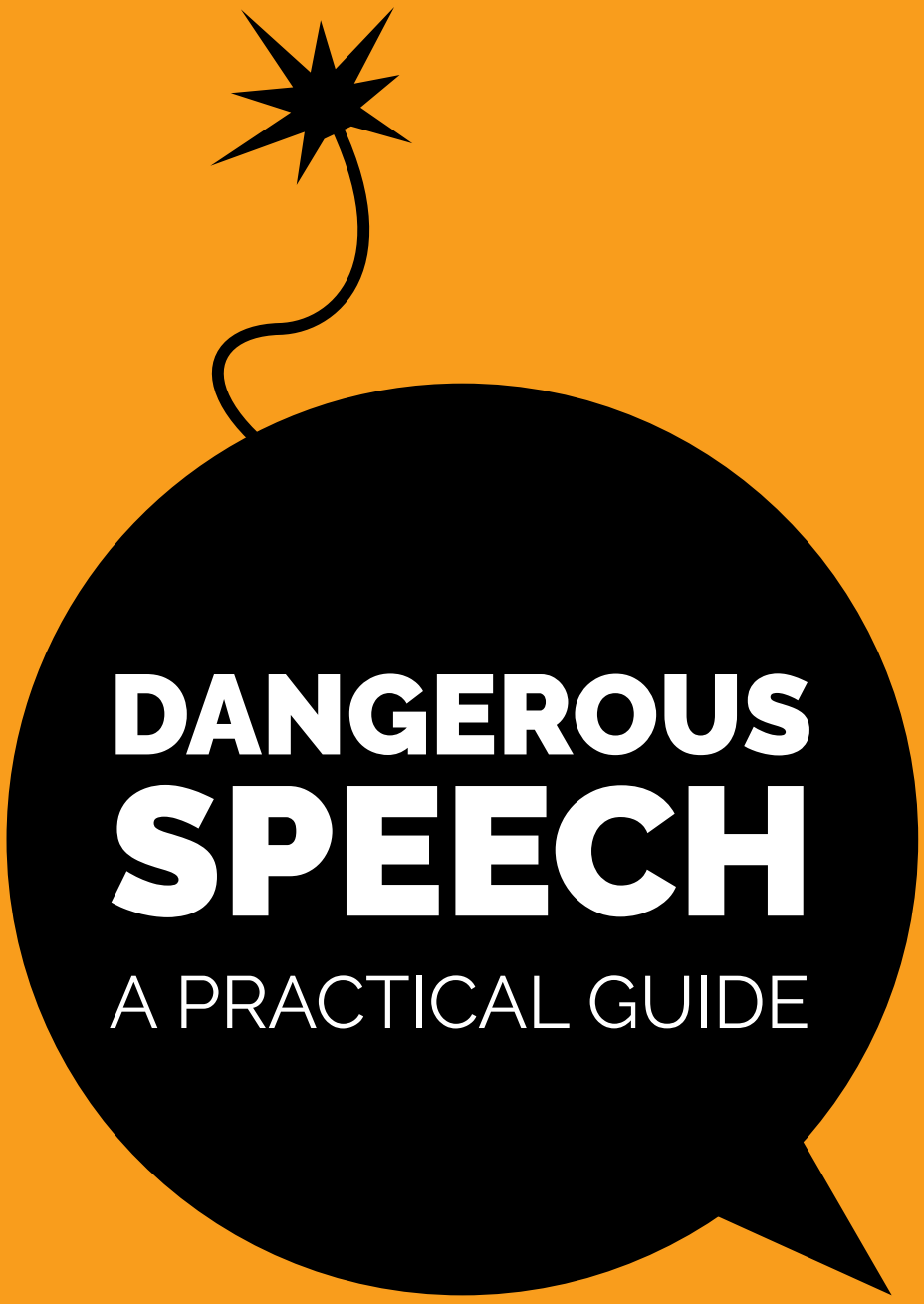


DANGEROUS SPEECH

A PRACTICAL GUIDE

BY THE **DANGEROUS SPEECH PROJECT**

CONTENTS

INTRODUCTION 3

DEFINING DANGEROUS SPEECH 5

DANGEROUS SPEECH IS AIMED AT GROUPS 6

DANGEROUS SPEECH PROMOTES FEAR 6

DANGEROUS SPEECH IS OFTEN FALSE 7

DANGEROUS SPEECH HARMS INDIRECTLY 7

DANGEROUS SPEECH AND HATE SPEECH 7

THE DANGEROUS SPEECH FRAMEWORK 10

1. MESSAGE 12

DANGEROUS SPEECH HALLMARKS 12

- Dehumanization 13

- Accusation in a Mirror 15

- Threat to Group Integrity or Purity 16

- Assertion of Attack Against Women and Girls 17

- Questioning In-Group Loyalty 17

2. AUDIENCE 18

3. CONTEXT 19

4. SPEAKER 20

5. MEDIUM 22

DANGEROUS SPEECH ONLINE—THE ROLE OF SOCIAL MEDIA 23

RESPONDING TO HATEFUL AND DANGEROUS SPEECH ONLINE 26

CONCLUSION 27

REFERENCES 28

INTRODUCTION

No one has ever been born hating or fearing other people. That has to be taught – and those harmful lessons seem to be similar, though they're given in highly disparate cultures, languages, and places. Leaders have used particular kinds of rhetoric to turn groups of people violently against one another throughout human history, by demonizing and denigrating others. Vocabulary varies but the same themes recur: members of other groups are depicted as threats so serious that violence against them comes to seem acceptable or even necessary. Such language (or images or any other form of communication) is what we have termed "Dangerous Speech."

Naming and studying Dangerous Speech can be useful for violence prevention, in several ways. First, a rise in the abundance or severity of Dangerous Speech can serve as an early warning indicator for violence between groups. Second, violence might be prevented or at least diminished by limiting Dangerous Speech or its harmful effects on people. We do not believe this can or should be achieved through censorship. Instead, it's possible to educate people so they become less susceptible to (less likely to believe) Dangerous Speech. The ideas described here have been used around the world, both to monitor and to counter Dangerous Speech.[1]

This guide, a revised version of an earlier text (Benesch, 2013) defines Dangerous Speech, explains how to determine which messages are indeed dangerous, and illustrates why the concept is useful for preventing violence. We also discuss how digital and social media allow Dangerous Speech to spread and threaten peace, and describe some promising methods for reducing Dangerous Speech - or its harmful effects on people.

1. Many of these efforts are described at www.dangerousspeech.org, the website of the Dangerous Speech Project.


Dangerous Speech is
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.

DEFINING DANGEROUS SPEECH

In the early 2000s, Benesch (2003) noticed striking similarities in the rhetoric that political leaders in many countries have used, during the months and years before major violence broke out. Since such messages seem to have special power to inspire violence, we have studied them, in search of ways to diminish their effect and preserve peace. We call this category Dangerous Speech and have defined it 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.***

Importantly, the definition refers to increasing the risk of violence, not causing it. We generally cannot know that speech[2] *caused* violence, except when people are forced by others to commit violence under a credible threat of being killed themselves. People commit violence for many reasons, and there is no reliable way to find them all or to measure their relative importance. Often even the person who commits violence does not fully comprehend the reasons why. To say that speech is dangerous, then, is to make an educated guess about the effect that the speech is likely to have on other people.

In the definition of Dangerous Speech, violence means direct physical (or bodily) harm inflicted on people, not other forms of harm such as doxing,[3] incitement to self-harm, discrimination, or social exclusion.[4] These other forms of harm are important, of course, and Dangerous Speech may inspire people to inflict many forms of harm. In our definition we focus on physical violence since it is easier to measure, and there is greater consensus on what constitutes physical violence.

Also, the definition mentions both committing and condoning violence. The reason for this is that even in the most large-scale violence between people, only a small proportion (usually young men) actually carry out violence. People close to them, however – e.g. siblings, friends, and teachers – often condone or even encourage it. Generally, when a society suffers major intergroup violence, a few commit it and a much larger number condone it.

---

2. We use the term 'speech' to refer to any form of human communication - in keeping with the definition of Dangerous Speech.
3. To dox is to harass or endanger someone by searching for, and then posting online, private or identifying information about that person.
4. Other definitions of violence do include non-physical harm. Peace and conflict studies scholar Johan Galtung, for example, includes discrimination, exclusion, and exploitation in what he calls "structural violence" (1969, p.171). The United Nations Declaration on the Elimination of Violence against Women (1993) defines violence against women as "gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women."

