findley, and James H. Kuklinski. "Rumor dynamics in ethnic violence."** ***The Journal of Politics.*** **71(3), 876-892.** **https://doi.org/10.1017/S002238160909077X**

  The authors used agent-based modeling (a model that examines the impact of individual decisions and behaviors on a system as a whole) to examine the role of audience and speaker characteristics in the emergence and persistence of 'ethnic-centered' rumors. One version of their model only considered 'within-group rumor dynamics,' while a second version examined 'across group interactions,' (880). In the former, rumors were most widely accepted when group leaders who frequently engaged with their audiences endorsed them. Further, extreme rumors were more widely accepted if the audience already held extreme views. When a context with rival ethnic groups was presented in order to study the interactions between groups, the level of rumor 'survival' was, in part, determined by whether the audience held extreme or moderate beliefs, and in part by whether the ethnic groups interacted. Leaders also played a role; for example, "when one group's leaders persist in advocating moderation, rumor propagation remains low in both groups" (876).

- **Das, Veena. 1998. "Specificities: Official Narratives, Rumour, and the Social Production of Hate."** ***Social Identities*****. 4(1), 109–30.** https://doi.org/10.1080/13504639851915

  In this seminal piece on rumor and violence, Das explores the "social production and circulation of hate" (109) by analyzing the massacre of Sikhs that followed the murder of India's former Prime Minister Indira Gandhi in 1984. Two of Gandhi's Sikh bodyguards killed her, and in the following days, rumors spread through Hindu communities that the Sikhs were plotting to take over the country — in part as revenge for Operation Blue Star (in which the Indian military attacked a Sikh temple where a prominent Sikh leader had been staying with a group of his heavily-armed followers). The rumors played on preexisting narratives that characterized Sikhs as fundamentally vengeful, fanatical, aggressive, and incapable of loyalty (124) and described Hindus as weak, making violence against the Sikhs seem like self-defense. Thousands of Sikhs were killed during the riots that followed the assassination (121). Thus, Das argues that rumors are particularly effective at motivating people to act because they contribute to a feeling of "mounting panic" (117).

- **Espeland, Rune Hjalmar. 2011. "Autochthony, Rumor Dynamics, and Communal Violence in Western Uganda."** ***Social Analysis*** **55(3), 18-34.** https://doi.org/10.3167/sa.2011.550302

  Espeland argues that rumors played a significant role in catalyzing violence in Western Uganda between the Banyoro and Bafuruki ethnic groups by reframing in moral terms what other narratives (such as in major Ugandan newspapers) had described as a land dispute. Rumors are not only a response to conflict, but also "constitutive of the situation," constructing a shared moral narrative (18). In this case, for example, rumors described the Banyoro as amoral and as practicing witchcraft, both of which made their Bafuruki attackers feel morally justified.

- **Osborn, Michelle. 2008. "Fueling the Flames: Rumour and Politics in Kibera."** ***Journal of Eastern African Studies*** **2(2), 315-327.** https://doi.org/10.1080/17531050802094836

  Osborn argues that rumors served as a crucial catalyst of Kenya's 2007-2008 post-election violence. Describing rumors that circulated over SMS, she explains how technology increased their reach. Osborn also draws out the differing goals of various actors who circulated rumors. Some, such as politicians, intentionally created and disseminated misinformation to mobilize voters. Others passed along warnings they believed to be true in order to protect loved ones. As Osborn illustrates, one need not intend to incite violence in order to do it by passing on a rumor.

3. DEHUMANIZATION

Although it is not the only dangerous rhetorical technique, dehumanization is one of the most familiar, and it is the subject of abundant scholarship. Instead of covering the topic exhaustively, we include only articles on dehumanization and violence here. Several pieces, including those by Beyond Conflict (2019), Giner-Sorolla et al. (2012), and Haslam (2006), include thorough discussions of relevant literature in addition to notable findings of their own that add nuance to the subject. For example, Haslam (2006) distinguishes between 'animalistic dehumanization' and 'mechanistic dehumanization.' Rai et al. (2017) explains the many different ways in which dehumanization can interact with someone's support of violence against another group; for example, hearing dehumanizing speech about another group can lead to increased support for violence against that group, and conversely, imagining oneself committing violence against a group can lead to viewing its members as less than human.

Although many scholars have argued that there is a strong relationship between dehumanizing rhetoric and mass violence, there is no consensus on this point. For example, several authors (Bloom 2017, Smith 2016, and Rai et al. 2017) have examined the humiliation that often accompanies violence, and because of this, some doubt that dehumanization really contributes to violence. Bloom, for example, notes that one can't humiliate people without acknowledging that they can feel shame and therefore must be human. Smith argues that tormentors often perceive their victims as simultaneously human and subhuman, and posits that framing the 'other' this way creates an "unsettling feeling of uncanniness" around them (431) — a feeling that can itself move people toward violence.

- **Bandura, Albert, Bill Underwood, and Michael E. Fromson. 1975. "Disinhibition of Aggression through Diffusion of Responsibility and Dehumanization of Victims."** ***Journal of Research in Personality.*** **9(4), 253-269.** **https://doi.org/10.1016/0092-6566(75)90001-X**

  In a lab experiment, the authors studied the impact of dehumanizing and humanizing language and personal vs. group responsibility on punitive behavior. Individual participants (brought into the study space in a small group) were told to administer shocks at a strength of their choosing to another group of participants called 'decision makers' (this group did not actually exist), as punishment for answering questions incorrectly. When the facilitators used dehumanizing language to describe the fictional decision makers, participants administered stronger shocks than when the group was described using humanizing language or not described at all. Participants also administered stronger shocks if they were told that their choice of shock level would be averaged with other study participants in their small group than when they were told that they were individually responsible for the shock received by one member of the decision-making team. For both types of responsibility, the

levels of shock were highest for dehumanized target groups and lowest for those labeled with humanizing language. Lastly, the study tested the impact of the effectiveness of punitive treatment on the willingness of participants to use it. When initial shocks seemed to produce improved answers from the decision makers, participants gradually increased the shock level throughout the experiment for the dehumanized and neutral group, while maintaining a low level of shock with the humanized group. When the shocks seemed not to be 'effective' however, participants quickly escalated the level of shock for dehumanized decision makers, while letting it remain moderate for the neutral group and decreasing to a low level for the group labeled with humanizing language.

- **Beyond Conflict. 2019. *Decoding Dehumanization: Policy Brief for Policymakers and Practitioners.***
  https://beyondconflictint.org/wp-content/uploads/2020/06/Decoding-Dehumanization-Policy-Brief-2019.pdf

  This report contains sections on (1) why we should care about dehumanization, (2) the science of dehumanization, (3) the connection between dehumanization and atrocities, and (4) suggestions for countering dehumanization. The report includes an extensive bibliography of literature on dehumanization in general, which is a useful resource for readers.

- **Bloom, Paul. 2017. "The Root of All Cruelty?" *The New Yorker.***
  https://www.newyorker.com/magazine/2017/11/27/the-root-of-all-cruelty

  Bloom argues here that dehumanization may not be a prerequisite for violence and atrocities at all. On the contrary, he points out that human cruelty often takes the form of humiliating other people, which wouldn't seem possible or satisfying to the tormentors if they didn't perceive their victims as human. As Bloom puts it, "The sadism of treating human beings like vermin lies precisely in the recognition that they are not." (See also David Livingstone Smith's 2016 article "Paradoxes of Dehumanization" for more on this).

- **Giner-Sorolla, Roger, Bernhard Leidner, and Emanuele Castano. 2012. "Dehumanization, Demonization, and Morality Shifting: Paths to Moral Certainty in Extremist Violence." *Extremism and the Psychology of Uncertainty,* edited by Michael A. Hogg and Danielle L. Blaylock. Oxford: Wiley-Blackwell.**
  http://umass.edu/bleidner/papers/Giner-Sorolla_Leidner_Castano_2011.pdf

  This article reviews the literature related to extremists' moral justifications of their actions. It synthesizes the psychological literature on this topic, arguing that extremists rationalize killing in three primary ways. They (1) dehumanize members of the out-group, or (2) demonize them, creating a 'moral mandate' to remove the threat

posed by the out-group (10). The third justification comes through what the authors call 'morality shifting': diverting one's own attention away from the violence and toward in-group loyalty (13).

- **Haslam, Nick. 2006. "Dehumanization: An integrative review." *Personality and social psychology review* 10(3), 252-264.**
  https://doi.org/10.1207/s15327957pspr1003_4

  This seminal article offers a new model of dehumanization — one that divides dehumanization into two discrete categories, based on two separate notions of humanness. Humanness, the author argues, can be characterized either by 'uniquely human traits' (which define the boundary between humans and animals) or by characteristics that are seen as central to human nature (which may be shared by some animals). Haslam calls speech portraying people as lacking uniquely human traits 'animalistic dehumanization,' and speech that questions someone's human nature 'mechanistic dehumanization.' He notes that these forms of dehumanization have distinct features (258); therefore identifying the division can help researchers better understand how they function in relation to human behavior. The article also provides a comprehensive literature review examining how various fields (including medicine, psychology, and disability and technology studies) engage with dehumanization.

- **Over, Harriet. 2021. "Seven Challenges for the Dehumanization Hypothesis."** ***Perspectives on Psychological Science*****. 16(1), 3-13.**
  https://doi.org/10.1177/1745691620902133

  Over questions the claim that dehumanization lowers the barriers to collective violence. Using behavioral and cognitive data, as well as historical evidence, she offers seven challenges to the generally accepted belief that dehumanization of an out-group makes committing violence against them easier, including that groups often compare their own members to nonhumans - sometimes even the same ones to which they are comparing out-group members. She also noted that out-group members are often described as having qualities that are eminently human but anti-social, "such as jealousy, spite, and cunning" (4), and being ascribed these traits increases the chance of violence against those people.

- **Neilsen, Rhiannon S. (2015) "'Toxification' as a more precise early warning sign for genocide than dehumanization? An emerging research agenda," Genocide Studies and Prevention: An International Journal: Vol. 9: Iss. 1: 83-95. DOI: http:/.doi.org/10.5038/1911-9933.9.1.1277**

  If dehumanization makes perpetrators feel that committing genocide is allowable, toxification, Neilsen argues, makes the extermination of other groups seem necessary. Toxification is "the cognitive perception of victims as malignant and carcinogenic pests that must be purged for the survival of the perpetrator, and/or the perpetrators' ideal society" (83). Using the Holocaust and the genocides in Rwanda, Armenia, and Cambodia as examples, Neilsen also offers further definitional distinction by describing two possible strains of toxification: one in which the target group is described as toxic to an ideal (such as a nation or a culture), and one in which the "perpetrators become convinced that the victims will, without fail and given the chance, murder the perpetrators" (87).

- **Smith, David Livingstone. 2016. "Paradoxes of Dehumanization." *Social Theory and Practice* 42(4), 416-443. https://doi.org/10.5840/soctheorpract201642222**

  In this article, Smith presents and examines a paradox: that during mass violence, perpetrators often view their victims as simultaneously human and subhuman. They use dehumanizing rhetoric, but also behave in ways that acknowledge their victims' humanness (for example by seeking to humiliate them). Instead of seeing this paradox as refuting a relationship between dehumanization and violence, Smith argues that the people seen as simultaneously human and less than human (subhumans in human form) produce an "unsettling feeling of uncanniness" (431). They are categorically distinct (neither fully human nor inhuman), which can make others feel that they "pose a threat to any social order" (430) and can be used to justify violence against them. Smith illustrates his concept by citing the European colonists who justified their violence against enslaved Africans by saying that they *resembled* men, but were "indeed no men" (421).

- **Rai, Tage S., Piercarlo Valdesolo, and Jesse Graham. 2017. "Dehumanization Increases Instrumental Violence, but not Moral Violence." *Proceedings of the National Academy of Sciences* 114(32), 8511-8516. https://doi.org/10.1073/pnas.1705238114**

  The authors conducted five psychological experiments on the relationship between dehumanizing speech and support for violence. They found that dehumanization "predicts, causes, and is caused by" (8514) instrumental violence (violence committed in pursuit of other goals), but is not related to support for moral violence (where the violence itself is the goal, because the victims 'deserve it'). In other words, if someone

has preexisting dehumanizing thoughts about another person, they are more likely to condone instrumental violence against them. They are also more likely to be moved to support violence against a person if they hear dehumanizing speech about that person and, conversely, more likely to think of a person as less than human if they imagine themselves perpetrating instrumental violence against that person. This is not the case for moral violence. The authors further argue that victims of morally justified violence are thought of as human, because one must be human in order to fully experience moral punishment.

4. ONLINE SPEECH AND HATE CRIMES

In the past few years, several efforts have emerged that directly search for a relationship between online speech and offline attacks, but their findings come with caveats. First, drawing a causal link between speech and action is nearly always difficult, since a variety of factors drive action. Scholars have also pointed out that the relationship between online speech and hate crimes may be reversed. For example, Olteanu et al. (2018) found an increase in online hateful speech *after* incidents of extremist violence.

Other challenges exist as well, largely related to data. Data about hateful speech online can be very difficult to obtain. Accurate hate crime data can be similarly elusive. For example, in the United States, data on hate crimes is collected by the Federal Bureau of Investigation (FBI), a federal agency, but reporting varies by jurisdiction. Phoenix, AZ has a very high rate of reported hate crimes, for example. This may be because more crimes are committed there, but it may also be because Phoenix has a special hate crime police unit, which means these crimes are likely reported and tagged as hate crimes more often than in other locales (Relia et al., 2019). Such flaws in the data make it difficult to conduct a robust analysis.

- **Müller, Karsten, and Carlo Schwarz. 2020. "Fanning the flames of hate: Social media and hate crime."** ***Journal of the European Economic Association.*** **https://doi.org/10.1093/jeea/jvaa045**

  Müller and Schwarz (2020) conducted a study in Germany investigating whether anti-refugee sentiment on Facebook predicted offline hate crimes. The authors found that anti-refugee hate crimes increased disproportionately in municipalities where a large amount of anti-refugee hateful content was shared through the far-right AfD party's Facebook page. They also used data on internet and Facebook outages (user-reported and news-reported) to support their claim that "Facebook disruptions reduce local hate crimes, particularly in areas with many AfD users" (3). Taken together, the authors claim that their data "suggests that social media has not only become a fertile soil for the spread of hateful ideas but also motivates real-life

action" (34). Their findings have been called into question by some,[4] however, as they made a number of assumptions and used questionable proxies. For example, they examined anti-refugee speech only on the AfD's Facebook page and used posts that included the word "*Fluchtling*" (refugee) as a proxy for anti-refugee speech on the page, a decision which could have resulted in over counting neutral or positive messages about refugees, while missing others that did not contain the word.