DANGEROUS SPEECH IS AIMED AT GROUPS

Dangerous Speech increases the risk that its audience (the "in-group" as it is often called by scholars) will commit or condone violence against another group (the "out-group"). The out-group must have a defining characteristic that is both different from and meaningful to the audience (whether this accurately describes or is meaningful to members of the out-group, or not). Common dividing lines include race, ethnicity, religion, class, or sexual orientation, but in some cases Dangerous Speech is aimed at groups defined by other characteristics, such as occupation, like journalists. Merely being in the same location or attending the same school would not define a group for the purposes of Dangerous Speech.

Speech targeting individuals is usually outside the scope of Dangerous Speech; however, in some cases an individual can symbolize a group so that targeting that person becomes a form of Dangerous Speech against the group they represent. For example, some Pakistanis have called for killing or maiming the Pakistani Nobel laureate Malala Yousafzai, attacking her as an individual and also as a leader of subversive or traitorous women. (Kugelman, 2017). Similarly, Hungarian Prime Minister Viktor Orban and his government attack the Hungarian-American philanthropist George Soros, as an individual and also as a Jew with money and influence, using familiar anti-semitic tropes such as referring to Soros as a puppet master (Wilson, 2018).

DANGEROUS SPEECH PROMOTES FEAR

A defining feature of Dangerous Speech is that it often promotes fear, as much as it expresses or promotes hatred. For example, one can assert that another group is planning to attack one's own group without expressing hatred, yet that message might easily convince people to condone or commit violence, ostensibly to fend off the attack. Violence would seem defensive, and therefore justified. For example contemporary rhetoric in many countries portrays immigrants as a catastrophic threat. Hungary's Prime Minister Viktor Orbán and United States President Donald Trump have referred to immigrants and refugees as a "trojan horse" which will necessarily increase criminal activity and terrorism (Brunsden, 2017; Kopan, 2015).

Frightening messages may also spread even more widely and quickly than purely hateful ones, since many people share them without malevolent intentions, or even the desire to incite violence. They feel genuine, heartfelt fear.

DANGEROUS SPEECH IS OFTEN FALSE

Dangerous Speech is commonly false - not surprisingly since it describes whole groups of human beings in appalling terms. Unfortunately, it can be difficult to refute falsehoods, especially when they are frightening. Dangerous Speech can be equally effective whether its messages are accurate, false, or greatly exaggerated (Leader Maynard and Benesch, 2016, p. 78).

DANGEROUS SPEECH HARMS INDIRECTLY

Speech can harm directly or indirectly, or both. It may directly offend, denigrate, humiliate or frighten the people it purports to describe – as when a racist shouts at a person of color.[5] Speech can also bring about harm indirectly – and with equal or even much greater brutality – by motivating others to think and act against members of the group in question. This is the work of Dangerous Speech. One message may, of course, harm both directly and indirectly.

DANGEROUS SPEECH AND HATE SPEECH

Dangerous Speech is also quite different from the term "hate speech" which, though it is a widely-used term, is hard to define clearly and consistently. This can endanger freedom of expression, which must always be vigorously protected since it is a fundamental human right – and also because silencing people can make them more likely to resort to violence, if they have no peaceful way of expressing and resolving their grievances.

"Hate speech" is oddly ambiguous. For example, what exactly is hatred? How strong or how durable must an emotion be to count? And does the "hate" in hate speech mean that the speaker hates, or seeks to persuade others to hate, or wants to make people *feel* hated?

Generally, hate speech means vilifying a person or group of people because they belong to a group or share an identity of some kind. This means it's not hate speech to say "I hate you," since there's no reference to a group.

Most definitions specify that to be considered hate speech, messages must be directed at particular types of groups, such as people of the same religion, race,

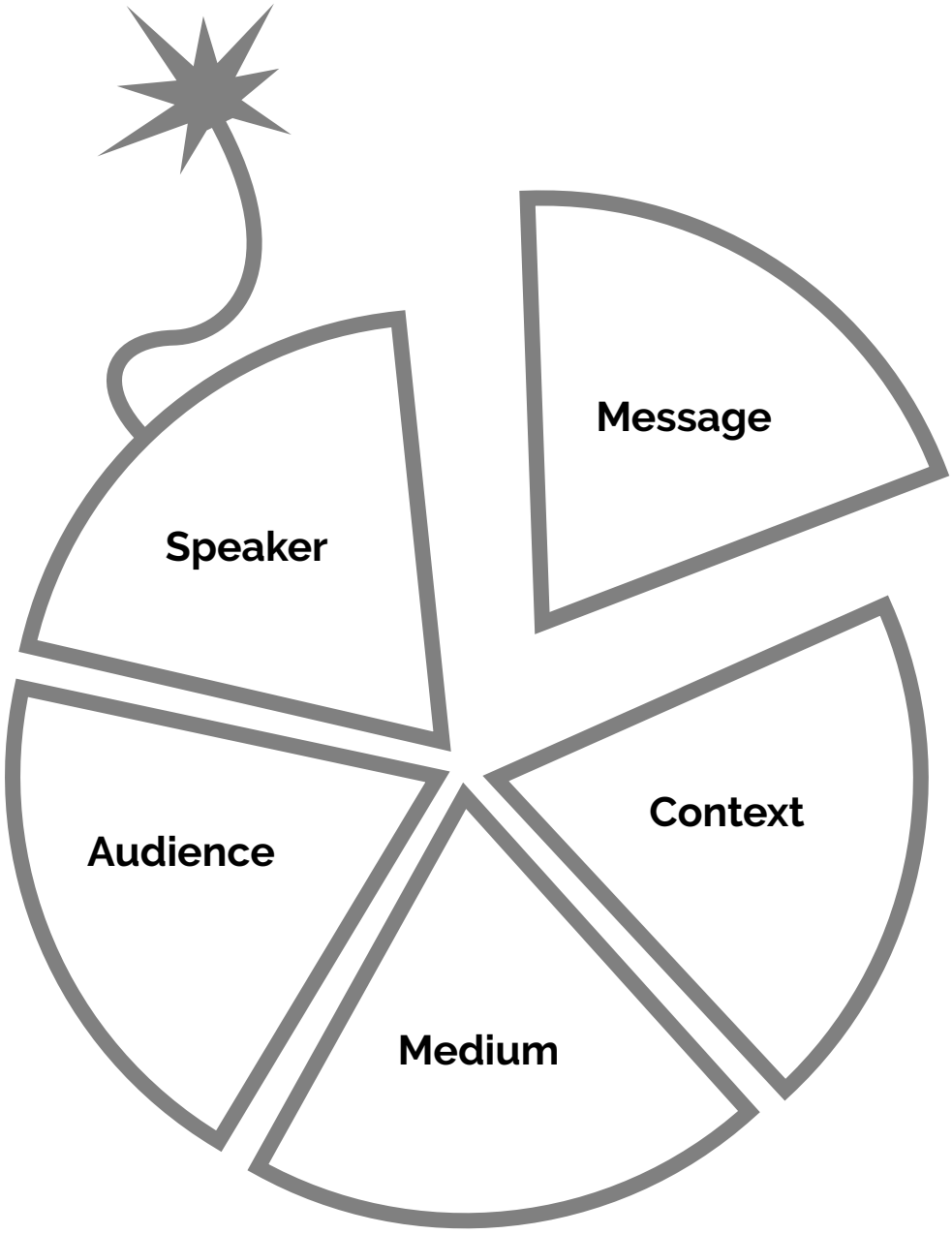
5. Hate speech can also cause other harm, e.g. to those at whom it is not aimed. Jeremy Waldron (2012) has argued that it does intolerable damage to entire societies, even in comparatively peaceful times.

or ethnicity. Some definitions also add disability, sexual orientation, gender, sex, age culture, belief, or life stance. For example section 135a of Norway's penal code defines hate speech as "threatening or insulting anyone, or inciting hatred or persecution of or contempt for anyone because of his or her a) skin color or national or ethnic origin, b) religion or life stance, or c) homosexuality, lifestyle or orientation" (*Norwegian Penal Code*). South Africa's hate speech law is one of the most detailed and comprehensive, specifying groups and attributes that are absent from other countries' laws such as pregnancy, marital status, conscience, language, color, and "any other group where discrimination based on that other ground (i) causes or perpetuates systemic disadvantage; (ii) undermines human dignity; or (iii) adversely affects the equal enjoyment of a person's rights and freedoms in a serious manner that is comparable to discrimination […]" (*Promotion of Equality*, 2000, pp. 3-5). Most countries' laws don't prohibit hate speech at all, instead criminalizing other related forms of speech such as incitement to hatred.