- **Newman, Benjamin, Jennifer L. Merolla, Sono Shah, Danielle Lemi, Loren Collingwood, and S. Karthick Ramakrishnan. 2020. "The Trump Effect."** ***British Journal of Political Science.*** **1-22. https://doi.org/10.1017/S0007123419000590**

  Since 2015, journalists and scholars have opined about the so-called 'Trump effect' – the notion that Donald Trump's hateful rhetoric has emboldend racists and is related to increased prejudice, discrimination, and hate crimes against minority groups in the United States. In 2016, the authors of this study conducted a two-part online survey experiment to test this notion. They found that for those who already held prejudiced beliefs, exposure to Trump's inflammatory statements (or occasionally even just his name) caused them to "be more likely to perceive engagement in prejudiced behavior as socially acceptable" (2). The effects were particularly strong among those who read statements in which other elites in the political system seemed to condone Trump's remarks.

- **Olteanu, Alexandra, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. "The Effect of Extremist Violence on Hateful Speech Online."** ***Twelfth International AAAI Conference on Web and Social Media*****.**

  The authors conducted a quantitative assessment of the relationship between offline "attacks involving Arabs and Muslims as perpetrators or victims, occurring in Western countries" (221) and online hateful speech on Twitter and Reddit. Based on a collection of more than 150 million messages that the authors identified as being "related to hate and counter-hate speech" (223) they found that incidents of extremist violence that were perpetrated by Muslims (but not Islamophobic attacks) led to increases in online hateful speech (especially explicit calls for violence), as well as counter-hate messages, in the week after an attack.

[4] See Masnick's (2018) critique on an earlier version of the paper. In the peer-reviewed version, which we include in this review, the authors are no longer using "Facebook likes of the 'Nutella Germany' page" as a proxy for German internet usage.
https://www.techdirt.com/articles/20180823/00122840491/dubious-studies-easy-headlines-no-new-report-does-not-clearly-show-facebook-leads-to-hate-crimes.shtml

- **Relia, Kunal, Zhengyi Li, Stephanie H. Cook, and Rumi Chunara. 2019. "Race, Ethnicity and National Origin-Based Discrimination in Social Media and Hate Crimes across 100 US Cities." In *Proceedings of the International AAAI Conference on Web and Social Media*, 13(1), 417-427. https://arxiv.org/abs/1902.00119**

  The authors created a sample of 532 million tweets posted from January 1, 2011 to December 31, 2016 in 100 American cities (identified by the 'place' attribute). They found a significant correlation between the number of hate crimes reported in each of those cities and the proportion of tweets from that city containing race, ethnicity and national-origin based discriminatory language. The authors were careful not to construe causality in either direction of this correlation. To identify tweets containing this language and reports of this kind of discrimination, they used a "tweet processing pipeline" (including an active learning classification, a "spatially-diverse training dataset" to account for regional differences, and a pronoun checker, which identified first-person pronouns to filter reports of discrimination).

- **Williams, Matthew L., Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *The British Journal of Criminology* 60(1), 93-117. https://doi.org/10.1093/bjc/azz049**

  Using police-recorded data on racially and religiously aggravated crimes from August 2013 to August 2014 (N = 6,572) and tweets from London during the same period (N = 21.7 million), the authors searched for a relationship between hate crimes and hateful speech on Twitter (used as an indicator of extreme polarization). Within the full set of tweets, a supervised machine learning classifier identified 294,361 tweets as 'hateful.' The authors note that their "models indicate a strong link between hateful Twitter posts and offline racially and religiously aggravated crimes in London," but stop short of stating a causal direction: they "cannot say if online hate speech precedes rather than follows offline hate crime" (107). The authors further express their doubt that hateful online speech causes offline hate crimes on its own, emphasizing instead the importance of factors like neighborhood ethnic make-up and local employment levels in determining the relationship between online speech and hate crimes (112).

5. CONTEXT-SPECIFIC REPORTS

In this section, we examine studies on the connections between the spread of dangerous speech through media (social media and in some cases radio) in specific contexts and episodes of intergroup violence. One cannot determine the impact of speech in the abstract. It is necessary to understand the social, historical, and cultural context in which that speech was made or disseminated to analyze its potential impact. A message could be benign in

one context and highly inflammatory in another. We focus on nine locales: Argentina, Côte d'Ivoire, Germany, Indonesia, Kenya, Myanmar, Sri Lanka, Rwanda, and the former Yugoslavia. In some cases, such as the 1994 Rwandan genocide and the 2007 post-election violence in Kenya, it has become commonly understood that messages broadcast to the public helped catalyze intergroup violence, though the literature does not provide hard evidence.

In many of these articles, the authors provide empirical evidence that *suggests* a causal relationship. For example, Sarjoon et al. (2016) note the connection between anti-Muslim speech and violence in Sri Lanka between 2009 and 2016. In one example, they describe a speech made by Gnanasara Thero, the Secretary General of Bodu Bala Sena, a Sinhalese Buddhist nationalist organization, in which he told the crowd they needed to fight against Muslims, and he threatened to destroy Muslim-owned businesses. Hours later, in the same city where he delivered the speech, anti-Muslim rioters destroyed over 100 Muslim-owned businesses and killed four people.

Other scholars have statistically analyzed the link between speech and violence. In Rwanda, for example, David Yanagizawa-Drott (2014) tested the relationship between the broadcast range of a Rwandan radio station notorious for calling for violence, and the number of individuals eventually prosecuted for committing genocide in villages. He found that exposure to broadcasts demanding the extermination of Tutsis increased Hutu participation in the killing, both among those who lived in the broadcast range and among those living in neighboring villages. Maja Adena and her colleagues (2015) also used data on radio subscription rates and strength of signal to support a connection between those broadcasts and public support for the Nazis and their policies against the Jews in the 1920s and 1930s.

Several studies in this section note the temptation to document a large amount of dangerous speech followed by violence, and then simply construe a causal link. This is especially common in the Rwandan case, in part because scholars and judges paid tremendous attention to the radio station RTLM, too easily assuming that it caused the genocide. Claims of causation often also assume an undifferentiated impact of speech on audiences — something strongly challenged by the findings of Li (2004), Mironko (2007), and Straus (2007).

Despite our caution regarding causation, below we list many articles that assert it, and provide abundant examples of the kinds of dangerous speech that circulate in the months or years before mass violence, as well as detailed information on how these messages spread through society.

Argentina

In 1976, the Argentine military took power in a coup and installed a junta that would rule the country until 1983. During this time, the Argentine military kidnapped, tortured, and killed

political dissidents and people accused of being associated with socialism (many of whom were students). The junta used language as a tool to exert power and elicit complicity among Argentines. As many as 30,000 people were killed or 'disappeared' during this period.

- **Feitlowitz, Marguerite. 1998. *A Lexicon of Terror: Argentina and the Legacies of Torture*. Oxford University Press.**

  In this book, Feitlowitz offers a detailed account of how the Argentine military regime used language to conceal, reframe, and elicit complicity from citizens, in the torture and killings they perpetrated during the Dirty War. Among many examples of euphemisms used by military leaders, torture was referred to as '*tratamiento*' (treatment), and citizens were told that the military regime would lead them on "a quest for the common good, for the full recovery of *el ser nacional,*" ("the collective national essence, soul, or consciousness") (21). The military regime also used language to dehumanize and create a sense of fear of leftists in the country, for instance describing them as "armed bands of subversive criminals" (50) in an effort to gain support among the civilian population for their actions.