Broad or vague definitions of hate speech and related crimes can jeopardize freedom of speech, since vagueness allows for subjective application. Indeed, laws against hate speech or hateful speech are often misused to punish and silence journalists, dissenters, and minorities, recently in countries as varied as Hungary, India, Rwanda, Kazakhstan, and Bahrain.

We focus instead on Dangerous Speech since it is a narrower, more specific category, defined not by a subjective emotion such as hatred, but by its capacity to inspire a harm that is all too easy to identify – mass violence – and that almost everyone can agree on wanting to prevent.

The Dangerous Speech Five Part Framework



THE DANGEROUS SPEECH FRAMEWORK

One cannot make a list of words that are dangerous, since the way in which any message will be understood – like its effect on the audience[6] – depends not only on its words but on how it is communicated: by whom, to whom, and under what circumstances. The very same words can be highly inflammatory, or benign.

To understand whether a message is dangerous when spread in a particular context, one must examine both content and context. It's important, also, to be able to compare the dangerousness of different messages. To this end we have developed a straightforward and systematic way to analyze speech in context – listing and describing all of the elements that can make a particular example of speech more dangerous. The result is a five-part framework (see Figure 1) which includes the message itself, the audience, the historical and social context of the message, the speaker, and the medium with which a speaker delivers a message. Analyzing each of these five elements is not only essential for identifying how Dangerous Speech operates, it is also useful for designing interventions to diminish the dangerousness of that speech.

To use the framework for a particular example of speech, one asks whether each of the five elements makes it dangerous, and if so, how dangerous. For example, one might ask whether a message came from a compelling or influential source. Because the social, historical, and cultural context in which speech was made or disseminated is essential for understanding its possible impact, this analysis must be carried out with extensive knowledge of the relevant language, culture, and social conditions – or at least with assistance from advisors who have such knowledge.

After considering all five elements in turn, one asks on the basis of that analysis: did/would this message make people more ready to commit or condone violence?

All five elements need not be significant in every case. For example, sometimes the speaker is irrelevant, when unknown (many messages are distributed anonymously, as in an online message or a printed flyer) or not influential with the audience. Such speech may still be dangerous, if its message is inflammatory and the audience is primed to accept it. Only those two elements are always required for speech to be dangerous: inflammatory content and a susceptible audience.

6. In linguistics a "speech act" is communication that brings about some sort of response or change in the world. The 20th-century British philosopher of language J.L. Austin (1962) pioneered speech act theory, in which he tried to capture and distinguish all the types of effects that language can have. "Perlocutionary force," Austin proposed, is the capacity of a speech act to bring about a response in its audience. We draw on this body of thought since Dangerous Speech is communication that has perlocutionary force.

Moreover, it isn't the case that speech is either dangerous or not dangerous at all. Rather, it can be not dangerous, slightly dangerous, very dangerous, or somewhere in between. These can be imagined along a spectrum. Once a moderately dangerous message becomes acceptable to people, a more dangerous message is likely to seem somewhat more acceptable also. In this way, normal social barriers to violence (and discrimination) erode as increasingly dangerous speech begins to saturate the social environment.[7]

In general, the Dangerous Speech that comes just before violence breaks out is easiest to identify since its meaning tends to be clear and it often calls for, or at least endorses, violence. Years or months earlier, speech is often expressed in ambiguous, coded language, so that both its meaning and its impact are less apparent. This doesn't mean that it can be safely disregarded.

Rwandans and scholars generally agree that speech helped to catalyze the 1994 Rwanda genocide in which thousands of Hutu men massacred between 500,000 and 800,000 people, mainly of the Tutsi ethnic group, and mainly by hand, using machetes: such a laborious way to kill that it seems they were highly motivated (Des Forges, 1999). Indeed, inflammatory speech against Tutsi had circulated in Rwanda for years before the genocide, and it was believed to have played such an important role that the International Criminal Tribunal for Rwanda (ICTR) made speech crimes a major focus of its cases. One of the best-known was *the Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza, Hassan Ngeze,* the so-called Media Trial, at which a newspaper editor and two executives of Radio Télévision Libre des Milles Collines (RTLM) – bitterly nicknamed Radio Machete – were all convicted. Much of the trial focused on ambiguous language, though, not explicit encouragement to kill.

During the trial, a witness recounted the spread of what we call Dangerous Speech,[8] over RTLM's existence from July 1993 to July 1994. "I monitored the RTLM virtually from the day of its creation to the end of the genocide, and, as a witness of facts, I observed that the operation of the genocide was not the work done within a day." The witness went on to describe RTLM's effect on its audience:

"[W]hat RTLM did was almost to pour petrol - to spread petrol throughout the country

7. This process can also be described with reference to the Overton Window, a theory of the way the acceptable range of political discourse, or policies, changes over time. The theory's originator Joseph Overton imagined a window containing views or policies that are acceptable to the opinion leaders or the majority, in a group. As the window moves, once-radical positions or ideas become more acceptable, and even ideas that were once unthinkable can eventually be found inside the window (Lehman, 2010).
8. The three defendants were convicted of incitement to genocide, among other grave crimes. Dangerous Speech is not a crime in any country's penal code, nor do we suggest that it should be criminalized. There are already related speech crimes in most bodies of law, and we believe that criminal law is generally not a very effective way of limiting speech or its harmful effects.

little by little, so that one day it would be able to set fire to the whole country."[9] As this implies, Dangerous Speech of all types should be analyzed carefully, to gauge its harmful effects and also to avoid defining it too broadly: some offensive or hateful speech isn't dangerous at all. The framework below is meant for identifying "drops of petrol," and making an educated, systematic guess as to where they fit along a spectrum of dangerousness.

1. MESSAGE

People express themselves in a seemingly infinite variety of ways, and Dangerous Speech is no exception. Quite often, it isn't explicit. For example, it may be expressed in language familiar to the in-group but not to the out-group, since shared jargon or code serves to bind the in-group together, and at the same time, allows leaders who use Dangerous Speech to deny it.

Regardless of the language or images with which it is expressed, we have found that Dangerous Speech often contains similar ideas, or what the scholar Jonathan Leader Maynard (2014) calls "justificatory mechanisms" – language used to justify violence against groups of people. We call such rhetorical patterns "hallmarks" of Dangerous Speech and describe some of them below. Note that a hallmark does not, by itself, make a message dangerous.

All groups of humans use these techniques, regardless of language, country, race, color, or class – just as virtually all groups also commit violence against other people. Similarly, this kind of rhetoric is found throughout human history.

DANGEROUS SPEECH HALLMARKS

We have identified five hallmarks of Dangerous Speech, which we call: dehumanization, accusation in a mirror, threats to group integrity or purity, assertions of attacks against women and girls, and questioning in-group loyalty. This list is not exhaustive – we expect it to grow and change as researchers gather more Dangerous Speech and observe patterns in it. All the examples of Dangerous Speech that we have found contain at least one of the hallmarks below.

9. *Prosecutor v. Nahimana et al. (Trial Judgment)*, para. 436. It's important to recognize that the witness made a subjective and unscientific attempt to gauge the effect of RTLM's broadcasts on a large number of people. Scholars who have since studied the impact of RTLM include David Yanagizawa-Drott (2014) and Scott Straus (2007).

Dehumanization:

By describing other groups of people as something other than human, or less than human, speakers can persuade their audiences to deny other people some of the moral consideration they give to those who are "fully" human (Leader Maynard and Benesch, 2016, pp. 80-81). Dehumanizing targets prepares audiences to condone or commit violence, by making their targets' death and suffering seem less significant, or even by making it seem useful or necessary.

There are several types of dehumanizing messages, each of which elicits certain emotional or practical responses.[10]

Speakers often describe an out-group as biologically subhuman: as animals, insects, or even microorganisms such as bacteria or viruses. Persistently, in cases of genocide and mass atrocity, supporters and perpetrators have referred to their victims as vermin (rats, cockroaches, foxes, or snakes), beasts (apes or baboons), or biological hazards (a virus, tumors, or an infection). Not at all language comparing people to animals or other non-human creatures is dehumanizing or dangerous, of course – it's possible to compare a person to an animal in a way that doesn't lower social barriers to violence.