Côte d'Ivoire

The articles in this section focus on dangerous speech that circulated in Côte d'Ivoire during the 2010-2011 post-election crisis. Then-President Laurent Gbagbo refused to concede the November 2010 election to Alassane Ouattara, which sparked a surge of dangerous speech (generally targeting opposition supporters), killings, and displacement. At least 3,000 were killed[5] and over 450,000 are estimated to have fled the country by March of 2011.

- **Human Rights Watch. 2011. "Côte d'Ivoire: crimes against humanity by Gbagbo forces." *HRW*, New York, 15 March.**
  https://www.hrw.org/news/2011/03/15/cote-divoire-crimes-against-humanity-gbagbo-forces
  This report was based on 100 interviews from early 2011, with victims of, and witnesses to, attacks and killings in Côte d'Ivoire. The violence followed the announcement of election results in December 2010, which in turn followed a long campaign of dangerous speech including a statement on state television by Charles Blé Goudé, then-President Gbagbo's Youth Minister, calling on Gbagbo supporters to set up roadblocks and "denounce every foreigner who enters." (Here 'foreigner' is a translation of a term used to describe ethnic groups from the northern part of Côte d'Ivoire or other West African nationals living in the country). Multiple victims reported hearing their perpetrators refer to Blé Goudé's 'order' while carrying out the attacks.

[5] Wormington, Jim. 2018. "Côte d'Ivoire's Forgotten Victims." *Human Rights Watch.* https://www.hrw.org/news/2018/02/23/cote-divoires-forgotten-victims#

- **United Nations Human Rights Council. 2011. "Report of the High Commissioner for Human Rights on the situation of human rights in Côte d'Ivoire." Geneva, 25 February. https://www.refworld.org/pdfid/4d8b3e162.pdf**
  This report describes extreme political rhetoric that surged prior to the 2011 presidential election in Côte d'Ivoire, and the violence that broke out after the results were announced. Section K focuses on speech and violence, discussing the "role of the media in inciting hatred and violence" (11). Radiodiffusion Télévision Ivoirienne (RTI), the publicly owned and state-controlled radio and television networks of Côte d'Ivoire, was a prominent offender. As the report notes, RTI aired messages telling Gbagbo supporters to 'resist the enemy,' running an "intensive and systematic campaign to incite intolerance and hatred against the United Nations, the African Union, the Economic Community of West African States (ECOWAS), the facilitator of the Ivorian dialogue and non-LMP leaders and supporters" (11).

Germany

From 1933, when Hitler established the "Reich Ministry of Public Enlightenment and Propaganda," to the Nazis' surrender in 1945, the Nazi party used an elaborate propaganda campaign to win the support of German citizens for its policies, which included a genocide in which the regime murdered 6 million Jewish people, among others who were portrayed as enemies of the German people.. Dangerous speech targeting Jews circulated on the radio, in films, and in print sources such as the Julius Streicher's notorious anti-Semitic newspaper, Der Stürmer.[6]

- **Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *The Quarterly Journal of Economics* 130(4), 1885-1939. https://doi.org/10.1093/qje/qjv030**
  The authors used quantitative analysis to study the impact of radio on support for the Nazis and for discrimination against Jews in the 1920s and 1930s. Using historical data about radio content, along with data on radio exposure (calculated using local radio subscription rates and strength of radio signal), the authors found that exposure to the radio increased Nazi party membership and support only after the regime took control of the media. Also "exposure to Nazi radio propaganda in its full strength increased the number of Jews deported to concentration camps and the number of anti-Semitic open letters" (1889), but the effectiveness of propaganda largely depended on the predisposition of the majority of the audience. In effect, "mass media does help dictators gain popular support and persuade people about the virtues of their most atrocious policies, but only if the majority does not disagree with

[6] See "Nazi Propaganda." *Holocaust Encyclopedia.* United States Holocaust Memorial Museum. https://encyclopedia.ushmm.org/content/en/article/nazi-propaganda

the propaganda message a priori" (1890).

- **Bytwerk, Randall. 2001. *Julius Streicher: Nazi Editor of the Notorious Anti-Semitic Newspaper Der Sturmer*. Cooper Square Press.**
  From 1923 to the end of WWII, Julius Streicher published *Der Stürmer,* a weekly German newspaper known for its violent anti-Semitism. Bytwerk's biography of Streicher examines the propaganda techniques Streicher used in his paper, such as publishing stories about, and letters from, people who claimed to have been cheated by Jewish businessmen (123-124), alleging that Jews posed a sexual threat to German women, and publishing offensive cartoons and doctored photographs of Jewish people. Chapter 9 of the book contains information about the impact of *Der Stürmer*, including several examples of letters written from readers who described how the newspaper taught them to see that Jewish people are "the deadly enemy of national life, (173).

- **Herf, Jeffrey. 2006. *The Jewish Enemy*: Nazi Propaganda during World War II and the Holocaust. Cambridge, MA: Harvard University Press.**
  Herf examines how antisemitism changed from a rationale for persecution to one for mass murder. He describes how the Nazis used propaganda to advance a narrative of German victimization at the hands of an international conspiracy led by Jews. The book includes many examples of such propaganda as well as discussion of how the Nazis used it to legitimize the war and to build support for genocide.

- **Klemperer, Victor. 2013. *The Language of the Third Reich.* London: Bloomsbury Academic.**
  Victor Klemperer kept detailed diaries throughout his life as a Jewish-descended German philologist and professor of Romance languages. The diaries, three volumes of which have been published in English, are valuable sources on the Nazi period, especially regarding the daily persecution and humiliation of Jews.[7] Klemperer also wrote this book on Nazi language based on notes in his diaries describing Nazi vocabulary and style of speech. As an eyewitness, Klemperer included examples taken from his conversations as well as books, newspaper articles, and radio broadcasts to illustrate how the language of the Third Reich contributed to its culture.

Indonesia

Both articles in this section focus on what is referred to as the 'Maluku sectarian conflict,' a period of ethno-religious violence between 1999 and 2002 on the islands making up the Maluku

[7] See Klemperer, Victor. 1999. *I Will Bear Witness (Vol. 1) 1933-1941*. Random House Digital, Inc.. and Klemperer, Victor. 2001. *I Will Bear Witness (Vol. 2) 1942-1945: A Diary of the Nazi Years* . Modern Library,

Archipelago in Indonesia. As the authors in this section note, rumors and conspiracy theories fueled violence between Muslims and Christians on the islands.

- **Bubandt, Nils. 2008. "Rumors, pamphlets, and the politics of paranoia in Indonesia." *The Journal of Asian Studies*. 67(3), 789-817. https://doi.org/10.1017/S0021911808001162**

  Bubandt examined the events surrounding an anonymous leaflet that circulated in North Maluku, Indonesia during a year-long ethno-religious conflict in 1999. Here he argues that anonymous leaflets were ideal for stoking conflict. As leaflets or pamphlets pass through communities, competing groups can easily interpret them in a way that conforms with their own preexisting beliefs. And in comparison to verbal rumors, the fact that they are written gives leaflets and pamphlets a 'testimonial authority,' (793) and encourages people to see them as tangible support for their own narrative of the conflict. For example, Muslims interpreted a leaflet as proof of a campaign to 'Christianize' the province, while Christians who saw the same leaflet took it as evidence that Muslim elites were inciting conflict — by forging the leaflet. Although rumors about a Christian campaign of 'religious cleansing' and a Muslim plot to incite violence had been circulating in the province for months, the author argues that the leaflet reinforced them and thus escalated the conflict (812).

- **Wilson, Chris. 2011. "Provocation or excuse?: Process-tracing the Impact of Elite Propaganda in a Violent Conflict in Indonesia." *Nationalism and Ethnic Politics*. 17(4), 339—360. https://doi.org/10.1080/13537113.2011.622629**

  Wilson tested several hypotheses on "why and how elite incitement can sometimes lead to widespread communal violence" (356). He used a method called 'process-tracing,' sequencing empirical evidence within a case study to theorize about causal mechanisms, paying attention to "the decision-making processes of actors influential to important outcomes and, where possible, of those groups of individuals involved in important events" (342-343). Wilson argues that different causal factors may drive different stages of conflict. For example, he notes that those who begin riots are often closely connected — through kinship, patronage, and/or politics —to elites (and are more influenced by the elites' propaganda).. As such conflicts grow, other actors join and fight for different reasons. Wilson notes that in this escalation phase, the 'tit-for-tat' violence of the initial conflict "bring[s] about the construction of more antagonistic group identities, making it rational to fear the other group," and inspiring moderates to become involved (342).

Kenya

Many human rights activists, media monitors, and journalists believe that broadcasts on vernacular radio stations incited hatred and violence in the days and weeks surrounding the 2007 Kenyan general election. Some have compared Kenyan radio during this time with RTLM in Rwanda during the 1994 genocide. In 2011, the International Criminal Court indicted Kenyan radio host Joshua Arap Sang for allegedly contributing to the commission of crimes against humanity through his broadcasts, but five years later the Court vacated the charges for insufficient evidence.

- **Ismail, Jamal Abdi, and James Deane. 2008. "The 2007 General Election in Kenya and Its Aftermath: The Role of Local Language Media."** ***The International Journal of Press/Politics*** **13, 319-327. https://doi.org/10.1177/1940161208319510**

  Based on data from 20 semi-structured interviews with "senior media figures in Kenya," the authors describe the role they believe local language radio stations played in the country's 2007-2008 post-election violence. For example, they argue that some local language radio stations were complicit in the violence by allowing callers to spread ethnic hatred on the air.

- **Mahoney, Chris, Eduardo Albrecht, and Murat Sensoy. 2019. "The Relationship Between Influential Actors' Language and Violence: A Kenyan Case Study Using Artificial Intelligence." The LSE-Oxford Commission on State Fragility, Growth and Development.**
  https://www.theigc.org/wp-content/uploads/2019/02/Language-and-violence-in-Kenya_Final.pdf

  The researchers used third-party software tools to gather and analyze 6,100 tweets by 30 of the most influential Kenyans between January 2012 and December 2017. They assigned sentiment scores to tweets and used those to assemble a model to predict increases and decreases in fatalities as reported by the Armed Conflict Location and Event Data Project (ACLED). By detecting variations in the rhetoric of influential actors, the model could predict "both increases and decreases in average fatalities" within 50 to 150 days, with almost 85% accuracy, according to the paper (14).

- **Odera, Edna Iplei. 2015. "Radio and Hate Speech: A Comparative Study of Kenya (2007 PEV) and the 1994 Rwanda Genocide." PhD diss.,University of Nairobi.**
  http://hdl.handle.net/11295/93846

  Odera used secondary data to describe how people can use language not only to incite acts of violence, but also to frame offline attacks in ways that lead to more

violence — whether retributory or not. The paper contains a detailed description of the post-election violence in Kenya in 2007-2008, an overview of the media landscape, and examples of dangerous speech from that time. Odera also compares the Kenyan and the Rwandan media's roles in inciting violence.

- **Somerville, Keith. 2011. "Violence, Hate speech and Inflammatory Broadcasting in Kenya: The Problems of Definition and Identification."** ***Ecquid Novi: African Journalism Studies*** **32(1), 82-101. https://doi.org/10.1080/02560054.2011.545568**

  Drawing on interviews, reports of the Kenyan post-election violence of 2007, and a limited number of radio transcripts, Somerville investigated the claim that Kenyan radio stations incited violence much the way that the Rwandan radio station RTLM had in 1993-4, finding evidence that Kenyan 'vernacular' or local radio stations did "periodically broadcast hate speech about perceived opponents from other communities, at times appeared to condone or even incite violence or the expulsion of people from particular areas, and demonstrated considerable partisanship" (97). The difference, he argues, is that there is no evidence of a coordinated campaign in Kenya to incite violence, as there was in Rwanda. Somerville also points out that it's difficult to study vernacular local radio stations retrospectively, even in communities where they seem to have major impact, since they generally don't record their broadcasts.

- **Waki Commission. 2008. Report of the Commission of Inquiry into Post-Election Violence Nairobi, Kenya. https://digitalcommons.law.seattleu.edu/cgi/viewcontent.cgi?article=1004&context=tjrc-gov**

  In February 2008, the Kenyan government established the Commission of Inquiry on Post-Election Violence (known as the 'Waki Commission' after its chairman Philip Waki, a Kenyan judge). The Commission's report notes examples of 'threatening terms' that "were routinely used against Kikuyu (one of the main ethnic groups in Kenya): madoadoa (spots), maharagwe (bean), bunyot (enemy), [and] sangara (wild grass)" (63). The report also documents specific uses of these terms, such as when a politician told members of the Maasai ethnic group to uproot Kikuyus as if they were weeds (13). Chapter 8 of the report describes how the media contributed to violence. It notes that "many recalled with horror, fear, and disgust the negative and inflammatory role of vernacular radio stations" in creating a "climate of hate" in Kenya (295).

Myanmar

The articles in this section explore the ways in which language inspired and amplified violence against Muslims in Myanmar. Dangerous speech targeting Muslims — especially on social

media — has received attention from scholars and human rights practitioners, and in 2018, United Nations investigators wrote that social media played a 'determining role' in the violence committed against Rohingya Muslims.[8]

- **Fink, Christina. 2018. "Dangerous Speech, Anti-Muslim Violence, and Facebook in Myanmar." *Journal of International Affairs* 71(1.5), 43-52. https://jia.sipa.columbia.edu/dangerous-speech-anti-muslim-violence-and-facebook-myanmar**

  Starting with an overview of anti-Muslim sentiment and action in Myanmar between 2012 and 2018, this article also describes how anti-Muslim content spread on Facebook. Fink highlights a 2014 case, arguing that "a Buddhist monk's Facebook post appeared directly connected" (46) to violence: the prominent monk Ashin Wirathu asserted on Facebook that a Muslim man who owned a teashop had raped one of his female Buddhist employees. Wirathu added that he "had called the proprietor to assert he would face justice" (46). Shortly after that, a Muslim man and a Buddhist man were killed in a riot in the town where the alleged rape took place. A police investigation later found that the rape claim was false.