Generally, speakers choose to compare out-group members with creatures that their audiences regard as repulsive, threatening, or deserving of violence (Leader Maynard, 2015, p. 197). It is almost instinctual knowledge, for example, how to deal with an infestation of vermin: try to eliminate the creatures completely. When Rwandan Hutu extremist media referred to the Tutsi ethnic group as cockroaches in the months preceding the 1994 genocide which left hundreds of thousands of Tutsis dead, they suggested the same action – extermination; one military training operation was even called "Operation Insecticide" (Des Forges, 1999, p. 666).

In the same way, government rhetoric during the Cambodian genocide warned that enemies of the Khmer Rouge regime were "microbes" and a "sickness" to be completely eliminated lest they "rot us from within" (Hinton, 2005, p. 147). One regime slogan declared, "What is infected must be cut; what is rotten must be removed" (Weitz, 2015, p. 156). Like depictions of humans as an infestation of insects, these messages were meant to disgust - but they also suggest that, like cancerous growth or bacterial infections, the Khmer Rouge's opponents had to be removed completely. Indeed, government soldiers killed more than one million Cambodians between 1975 and 1979, by forced labor, torture, and mass execution.

---

10. Scholars have described dehumanization in some detail, observing distinct forms of it and seeking to explain it. Psychologist Nick Haslam proposed two categories: animalistic dehumanization (viewing other people as animals) and mechanistic dehumanization (asserting that other people lack typical human qualities) (2006, p.258). In a monograph on dehumanization, philosopher David Livingstone Smith suggests that humans are prone to dehumanizing others because of what he describes as our "cognitive architecture" (2011).

Speakers also refer to out-groups using supernatural terms. Unlike forms of dehumanization which make targets seem lesser or weak, supernatural dehumanization makes them seem stronger than humans and threatening to them. For example, during World War II, Japanese propaganda portrayed American and British leaders as "demons," "evil spirits," and "monsters" (Smith, 2011, p. 22). U.S. war propaganda posters similarly demonized Japanese and German people (Brcak and Pavia, 1994, p. 682; Lane, 2014, pp. 49-53). And in the decades following the United States' Civil War and the emancipation of slaves in the country, newspapers covered lynchings of black people by white supremacists by describing the victims as "inhuman" or "unnatural" monsters who terrorized white communities (Smith, 2018).

The language of environmental threats such as floods and pollution can also be used to dehumanize people. This is now common around the world, amid anxiety about climate change. Speakers in many countries have compared present-day mass migration to environmental catastrophe, from Israel's Prime Minister Benjamin Netanyahu, who said that if Israel took down its border fence with Egypt, it would face "attacks by terrorist groups in the Sinai and the worst thing: a flood of illegal infiltrators from Africa" (Zikri, 2018), to the United Kingdom's Daily Mail newspaper, which ran a headline comparing the supposed threat of a "tidal wave of migrants" to that of the Second World War (Burleigh, 2015). Comparisons like these are not a new phenomenon: in 1920, American eugenicist Lothrop Stoddard warned of "a rising tide of color" which would destroy the privileged status of white people in the United States once it permitted non-white immigration to increase (Smith, 2011, p. 97). In 1914, when 376 people from India (unsuccessfully) attempted to immigrate to Canada on the S.S. Komagata Maru, the Vancouver Sun newspaper ran a cartoon with the title "Will the Dyke Hold?" which depicted a tidal wave shaped like a man in a turban, racing toward the Canadian coastline (Mackie, 2014). While these examples of "flooding" human beings were meant to justify government policy to exclude people, similar rhetoric is also used for forcing them out violently. In 1915, clandestine plans to ethnically cleanse Armenians from the Ottoman Empire referred to uprooting "malignant weeds" (Kuper, 1981, p. 91), just as radio broadcasts during Kenya's 2008 election encouraged the Kalenjin tribe to "clear the weeds" in reference to their rival tribe, the Kikuyu (McCrummen, 2008). In both cases, these messages preceded widespread violence, killings, and mass displacement.

Dehumanizing rhetoric needn't refer explicitly to people as something other than human; a speaker may instead use terms that imply dehumanization. For example, when Brazilian politician - now President - Jair Bolsonaro visited a quilombo (a community inhabited primarily by the descendants of African slaves) in 2017, he mockingly described a black man as weighing 7 *arrobas* - using a weight unit used in the country's agriculture industry, especially for cattle (Simões, 2018).

Lastly, like all other hallmarks, dehumanization is neither necessary nor sufficient for Dangerous Speech. People can inflict violence on others while perceiving them

as human. Paul Bloom (2017) writes that people need not dehumanize others in order to mistreat or even torture them. On the contrary, he argues, one can only take full satisfaction from inflicting cruelty when one's victims can feel humiliated and debased - which are human qualities. "The sadism of treating human beings like vermin lies precisely in the recognition that they are not."

Accusation in a Mirror

Combatants in intergroup conflicts often try to frame violence as a necessary means to protect against greater harm. Dangerous Speech often includes a specific kind of collective justification of violence that has become known to scholars as "Accusation in a Mirror" and sometimes abbreviated as AiM. The term comes from an anonymous manual for propaganda and recruitment found in Butare, Rwanda after the 1994 genocide. The document advises attributing to one's enemies the very acts of violence the speaker hopes to commit against them. "In this way," the author writes, "the party which is using terror will accuse the enemy of using terror" (Des Forges 1999, p. 66).

To predict violence from another group is especially powerful (whether the threat is real, false, or exaggerated) since it makes violence against that group seem defensive and *necessary*. In this sense, accusation in a mirror is a collective analogue of the defense to homicide that is available in virtually all legal systems: self-defense. To believe that you, your family, your group, or even your culture faces an existential threat from another group makes violence to fend off that threat seem not only acceptable (as dehumanization does), but necessary.

One of the Rwandan propagandists who famously used this technique is Léon Mugesera, whom Canada deported after the Canadian Supreme Court found sufficient "reasonable grounds to believe" that he had committed incitement to genocide, based on a speech he gave in Rwanda in November 1992 (17 months before the genocide began) in which he told his Hutu audience that they were in mortal danger. For instance, he said a Hutu man had been summarily shot by armed men – Tutsi, his audience was meant to understand. Then he predicted much worse: "they only want to exterminate us: they have no other aim." (*Mugesera v. Canada*, 2005; Straus, n.d.). Mugesera was later convicted of genocide crimes in Rwanda based on his public speech before the genocide and sentenced to life in prison.

The technique of AiM was hardly invented by Hutu extremists: it is one of the most common hallmarks of Dangerous Speech. In Nazi Germany, for example, anti-Semitic propaganda repeatedly and relentlessly accused Jewish people of hatching a *Mordplot* (murderous plan) to eliminate all non-Jews (Streicher, 1934, p. 1). This assertion was especially preposterous since the Jews had no military or guerrilla force at all, yet it was apparently convincing.

Some of the most powerful AiM messages come from speakers who suggest that their own group is in danger of being totally annihilated: that it faces genocide.

For example, Nazi SS Reichsführer Heinrich Himmler told senior officers in 1943 that "we had the moral right ... to wipe out [the Jewish people] bent on wiping us out" (Leader Maynard, 2015, p. 203). And General Ratko Mladić, who became known as the "Butcher of Bosnia" for directing killings including the massacre of more than 8,000 Bosnian Muslim men and boys at Srebrenica in 1995 (Osborne, 2017), had earlier claimed that Muslims, Germans, and Croatians were planning for "the complete annihilation of the Serbian people" (Kiernan, 2009, p. 591).

Threat to Group Integrity or Purity

Another rhetorical technique, or hallmark of Dangerous Speech, is to assert that members of another group can cause irreparable damage to the integrity or purity of one's own group. A 1931 German cartoon from Julius Streicher's Nazi newspaper *Der Stürmer* shows an apple sliced open with a knife marked with a swastika. Inside the apple is a worm that has a stereotypically Jewish face. The caption reads "*Wo etwas faul ist, ist der Jude die Ursache*" ("Where something is rotten, the Jew is the cause") (Bytwerk, n.d.). Similarly, in the ethnic attacks following the December 2007 presidential election in Kenya, members of the Kalenjin (the President's ethnic group) referred to Kikuyu people as "madoadoa" (spots) that had to be removed (Thuku, 2014).

By portraying members of the target group as a threat to the audience group, this type of message reinforces fear. Moreover, these messages indirectly (and sometimes directly) instruct people to rid their group of the supposed contaminant, to preserve the health of their own group.