- **Lee, Ronan. 2019. "Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis." *International Journal of Communication* 13: 3203-3224. https://ijoc.org/index.php/ijoc/article/view/10123/2718**

  Lee notes that social media companies are often blamed for being conduits of dangerous speech and violence, and governments are often touted as a solution. However government speech can be just as dangerous, Lee argues. He analyzed the reporting of *Global New Light of Myanmar*, the state-run newspaper, on events in Rakhine state, where Myanmar's Rohingya Muslim population has long been concentrated. According to Lee, the newspaper focuses either on the government's economic victories in Rakhine or on security threats ostensibly posed by 'extremist terrorists' living there. This, he argues, "shows how official media contributed to a political environment where anti-Rohingya speech was made acceptable and where rights abuses against the group were excused" (3203).

- **Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar. 2018. U.N. Doc. A/HRC/39/CRP.2 https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf**

[8] Miles, Tom. 2018. "U.N. Investigators Cite Facebook Role in Myanmar Crisis." *Reuters. https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN*

In 2017 and 2018, members of the Independent International Fact-Finding Mission on Myanmar conducted 875 interviews with victims and eyewitnesses of violence that Myanmar's armed forces (known as the 'Tatmadaw') committed against civilians — primarily Rohingya — living in Myanmar. Their 441-page report contains a wealth of information about the violence, and of particular interest to this review is section VI.B, titled "the Issue of Hate Speech" (320-343). In it, the authors discuss methods of dissemination of hateful speech, prominent speakers (e.g. ultranationalist Buddhist monks), a list and description of "derogatory terms used to refer to the Rohingya, or to Muslims in general" (322), and a collection of examples of hateful speech. A subsection describes "coordinated hate campaigns and possible links to outbreaks of violence" (330), noting that although it is "difficult to establish" a direct connection between the campaigns and outbreaks of violence, the "Mission received information suggesting that the linkage between offline and online hate speech and real world acts of discrimination and violence is more than circumstantial" (331). There is also a useful section describing how social media (primarily Facebook) disseminates hateful speech in Myanmar (339-343).

- **Schissler, Matt, Matthew J. Walton, and Phyu Phyu Thi. 2017. "Reconciling contradictions: Buddhist-Muslim violence, narrative making and memory in Myanmar."** ***Journal of Contemporary Asia*** **47(3), 376-395.** https://doi.org/10.1080/00472336.2017.1290818

  From 68 interviews conducted in 2015 in six cities in Myanmar, this team of researchers expert on Myanmar illustrate how Burmese people talk about and conceptualize inter-group violence in their everyday lives, and how discourse in Myanmar constructs Muslims as a "fearsome Other" (376). Many of the interviewees saw Muslims as an existential religious and racial danger to Myanmar, as well as a personal threat to their families and communities. The authors argue that these perceptions are easily used to justify "virtuous self-defense" (385). The interviews also contained some counternarratives. For example, some interviewees noted both "contemporary tension and memories of co-existence" with Muslims (391). This is important, the authors argue, as working for peace becomes even more difficult "if the two communities are understood to have always been at odds" (390).

- **Wade, Francis. 2017. *Myanmar's Enemy Within: Buddhist Violence and the Making of a Muslim 'Other'*. London: Zed Books Ltd.**

  Wade traces the roots of Buddhist violence against Rohingya Muslims in Myanmar through the long process of preparing for intergroup violence, illustrating how policy decisions, rumors, and propaganda can fracture a nation. Together, these factors created a context of fear in which one group finds it acceptable, and is even compelled, to commit violence against another. Wade also describes how authoritative speakers, such as monks and politicians, garnered support for attacks on Muslims.

- **Win, Ye Myint. 2015. "Rise of Anti-Muslim Hate Speech Shortly before Outbreaks of the Mass Violence against Muslims in Myanmar." *ICIRD 2015 Mahidol Online Proceedings*.**
  https://www.burmalibrary.org/docs21/Religion/Ye-Myint-Win%28Nickey%20Diamond%29-2015-The_Rise_of_Anti-Muslim_Hate_Speech_Shortly_Before_the_Outbreaks_of_Mass_Violence_Against_Muslims_in_Myanmar-en.pdf

  This master's thesis draws from interviews and pamphlets, DVDs, CDs, and the texts of Buddhist sermons, to argue that anti-Muslim rhetoric has always escalated just before large-scale violence against Muslims in Myanmar. Win describes several cases in which anti-Muslim speech preceded mass violence. The thesis is an example-filled narrative of anti-Muslim speech in Myanmar between 1997 and 2013, a period that has received scant attention in scholarship on speech and violence in the country.

Rwanda

Perhaps more than in any other case, scholars and practitioners have frequently noted a connection between speech and intergroup violence during the 1994 Rwandan genocide, in which Hutus murdered around 800,000 Tutsis. The radio station Radio Télévision Libre des Mille Collines (RTLM) in particular is often identified as the source of violence-inciting speech. In 1997, the United Nations International Criminal Tribunal for Rwanda (ICTR), indicted three Rwandans for 'inciting genocide': Ferdinand Nahimana and Jean-Bosco Barayagwiza, who co-founded RTLM, and Hassan Ngeze, who founded a pro-Hutu newspaper called 'Kangura.' All three were convicted. As can be seen below, however, there isn't consensus within the scholarship about whether and how speech contributed to the genocide.

- **Des Forges, Alison. 2007. "Call to Genocide: Radio in Rwanda, 1994." In *Media and the Rwanda genocide*. Edited by Allan Thompson. IDRC, Ottawa, ON, CA. p41-54.**

  Drawing on secondary sources as well as transcripts from the radio station RTLM, Des Forges, a longtime expert on Rwanda who covered the country for Human

Rights Watch, laid out how Hutu political actors sought to use radio to incite violence in Rwanda. The chapter is packed with specific details to argue that radio had an impact on the Rwandan genocide in a variety of ways (including directing killings and spreading false rumors that Tutsis were planning to exterminate Hutus).

- **Human Rights Watch and the International Federation for Human Rights. 1999. *Leave None to Tell the Story: Genocide in Rwanda.* New York: Human Rights Watch. https://hrw.org/sites/default/files/media_2020/12/rwanda-leave-none-to-tell-the-story.pdf**

  This 789-page report was a major effort in the wake of the genocide, by a team of researchers led by Alison Des Forges and Eric Gillet. It is an authoritative source on many aspects of the 1994 genocide, and we include it here for its chapter on propaganda and practice, which includes many examples of dangerous speech that circulated before and during the genocide.

- **Kellow, Christine L., and H. Leslie Steeves. 1998. "The Role of Radio in the Rwandan Genocide." *Journal of Communication*. 48(3), 107-128. https://doi.org/10.1111/j.1460-2466.1998.tb02762.x**

  The authors used a qualitative textual analysis of translated radio transcripts to assess the role of radio in the Rwandan genocide. Rwandans depended heavily on radio for news and information before and during the genocide, they noted, because "interpersonal networks were insufficient for [receiving] political information, [and] illiteracy was widespread (125). Radio messages that capitalized on Hutus' growing fear of Tutsis were therefore very persuasive. This study is significantly limited in that the researchers did not interview any Rwandans. They used only translated radio transcripts and secondary sources.