Notably, this hallmark need not include any prediction of physical violence. A culture, group identity, or political project may be threatened instead (Chirot and McCauley, 2010, p. 62). While such messages may not invoke fears of bodily harm, they appeal to powerful emotional connections that connect people to their identity groups and belief systems. Norwegian mass murderer Anders Breivik, who killed 77 people in July 2011, was motivated by what he called a European "cultural suicide" brought upon by the influences of multiculturalism, Islam, and "cultural Marxism"(Berwick, 2011, p. 12). In his manifesto (written under the pseudonym Anders Berwick), Breivik wrote that "the fate of European civilization" depends on men like him resisting these influences (Berwick, 2011, p. 38). Communists in the Soviet Union appealed to similar threats while justifying violence against kulaks, landowning peasants who resisted collectivization. One Bolshevik leader instructed Communist Party organizers: "beat down the kulak agent wherever he raises his head. It's war - it's them or us" (Figes, 2008, p. 85).

Assertion of Attack Against Women and Girls

Related to the previous hallmark is the suggestion that women or girls of the in-group have been or will be threatened, harassed, or defiled by members of an out-group. In many cases, the purity of women symbolizes the purity, identity, or way of life of the group itself.

This hallmark is very common in Dangerous Speech around the world and throughout history, likely because it is difficult to ignore a warning of violence against members of a group who are traditionally viewed as vulnerable and needing protection. For most societies, this includes children (especially girls) and women; almost universally, men are instructed to protect women and children at all costs, up to and including killing an attacker.

In the United States, false claims of attacks against white women often led to lynchings and other violence against black people, especially in parts of the country where Africans had been enslaved. In Tulsa, Oklahoma, for example, after a report that black men had assaulted white women in 1921, mobs of whites destroyed the homes of black residents (Johnson, 1998, pp. 258-259). Narratives and images of black men attacking white women also appeared in popular media such as the 1915 film *Birth of a Nation.* Like the book *The Clansman* on which it is based, the film depicts a black man attempting to rape a white woman, who escapes only by jumping to her death.

In one of many present-day examples, rumors that Rohingya Muslim men had raped a Buddhist woman in 2012 in Myanmar[11] sparked riots (Gowen, 2017). In February of 2016, the conservative mass-market Polish weekly wSieci published a striking cover image of a beautiful young blonde, blue-eyed woman wearing a dress made from the flag of the European Union. Six dark-skinned male hands grab and tear at her body (and the dress) as she screams in terror. Though the image makes its meaning obvious, it was accompanied by the headline "Islamski gwałt na Europie" (Islamic rape of Europe). In each of these cases, men from the out-group are portrayed as criminal and/or barbaric, heightening a sense of threat.

Questioning In-Group Loyalty

Though Dangerous Speech usually describes members of the out-group or target group, some of it never mentions them, instead characterizing members of the in-group as insufficiently loyal, or even traitorous, for being sympathetic to the out-group. During atrocities, in-group members seen as disloyal are often punished as severely, if not more severely, than members of the out-group. In the Rwandan

11. Myanmar and Burma are the same country. The British who colonized the country called it "Burma," and the ruling military junta changed that name to "Myanmar" in 1989, but both names are still used.

genocide, for example, for the most part Hutus killed Tutsis, but so-called "moderate" Hutus were also often killed by their fellow Hutus, for helping Tutsis or apparently wanting to do so. The radio station RTLM spread the message "kill or be killed," which both supported the idea that killing Tutsis was an act of self-defense and also the notion that Hutus who did not take part in the killing would themselves be killed (Yanagizawa-Drott, 2014, p. 1946). As Mary Kimani (2007, p. 113) notes, "RTLM, as well as political leaders, made it clear that killing 'the enemy' was the duty of every Rwandan."

Such messages were also common in the years leading up to the genocide. In December of 1990, *Kangura,* a pro-Hutu newspaper whose editor was later convicted for incitement to genocide in the Media Trial described above, published the "Hutu Ten Commandments," which called Tutsi a "common enemy" and asserted that Hutus who formed romantic or business relationships with Tutsis were traitors.[12] Hutus sympathetic to Tutsis, in other words, posed a threat to the unity and survival of the Hutu people.

2. AUDIENCE

Even the most inflammatory message is unlikely to inspire violence if its audience is not already susceptible to such messages – for any number of reasons. A group may be fearful about past or present threats of violence, or may be "on edge" due to a social environment that is already saturated with fear-inducing messages. For example, mobs of people have lynched 33 innocent victims in India since 2017 after false rumors of roving child traffickers spread throughout the country (Saldanha, Hazare, and Rajput, 2018). Economic hardship, alienation, unresolved collective trauma, or social norms in favor of obedience to authority may also make people more susceptible to Dangerous Speech.

Dangerous Speech is often false, so audiences are more vulnerable to it when they can be duped into believing what's false – or are not skilled at distinguishing lies from truth. As false content propagates more and more widely online, it can lead to violence, and it seems to diminish participation in civic life. Researchers are trying to understand why people are more or less easily convinced by lies - to learn how to change this for the better. A study published in September 2018 (Shen et al.) indicates that Internet skills, photo-editing experience, and social media use were significant predictors of image credibility evaluation. In other words, people with less experience on digital media are more likely to be duped by false content.

12. The Hutu 10 Commandments (or "Ten Commandments of the Bahutu") were originally published in Kinyarwanda. This translation was taken from Berry, J.A. and Berry, C.P. eds. (1999).

Sometimes, speakers use specific language that isn't dangerous in itself, but can render other messages more dangerous, by binding the members of a group more tightly to each other, to the group itself, and/or to its leader, or by strengthening distinctions between the in-group and the out-group. A common form of this binding speech is "kinship talk," language that gives a sense of familial belonging to members of a group. In some cases, for instance, this talk tells them that they are bound by blood, not just politics. Such messages can amplify the effects of hallmarks of Dangerous Speech.

Most messages reach many types of people, and each receives them somewhat differently. Some people are much more willing and able to commit violence, for instance, though almost anyone can do so under certain circumstances, especially when they perceive an imminent threat to themselves or their fellow human beings (Leader Maynard and Benesch, 2016, p. 78). When analyzing speech for dangerousness, we try to predict its effect on the groups or individuals who are most susceptible, or most likely to commit violence.

Even where a group does not seem susceptible to Dangerous Speech, a few of its members usually are. So-called "lone wolf" attackers can be understood either as the most susceptible members of a group, or as individual "audiences," moved to commit violence on their own. One lone wolf inspired by Dangerous Speech is Timothy McVeigh, who killed 168 people by bombing a U.S. government building in the state of Oklahoma in 1995, motivated and guided (in part) by *The Turner Diaries*, a racist, anti-Semitic novel in which characters commit a similar attack (Thomas, 2001).

3. CONTEXT

The social and historical context in which speech spreads also affects the extent to which it is dangerous, since any message may be understood in dramatically different ways in one place or time versus another. Any number of aspects of context may be relevant. When conducting a Dangerous Speech analysis, one should consider as many of those as possible.

For example, is there a history of violence between the groups? Messages encouraging violence, or describing another group as planning violence, are more inflammatory where groups have exchanged violence in the past, or where there are longstanding, unresolved grievances between them. Former attacks tend to weaken or remove psychological barriers to violence. The Israeli-Palestinian conflict is a striking example of this, as is recurring intercommunal violence in many parts of India. Unfortunately, there are dozens of other such cases around the world, in which old fighting and violence always form a kind of collective psychological backdrop, and it is all too easy to catalyze new violence with words.

Another question to consider is whether there are social norms, laws, and/or policies that put one group at special and persistent risk. Systemic discrimination can create a context in which it seems entirely normal – because it is officially and widely sanctioned – to regard a group of people as inferior, deficient, or wicked. For example in Pakistan the Ahmadi, a religious minority, are denounced in the law, by clerics, political leaders, and even by journalists as traitors to Islam, the national religion. As the Ahmadis' beliefs are legally considered blasphemous, they often face social boycott and much worse on account of their religion (Khan, 2003) or even their efforts to defend themselves against Dangerous Speech.

The Pakistani Supreme Court condemned three Ahmadi men to death in October 2017 for taking down an anti-Ahmadi sign (Hashim, 2017), and a fourth man would have faced death at the hands of the state also, but a teenager had walked into the police station where he was being held in 2014 and shot him to death (Houreld, 2014).

Within this context, anti-Ahmadi speech is even more dangerous as the state has already proven its unwillingness to protect the Ahmadi or treat them as equal citizens. Discriminatory legal systems normalize persecution and create a context in which members of the in-group (usually the majority) feel protected for their personal acts of discrimination and even violence against members of the out-group.

Other aspects of social or historical context, such as whether there is competition between groups for resources like land or water, are also important to consider.

4. SPEAKER

When a speaker is unusually influential, this can make their speech more dangerous. Influence or authority can come from a variety of sources, including personal charisma, high social status, or official status such as political office – which may also come with control of resources needed by the audience, and the power to deploy force against uncooperative audience members. In other cases, a speaker's influence may derive from cultural stature as an unelected community leader, popular entertainer, or star athlete; indeed, religious and cultural leaders often have more influence over an audience than politicians.

A close family member or trusted friend might also be highly influential. This is especially relevant to a social media platform like Facebook or a digital messaging system like WhatsApp, where users connect to such people. But the source of a message may also be unknown, or there may be multiple sources of the same message.

The source of Dangerous Speech need not be a person, of course – it may be an organization, company, group, or government. In fact, governments often have

disproportionate influence, and are powerful disseminators of Dangerous Speech. Moreover, governments speak not only in official statements, but also through law. For example, Russia's 2018 law banning the distribution of "homosexual propaganda" to minors endangers LGBTQ people by vilifying their existence. The law seems designed to reinforce existing discriminatory attitudes and fears among the Russian population. An all-too-common phenomenon, this law both emerges from and reinforces discriminatory and even dangerous social norms.

THE SECOND SPEAKER

In many cases, a speaker makes a message dangerous not by creating it, but by distributing, and often distorting, someone else's content. In mid-2017, a video clip began circulating virally in India on WhatsApp, a platform which was then used by 200 million people in that country (Elliott, 2018). The clip seemed to show security camera footage of a child being kidnapped. What most of the furious, frightened people who shared it didn't know is that the clip was part of a longer video showing a mock kidnapping in which the child is safely returned – made by a Pakistani charity to raise awareness about child abductions (Rebelo, 2017). The distorted version omitted the name of the charity, the campaign, and the safe return of the child. Instead, it falsely accused people in India of kidnapping, and inspired gruesome vigilante lynchings. As many such rumors circulated online and offline, mobs killed 33 people in India between January 2017 and July 2018 (Sanghvi, 2018).

"Second" speakers may also play an important role by carrying messages to a new audience,[13] or to a much larger one than the original speaker could reach. In November 2017, U.S. President Donald Trump retweeted a series of shockingly violent videos. One of them was falsely titled, "Muslim migrant beats up Dutch boy on crutches!" – the Embassy of the Netherlands in the United States indicated via its own Twitter account that the boy who did the beating was not a Muslim migrant (Netherlands Embassy, 2017).

The videos were originally shared by Jayda Fransen, deputy leader of the far-right extremist group Britain First. Fransen then had 52,776 followers; Trump had over 42 million (Data Team, 2017). By retweeting the messages, the president not only disseminated Dangerous Speech to a much larger audience, but increased the legitimacy of the extremist message by endorsing it. Trump did not create the content; he gave it his highly influential voice.

13. Those who carry information across the social or cultural boundaries between groups are sometimes called "bridge figures." For further description of this, see Benesch, 2015.

5. MEDIUM

Speech may take any number of forms, and can be disseminated by myriad means. It may be shouted during a rally, played on the radio as a song, captured in a photograph, written in a newspaper or on a poster, or shared through social media. The form of the speech and the manner in which it is disseminated affect how the message is received and therefore, how dangerous it is.

There are several factors to consider when analyzing a medium. The first is whether the speech was transmitted in a way that would allow it to reach a large audience. Private conversation around a dinner table, for example, will not reach as many people as a post on a public Facebook page with many followers.

A second fact is whether the speech was transmitted in a way that would reinforce its capacity to persuade. For example, was it repeated frequently? Repetition tends to increase the acceptance of an idea. Or was the speech published in or broadcast on a media source that is particularly influential or respected among the intended audience? In the same way that an influential speaker lends legitimacy to a message, a media source that is trusted by a particular audience will lend credibility to the messages it spreads.

The particular language used by the speaker may also play a role. In fieldwork on violence prevention efforts in Kenya following the 2007-2008 post-election violence there, one of us (Benesch, 2014) was told independently by more than one Kenyan that if they heard a message in English or Kiswahili (Kenyan national languages), they heard it with their heads. If the same message came in the listener's vernacular language (or "mother tongue"), they said they heard it with their hearts (Benesch, 2014, p. 25).

Messages also tend to have a greater capacity to persuade if there are no alternative sources of news available, or if other sources don't seem credible. In Myanmar, most people relied on government-controlled radio, television, and newspapers for decades until the country emerged from military rule in 2012. Only 1.1 percent then had access to the internet. Within only four years, half the population had a mobile phone – and most of those had free access to Facebook (Stecklow, 2018) which for many became synonymous with the internet itself (Beech and Nang, 2018). As a result, Facebook became a highly influential medium, used to spread frightening, false messages intended to turn the majority population against minority Rohingya Muslims, even as the country's military has carried out a vicious campaign to drive the Rohingya out, including rape, killing, and burning villages (Specia and Mozur, 2017). A Burmese administrator of a village that has banned Muslims from even spending the night there told The New York Times, "I have to thank Facebook because it is giving me the true information in Myanmar" (Beech, 2017).

For generations, the Rohingya have faced discrimination and exclusion, and have been denied legal citizenship. Violence against them increased as government officials, influential Buddhist monks, and anonymous online sources described them as dangerous. Many also spread false rumors of upcoming attacks by Rohingya (Ingram, 2017) and dehumanized them, calling them "dogs," "maggots," "rapists," or "pigs," and calling for violence against them. Some posts even called for genocide – one Facebook page was called "We will genocide all of the Muslims and feed them to the dogs" (Stecklow, 2018). This rhetoric, much of which Facebook's content moderators failed to detect, intensified as Myanmar escalated its campaign of forced relocation, driving almost one million Rohingya into Bangladesh. A Facebook post from September 2017 states "These non-human kalar dogs, the Bengalis, are killing and destroying our land, our water, and our ethnic people…We need to destroy their race" (Stecklow, 2018).[14]

DANGEROUS SPEECH ONLINE — THE ROLE OF SOCIAL MEDIA

Digital media and the internet have immeasurably changed the way people spread all kinds of messages, from the innocuous to the incendiary. Those who seek to turn groups of people violently against each other can spread Dangerous Speech quickly – especially in places where there is already a risk of mass violence. Ideas and narratives once confined to the fringes of popular discourse – including extremist ideas – are now widely available. Speakers who could hardly find an audience offline, even those who espouse the most widely-derided ideologies, can find at least a few fellow-thinkers across the world, and can form so-called "echo chambers" in which they bolster and further radicalize each other. By forging such bonds, people can collectively disseminate harmful content further than they could have alone and with the fervor of solidarity. Others are motivated neither by hatred nor conviction, but by simply wanting more followers and/or more money (from subscribers or advertisers).

Online, people can also communicate anonymously. On social media platforms like Twitter or Reddit, or messaging platforms like WhatsApp or Discord, they can spread ideas that they might not dare to express offline, where their identities would be known.

As it has become increasingly obvious that online content leads to serious offline harm, governments, researchers, activists, and internet companies have sought ways to diminish the problem. The first, most obvious response is simply to remove bad content or censor it. Each country has laws prohibiting certain forms of speech (they

14. The term "*kalar*" is a slur commonly used in Myanmar to denigrate Rohingya. It implies dark skin, and foreignness (OHCHR, 2018, p. 168). Rohingya are also often called "Bengalis" to refer to their Bangladeshi ancestry and imply that they do not belong – and have no right to stay – in Myanmar.

vary) and social media companies like Facebook and Twitter also have their own rules forbidding certain kinds of content, such as hate speech, nudity, or incitement to violence (Facebook, Inc., 2018; Twitter, Inc., 2018).

Censorship, whether by governments or private companies, poses significant risks to democracy and freedom of expression since it's almost impossible to do it without making serious mistakes. First, although some content is obviously harmful or even illegal, most is quite context-dependent or ambiguous, and it's often difficult to agree on where to draw the lines.

Second, policing the internet for harmful content is a job so huge that its scale is hard even to imagine: every day, 1.47 billion people log on to Facebook alone and post billions of pieces of information (Zephoria Digital Marketing, 2018). Although internet companies train thousands of people (often ill-paid, and psychologically battered from looking at terrible content all day) to decide which posts to take down, at such a scale mistakes are inevitable and numerous (Roberts, 2014, pp. 15-16; Ohlheiser, 2017; Shahani, 2016).