- **La Mort, Justin. 2009. "The Soundtrack to Genocide: Using Incitement to Genocide in the Bikindi Trial to Protect Free Speech and Uphold the Promise of Never Again." *Interdisciplinary Journal of Human Rights Law*. 4(1), 43-66. https://ssrn.com/abstract=1523568**

  Primarily a legal history and analysis of the international law crime of "direct and public incitement to commit genocide," the article uses the case of Simon Bikindi, the Rwandan singer and radio personality who was convicted of incitement to genocide by the International Criminal Tribunal for Rwanda (ICTR) in 2008, to illustrate the forms of incitement considered by the Tribunal. LaMort gives examples of Bikindi allegedly calling his fellow Hutus to violence, using coded language, and asserts that Bikindi's songs "poisoned the hearts and minds of his listeners" (45), although none of the songs mentioned killing, nor called for it. The author notes that witnesses testified that some Hutus sang Bikindi's songs while they murdered Tutsis.

- **Li, Darryl. 2004. "Echoes of Violence: Considerations on Radio and Genocide in Rwanda," *Journal of Genocide Research* 6(1), 9–28. https://doi.org/10.1080/1462352042000194683**

  As Li notes, most popular accounts of the Rwandan genocide assume that the radio station RTLM incited the genocide. Drawing on interviews that Li conducted after the genocide, he attempts to explain how listeners used the radio station and how it became vital to Hutus during the genocide. Li argues that people were informed, but not controlled, by what they heard on RTLM. It became such an influential source of information, he claims, because it 1) amplified the dominant ideological narratives of the time, 2) used performances to create a relationship with listeners, and 3) used the mundane nature of radio to routinize the violence. Li describes how individuals amplified the messages they heard on the radio, such as gathering friends and telling them about what they had heard, or singing Simon Bikindi's songs as they walked. These interviews advanced the findings of other studies on the topic, such as the articles by Straus (2007) and Yanagizawa-Drott (2014).

- **Mironko, Charles. (2004). "Igitero: Means and Motive in the Rwandan Genocide." *Journal of Genocide Research*, *6*(1), 47-60. https://doi.org/10.1080/1462352042000194700**

  After conducting nearly 100 interviews with confessed genocide perpetrators in Rwandan prisons in 2000, Mironko tried to explain why they killed. He argues that the similarity of language used by the perpetrators to describe their crimes (for example, often talking about killing using hunting terminology) suggests not only the power of government propaganda, but also the group thinking that existed during the genocide. He writes, "while state actions in Rwanda in 1994 may have speeded the process of genocide, people themselves, thinking and acting in mobs, assumed a degree of initiative in the violence" (47).

- **Mironko, Charles. 2007. "The Effect of RTLM's Rhetoric of Ethnic Hatred in Rural Rwanda." *The Media and the Rwanda Genocide*. Edited by Allan Thompson. London: Pluto Press. 125-135.**

  This article also draws on Mironko's interviews with nearly 100 'confessed perpetrators' to demonstrate the difficulty in drawing a clear causal link between speech on the radio and genocide. RTLM did not have the same impact on all Hutus, obviously. Many of Mironko's interviewees said they considered RTLM to be for urban and well-educated people, not for them. The interviews also revealed important nuances in how people received the messages in RTLM broadcasts. For example, although rural Rwandans may have been familiar with the meanings of songs or influential speeches calling for violence, they often heard about them from other people rather than on the radio. Although this does not disprove that the broadcasts

had an impact, it does suggest that the dissemination of dangerous speech was complicated.

- **Schabas, William A. 2000. "Hate speech in Rwanda: The road to genocide." *McGill Law Journal.* 46: 141. http://doi.org/10.4324/9781351157568-8**

  In a legal analysis of the crime of incitement to genocide, Schabas assumes a causal relationship between speech and violence, instead of trying to substantiate it. He notes that several people closely acquainted with the Rwandan genocide also believe in a causal link, including General Romeo Dallaire, chief of the UN Peacekeeping Mission in Rwanda during the genocide, who said that if RTLM had been shut down, "many lives might have been spared" (148).

- **Straus, Scott. 2007. "What is the Relationship between Hate Radio and Violence? Rethinking Rwanda's 'Radio Machete.'" *Politics & Society.* 35(4), 609-637. https://doi.org/10.1177/0032329207308181**

  This article challenges the often-repeated belief that radio broadcasts catalyzed or directly caused the Rwandan genocide. Straus analyzed timing, exposure, and content data and also surveyed 210 confessed perpetrators. His analysis suggests that radio had "marginal and conditional" effects. He argued that "radio alone cannot account for either the onset of most genocidal violence or the participation of most perpetrators" (611), but it did reinforce messages people heard in face-to-face interactions, contribute to an environment where would-be perpetrators felt they had limited choices, and likely convinced a small number of actors to commit violence.

- **Yanagizawa-Drott, David. 2014. "Propaganda and Conflict: Evidence from the Rwandan Genocide." *The Quarterly Journal of Economics*. 129(4), 1947-1994. https://doi.org/10.1093/qje/qju020**

  Yanagizawa-Drott took note of Rwanda's very hilly terrain in designing a study to look for correlation between radio broadcast reception and levels of killing during the genocide. He assumed that villages at the top of hills were able to receive the signal of the notorious station RTLM, while demographically similar villages in adjacent valleys could not get the broadcasts. Constructing pairs of those villages, Yanagizawa-Drott compared participation in genocidal violence, though since actual deaths were not recorded, he used the number of people later prosecuted for genocide as a proxy. He found that radio broadcasts increased participation in the killing, both among those who lived in the broadcast range and among those living in neighboring villages. He argues "approximately 10 percent of overall participation can be attributed to the radio station's broadcasts, and almost one-third of the violence by militias and other armed groups" (29-30).

Sri Lanka

The articles in this section explore the use of dangerous speech in Sri Lanka since the end of the country's civil war (fought from 1983 to 2009). The authors describe dangerous speech that circulated around specific flashpoints: anti-Muslim riots in the communities of Grandpass (2013) and Aluthgama (2014), and following the sentencing of Sgt. Sunil Rathnayake (2015), an army officer who was convicted of murdering eight internally displaced refugees in the Sri Lankan town of Mirusuvil.

- **Samaratunge, Shilpa, and Sanjana Hattotuwa. 2014. *Liking violence: A study of hate speech on Facebook in Sri Lanka*. Centre for Policy Alternatives.** https://cpalanka.org/wp-content/uploads/2014/09/Hate-Speech-Final.pdf

  The authors analyzed posts from 20 Facebook pages of Sri Lankan extremist Buddhist groups (included in an appendix). They collected the posts in the weeks surrounding two anti-Muslim riots in Grandpass (2013) and Aluthgama (2014), included engagement information for the posts, and distinguished between hate speech and dangerous speech. The authors noted that increases in hate speech occurred alongside increasing violence against minorities (sexual, ethnic, and religious) in the country, and wrote that hate speech "risks fanning even greater violence in the future" (5).

- **Sarjoon, Athambawa, Mohammad Yusoff, and Nordin Hussin. 2016. "Anti-Muslim sentiments and violence: A major threat to ethnic reconciliation and ethnic harmony in post-war Sri Lanka." *Religions* 7(10), 125.** https://doi.org/10.3390/rel7100125

  This paper describes how anti-Muslim sentiment has developed in Sri Lanka in the years following the 1983-2009 Sri Lankan Civil War. Drawing on secondary data collected from newspapers, civil society reports, and academic articles, the authors trace the changing relationship between Sinhalese Buddhists and Muslims in Sri Lanka. They provide many examples of hateful speech against Muslims in the country, such as a speech delivered by Gnanasara Thero, a leader in a Sinhalese Buddhist nationalist organization, in Aluthgama in 2014. In it, he threatened to destroy Muslim-owned businesses and "asked his audience to fight against the minorities" (125). Later that day, an anti-Muslim riot broke out in Aluthgama, during which many businesses were destroyed and four Muslims were killed.