Social media companies are increasingly turning to automated methods (software) to detect a variety of types of content they want to take down, such as terrorist recruiting and hate speech. Although this might seem like an efficient solution, it doesn't work well, and it also threatens freedom of expression. First, software makes lots of mistakes. People express hatred, denigrate others, or promote fear in a wide and creative variety of ways. Moreover, computers can't make some distinctions that humans can, such as to distinguish hate speech from a post denouncing it (Saleem et al., 2016).

Another reason not to rely on deleting harmful content is that it can foreclose other kinds of constructive responses. The simplest response – to express disagreement – can usefully demonstrate that the majority disagrees with hateful views. In fact, the presumed power of "counterspeech," which we define as "direct responses to hateful or harmful speech" (Wright et al., 2017) is one of the main reasons why United States law protects freedom of speech so vigorously, refusing even to prohibit hate speech. If the "marketplace of ideas" is left as open as possible, the theory suggests, the best and safest ideas will eventually prevail (Brandenburg v. Ohio, 1969).

Evidence to prove or disprove this theory is scarce, but there are many intriguing uses of counterspeech, offline and online. For example, when a hate group sought to post anti-Muslim signs on public buses and trains in several U.S. cities in 2010, some cities tried to refuse. The group sued, and some courts allowed cities to reject the signs while others ruled that they must be displayed. In Detroit, where the ads were suppressed, public attention focused on the signs' author, as a victim whose free speech rights were violated. In New York where the ads appeared, members of the public spoke against them and produced Muslim-defending ads to hang alongside the inflammatory ones (Abdelkader, 2014, pp. 81-82). A striking example

of online successful counterspeech is the case of Megan Phelps-Roper. Although she grew up as a fervently loyal member of the extremist homophobic Westboro Baptist Church (founded by her grandfather), Phelps-Roper changed her beliefs, mainly thanks to a few long-running individual conversations with counterspeakers on Twitter (Chen, 2015).

At this writing, some internet companies are also experimenting with other alternatives to deletion, intended to limit the circulation of Dangerous Speech and other forms of harmful content. For example, after inflammatory rumors spread in India as described in the section entitled "Speaker" above, WhatsApp took steps to limit the spread of dangerous messages. It restricted the number of groups or individual accounts to which one can forward a particular message to 20 or fewer – and no more than five in India (WhatsApp, 2018).

RESPONDING TO HATEFUL AND DANGEROUS SPEECH ONLINE

There are also many other ways to diminish harmful content or its damaging effects. One might try to persuade people to stop posting such content in the first place (a preventive approach, rather than a reactive one like deletion), or support those who are attacked by it.

Internet users themselves (not governments or companies) are conducting many ingenious experiments in responding to harmful content online. At the Dangerous Speech Project we are searching out and studying such efforts, and will publish two major reports on them in 2019.

There are also many educational resources to help individuals respond to hateful and harmful speech in productive ways – while protecting themselves from attack. Here are a few examples: "Seriously," an online program created by the French organization Renaissance Numérique, educates people on which tone and content make the best counterspeech. Over Zero, a nonprofit located in Washington, D.C., trains people to apply the Dangerous Speech framework for designing interventions to make the speech less dangerous, in context (Brown, 2016). In 2017 our Dangerous Speech Project, along with #ICANHELP, iCanHelpline.org, HeartMob, and Project HEAR, created a comic for youth, illustrating several "dos" and "don'ts" for effective counterspeech.[15]

15. Comic available at https://dangerousspeech.org/counterspeech-tips/

CONCLUSION

The Dangerous Speech ideas offered in this chapter have been used in countries as varied as Nigeria, Sri Lanka, Denmark, Hungary, Kenya, Pakistan, and the United States, in two basic ways that seem promising. First, it's useful to collect and study Dangerous Speech systematically, looking for changes in its nature and volume over time, since this can serve as an early warning for violence. Second, it's valuable to find the most effective ways to diminish Dangerous Speech or its harmful effects – without impinging on freedom of speech. We have made efforts of both kinds and look forward to continuing, with colleagues in many countries where, unfortunately, the topic is all too relevant.

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.

We warmly welcome critique and feedback on the ideas offered above. To contact us, please visit dangerousspeech.org/contact

Contributors to this Guide
Susan Benesch, Founder and Executive Director
Cathy Buerger, Senior Researcher
Tonei Glavinic, Director of Operations
Sean Manion, Communications Fellow

Acknowledgments
We are very grateful to many people who have made invaluable contributions to our thinking, and therefore to this Guide. Many are mentioned in the text, but some must remain anonymous for their safety. We are especially grateful to those who are working in interesting and innovative ways to undermine Dangerous Speech around the world.

We also wish to thank the John D. and Catherine T. MacArthur Foundation, whose support made this report possible.

Design by CstudioDesign.com

REFERENCES

Abdelkader, E. (2014). Savagery in the Subways: Anti-Muslim Ads, the First Amendment, and the Efficacy of Counterspeech. *Asian American Law Journal.* 21. pp.43-87.

Allen, R. (2017). What Happens Online in 60 Seconds. *Smart Insights*. Available at: https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/. [Accessed 10 Oct. 2018].

Austin, J. L. (1962). *How to do Things with Words*. Cambridge: Harvard University Press.

Beech, H. (2017). Across Myanmar, Denial of Ethnic Cleansing and Loathing of Rohingya. *The New York Times*.Available at: https://www.nytimes.com/2017/10/24/world/asia/myanmar-rohingya-ethnic-cleansing.html. [Accessed 10 Oct. 2018].

Benesch, S. (2003). Vile Crime or Inalienable Right, Defining Incitement to Genocide. *Virginia Journal of International Law, 48*(3), pp. 485-528.

Benesch, S. (2013). Dangerous Speech: A Proposal to Prevent Group Violence. *Dangerous Speech Project.* Available at: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf. [Accessed 10 Oct. 2018].

Benesch, S. (2014). Countering dangerous speech to prevent mass violence during Kenya's 2013 election. Available at: https://dangerousspeech.org/countering-dangerous-speech-kenya-2013/. [Accessed 10 Oct. 2018].

Benesch, S. (2015). Charlie the Freethinker: Religion, Blasphemy, and Decent Controversy. *Religion & Human Rights 10*, pp. 244-254.

Berry, J.A. and Berry, C.P. eds. (1999). *Genocide in Rwanda: A collective memory*. Howard University Press. pp. 113–115.

Berwick, A. (2011). 2083: A European Declaration of Independence. Available at: https://publicintelligence.net/anders-behring-breiviks-complete-manifesto-2083-a-european-declaration-of-independence/. [Accessed 10 Oct. 2018].

Bloom, P. (2017). The Root of All Cruelty. *The New Yorker.* Available at: https://www.newyorker.com/magazine/2017/11/27/the-root-of-all-cruelty [Accessed 14 Dec. 2018].

Brandenburg v. Ohio (1969), 395 U.S. 444. Available at: https://cdn.loc.gov/service/ll/usrep/usrep395/usrep395444/usrep395444.pdf [Accessed 9 Oct. 2018].

Brcak, N. and Pavia, J.R. (1994). Racism in Japanese and US Wartime Propaganda. *Historian*, *56*(4), pp. 671-684.

Brown, R. (2016). *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech.* Available at: https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf [Accessed 25 Sept. 2018].

Brunsden, J. (2017). Europe refugee policy is 'Trojan horse of terrorism', says Orban. *Financial Times*. Available at: https://www.ft.com/content/538b2a0a-154e-11e7-80f4-13e067d5072c. [Accessed 10 Oct. 2018].

Burleigh, M. (2015). Migrants could be biggest threat to Europe since the war. *Daily Mail Online*. Available at: https://www.dailymail.co.uk/news/article-3141005/Tidal-wave-migrants-biggest-threat-Europe-war.html.

Bytwerk, R. (n.d.). Caricatures from Der Stürmer: 1927-1932. *German Propaganda Archive.* Available at: http://research.calvin.edu/german-propaganda-archive/sturm28.htm. [Accessed 9 Oct. 2018].

Chen, A. (2015). Unfollow: How a prized daughter of the Westboro Baptist Church came to question its beliefs. *New Yorker*. Available at: http://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper [Accessed 9 Oct. 2018].

Chirot, D. and McCauley, C. (2010). *Why not kill them all?: The logic and prevention of mass political murder*. Princeton, NJ: Princeton University Press.