- **Wickremesinhe, Roshini, and Sanjana Hattotuwa. 2015. "Saving Sunil: A study of dangerous speech around a Facebook page dedicated to Sgt. Sunil Rathnayake." *Colombo: Centre for Policy Alternatives* (2015).**
  https://www.cpalanka.org/wp-content/uploads/2015/10/SS-Final-RW-SH-formatted.pdf

  This study analyzes comments on a Facebook page dedicated to 'saving' Sgt. Sunil Rathnayake, an army officer who was convicted of murdering eight internally displaced refugees. On the day when Rathnayake was sentenced to death,[9] the Facebook page was created, and the authors collected comments from it for the following four weeks. They analyzed the comments, of which screenshots and translations appear in the paper, using the dangerous speech framework, and they pointed out common themes in them. Most of the messages attacked members of Sri Lanka's political leadership (those responsible for Sgt. Rathnayake's conviction), while other messages were aimed at Muslims, Tamils, and members of the LGBTQ community. Most of those who posted the hateful comments were young and male. The authors posit that although individual Facebook pages may not each have a substantial impact on violence in the country, they exist in a larger ecosystem of hate on social media where hateful rhetoric comes in 'waves,' often sparked by incidents such as the trial of Sgt. Rathnayake.

The Former Yugoslavia

The articles in this section described the manner in which dangerous speech was used before and during the Yugoslav Wars, a series of ethnic conflicts that took place from 1991 to 2000 after Slovenia, Croatia, Bosnia and Herzegovina, and Macedonia declared their independence from Yugoslavia. Over 140,000 people were killed in the wars, and the United Nations established the International Criminal Tribunal for the former Yugoslavia (ICTY) to deal with the crimes committed during the conflict.

- **De le Brosse, Renaud. 2003. *Political Propaganda and a Plan to create a "State for All Serbs": Consequences of Using the Media for Ultra-Nationalist Ends.* Report Compiled at the request of the Office of the Prosecutor of the International Criminal Tribunal for the Former Yugoslavia (ICTY). 4 February. Case Slobodan Milošević, Case No. IT-02-54.**
  http://balkanwitness.glypx.com/de_la_brosse_pt1.pdf

  This report describes the way the Serbian government used propaganda in the 1980s and 1990s to condition its audience to condone atrocities committed by Serbian forces. De la Brosse offers examples of dangerous speech from the war, and quotes

[9] In 2020, Sgt. Rathnayake was pardoned for his crimes.

from interviews with individuals who were working in media or who were close to then-Serbian President Slobodan Milošević, discussing how media — especially television — was used to build support for mass atrocities. The focus of the report is not causation, but rather the form and content of propaganda during this period.

- **Kiper, Jordan, Yeongjin Gwon, and Richard Ashby Wilson. 2020. "How Propaganda Works: Nationalism, Revenge and Empathy in Serbia." *Journal of Cognition and Culture* 20(5), 403-431. https://doi.org/10.1163/15685373-12340091**

  The authors ran an experiment to determine how exposure to war propaganda changed a person's feelings of in-group and out-group empathy, as well as their support of violence. The study randomly exposed participants (hired through Amazon's MTurk webservice) to one of nine excerpts based on public speeches and texts by Vojislav Šešelj, a Serbian convicted war criminal. The excerpts were altered to replace the in-group and out-group names with those of fictitious nationalities. Each message was chosen to exemplify one of nine types of propaganda identified by Oberschall (2006).[10] The authors tested the impact of only a single exposure to the speech and found that it did not increase support for violence. "References to past atrocities, victimization, revenge and dehumanization" (423) all increased in-group empathy, but only propaganda based on revenge decreased out-group empathy.

- **Mazowiecki, Tadeusz. 1994. Special Report on the Media by Tadeusz Mazowiecki, Special Rapporteur. E/CN.4/1995/54. 13 December 1994**
  https://ap.ohchr.org/documents/alldocs.aspx?doc_id=500

  This report, prepared by the Special Rapporteur of the UN Commission on Human Rights, described the media landscape in Bosnia and Herzegovina, Croatia, the Federal Republics of Yugoslavia (Serbia and Montenegro), and Macedonia. Most relevant for this bibliography, the author includes examples of "incitement to nationalist hatred" in media from the different regions. Although the content in the report is useful for understanding the general media environment in the former Yugoslavia between 1990 and 1994, it is not an exhaustive survey. As noted in the report, "lack of access to certain regions of the former Yugoslavia because of the war and denial of access to others by the Government of the Federal Republic of Yugoslavia made [an exhaustive survey] impossible" (3).

[10] Direct threat or paranoia, referencing past atrocities, victimization, justice, revenge, religion, nationalistic speech, negative outgroup stereotyping, dehumanization. See Oberschall, Anthony. 2006. "Vojislav Seselj's Nationalist Propaganda: Contents, Techniques, Aims and Impacts, 1990–1994." United Nations.

- **Oberschall, Anthony. 2006. "Vojislav Seselj's Nationalist Propaganda: Contents, Techniques, Aims and Impacts, 1990-1994." Expert report for the United Nations International Criminal Tribunal for the Former Yugoslavia, Case No. IT-03-67, MFI. http://www.baginst.org/uploads/1/0/4/8/10486668/vojislav_seseljs_nationalist_propaganda-_contents_techniques_aims_and_impacts.pdf**

  In this expert report prepared for the ICTY, Oberschall argues that mass media propaganda convinced Serbs to accept and participate in collective violence. The report describes social science research on propaganda and gives specific historical notes on the use of propaganda in Serbia. At its core is a content analysis of 242 media messages by Vojislav Seselj, the former deputy prime minister of Serbia, presented with data indicating that "a majority of the Serb public was immersed in and believed the xenophobic nationalist propaganda promoted by Seselj and other nationalists" (45). Of the 242 Seselj texts that Oberschall analyzed, he identified many (38%) that conveyed a message of Serb victimhood or told Serbs that they and Serbia were "besieged and under attack, as it was in the past, by foreign and internal enemies, and by the other people in the former Yugoslavia, especially the Croats" (22). He also notes how Seselj developed sanitized ways of describing atrocities to garner support, such as referring to the forced expulsion of a population, as "a civilized exchange of population" (24). The report is full of useful examples of speech that circulated between 1990-1994 in the former Yugoslavia, such as a false rumor that circulated claiming that Croatian soldiers had murdered 41 Serb schoolchildren (40).

Contribute to this Literature Review

We hope you have found this literature review helpful, and we welcome feedback on how to improve it. If there is another topic you would like to see covered, please let us know. We would appreciate citations for any and all additional literature that contains findings relevant to the study of speech as a driver of intergroup violence.

Please send ideas and inquiries to **Cathy@DangerousSpeech.org**

Dangerous Speech Project

The Dangerous Speech Project is a team of experts on how speech leads to violence. We use our research to advise internet companies, governments, and civil society on how to anticipate, minimize, and respond to harmful discourse in ways that prevent violence while also protecting freedom of expression.