Crushing, T. (2018). For The Second Time In A Week, German Hate Speech Laws Results In Deletion Of Innocent Speech. *Techdirt.* Available at: https://www.techdirt.com/articles/20180111/15543538989/second-time-week-german-hate-speech-laws-results-deletion-innocent-speech.shtml [Accessed 9 Oct. 2018].

Data Team, The (2017)."Donald Trump is crushing it on Twitter." (2017). *The Economist.* Available at: https://www.economist.com/graphic-detail/2017/11/10/donald-trump-is-crushing-it-on-twitter [Accessed 9 Oct. 2018].

"Declaration on the Elimination of Violence against Women." 1993. United Nations General Assembly. Available at: http://www.un.org/documents/ga/res/48/a48r104.htm [Accessed 13 Dec. 2018].

Des Forges, A. (1999). *"Leave none to tell the story:" Genocide in Rwanda*, New York, New York: Human Rights Watch. Available at: https://www.hrw.org/reports/1999/rwanda/ [Accessed 10 Oct. 2018].

Elliott, J. (2018). "India WhatsApp killings: Why mobs are lynching outsiders over fake videos." *Global News.* Available at: https://globalnews.ca/news/4333499/india-whatsapp-lynchings-child-kidnappers-fake-news/ [Accessed 25 Sept. 2018].

Ellman, M. (2005). The role of leadership perceptions and of intent in the Soviet Famine of 1931–1934. *Europe-Asia Studies*, *57*(6), pp.823-841.

Facebook, Inc. (2018). *Community Standards*. Available at: https://www.facebook.com/communitystandards/ [Accessed 9 Oct. 2018].

Figes, O. (2008). *The Whisperers: Private life in Stalin's Russia.* 2nd ed., New York: Metropolitan Books.

Galtung, J. (1969). Violence, peace, and peace research. *Journal of Peace Research*, *6*(3), pp. 167-191.

Gowen, A. (2017). "We are going to kill you": Villagers in Burma recount violence by Rohingya Muslim militants. *Washington Post.* November 15, 2017. Available at: https://www.washingtonpost.com/world/asia_pacific/we-are-going-to-kill-you-villagers-in-burma-recount-violence-by-rohingya-muslim-militants/2017/11/14/409ff59b-849d-4459-bdc7-d1ea2b5ff9a6_story.html [Accessed 20 Sept. 2018].

Handy, J. (1984). *Gift of the Devil: a History of Guatemala*, Boston, Massachusetts: South End Press.

Harris, B. (1999). Guatemala: Bill Clinton's Latest Damn-Near Apology. *Mother Jones*. Available at: https://www.motherjones.com/politics/1999/03/guatemala-bill-clintons-latest-damn-near-apology/. [Accessed 22 Sept. 2018].

Hashim, A. (2017). Three Ahmadis sentenced to death for blasphemy. Al Jazeera. Available at: https://www.aljazeera.com/news/2017/10/ahmadis-sentenced-death-blasphemy-171012081709423.html [Accessed 10 Oct. 2018].

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review, 10*(3), 252-264.

Hinton, A.L. (2005). *Why did they kill? Cambodia in the shadow of genocide*, Berkeley, California: University of California Press.

Houreld, K. (2014). Teenager kills man accused of blasphemy in Pakistan police station. Reuters. Available at: https://www.reuters.com/article/us-pakistan-blasphemy-killing/teenager-kills-man-accused-of-blasphemy-in-pakistan-police-station-idUSBREA4F0HI20140516 [Accessed 10 Oct. 2018].

Jan, T. and Dwoskin E., (2017). "A white man called her kids the n-word. Facebook stopped her from sharing it," *Washington Post.* July 31. Available at: https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html?utm_term=.34f332f66679 [Accessed 9 Oct. 2018].

Janowitz, M. (1946). German reactions to Nazi atrocities. *American Journal of Sociology*, *52*(2), pp.141-146.

Johnson, M.S. (1998). Gender, Race, and Rumours: Reexamining the 1943 Race Riots. *Gender & History*, *10*(2), pp. 252-277.

Khan, A.M. (2003). Persecution of the Ahmadiyya community in Pakistan: An analysis under international law and international relations. *Harvard Human Rights Journal 16*, pp. 217-244.

Kiernan, B. (2009). *Blood and Soil: A World History of Genocide and Extermination from Sparta to Darfur* 2nd ed., New Haven, Connecticut: Yale University Press.

Kimani, M. (2007). RTLM: the Medium that Became a Tool for Mass Murder. In L. Waldorf and A. Thompson, eds. *The Media and the Rwandan Genocide.* 1st ed. London: Pluto Press.

Kopan, T., (2015). Donald Trump: Syrian refugees a 'Trojan horse'. *CNN*. Available at: https://www.cnn.com/2015/11/16/politics/donald-trump-syrian-refugees/index.html [Accessed 10 Oct. 2018].

Kottasová, I. (2018). *Is Germany's new hate speech law killing press freedom?* CNN. January 4. Available at: https://money.cnn.com/2018/01/04/media/twitter-satire-free-speech-germany/index.html. [Accessed 10 Oct. 2018].

Kugelman, M. (2017) Why Pakistan hates Malala. *Foreign Policy.* Available at: https://foreignpolicy.com/2017/08/15/why-pakistan-hates-malala/ [Accessed 17 Dec. 2018].

Kuper, L. (1981). *Genocide: Its political use in the twentieth century*. 1st ed., New Haven, Connecticut: Yale University Press.

Lane, J. (2014). 'Be afraid. Be very afraid: Exploring the rhetoric of the monster in political and horror posters of the 20th century', Edith Cowan University, Perth, Australia. Available at: https://ro.ecu.edu.au/theses_hons/198/ [Accessed 17 Dec. 2018].

Larimer, S. (2016). The case of the veterinarian who shot a cat with a bow and arrow, then posed with its body. *Washington Post*. Available at: https://www.washingtonpost.com/news/animalia/wp/2016/10/19/the-case-of-the-veterinarian-who-shot-a-cat-with-a-bow-and-arrow-then-posed-with-its-body/?utm_term=.8c9a7201813e [Accessed 9 Oct. 2018].

Lehman, J. (2010). A brief explanation of the Overton window. *Mackinac Center for Public Policy.* Available at: https://www.mackinac.org/overtonwindow [Accessed 17 Dec. 2018].

Leader Maynard, J. and Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal*, *9*(3).

Leader Maynard, J. (2015). "Combating Atrocity-Justifying Ideologies," in Sharma, S.K. and Welsh, J.M. eds., The responsibility to prevent: overcoming the challenges of atrocity prevention. Oxford: Oxford University Press.

Leader Maynard, J. (2014). Rethinking the role of ideology in mass atrocities. *Terrorism and Political Violence*, *26*(5), pp.821-841.

Lovett, I. (2012). Man Linked to Film in Protests Is Questioned. *The New York Times*. Available at: https://www.nytimes.com/2012/09/16/world/middleeast/man-linked-to-film-in-protests-is-questioned.html?_r=1&ref=internationalrelations. [Accessed 10 Oct. 2018].

Mackie, J. (2014). A century ago, the Komagata Maru arrived in Vancouver and challenged racist policies. *Vancouver Sun*. Available at: http://www.vancouversun.com/life/century+Komagata+Maru+arrived+Vancouver+challenged+racist+policies/9868708/story.html. [Accessed 10 Oct. 2018].

Manion, S. & Benesch, S. (2018). WhatsApp Responds after Lynchings Linked to Dangerous Speech in India. *Dangerous Speech Project*. Available at: https://dangerousspeech.org/whatsapp-updates-highlight-indias-problem-with-dangerous-fake-news/. [Accessed 10 Oct. 2018].

Marsden, S. (2013). Internet troll who abused Mary Beard apologises after threat to tell his mother. *The Telegraph*. Available at: https://www.telegraph.co.uk/news/uknews/law-and-order/10209643/Internet-troll-who-abused-Mary-Beard-apologises-after-threat-to-tell-his-mother.html [Accessed 9 Oct. 2018].

Martin, D. (2018). German satire magazine Titanic back on Twitter following 'hate speech' ban. *Duetsche Welle.* January 6. Available at: https://www.dw.com/en/german-satire-magazine-titanic-back-on-twitter-following-hate-speech-ban/a-42046485. [Accessed 10 Oct. 2018].

McCrummen, S. (2008). No Quick Fix for What Still Ails Kenya. *The Washington Post Foreign Service*. Available at: http://www.washingtonpost.com/wp-dyn/content/article/2008/03/06/AR2008030603766.html?sid=ST2008030603799. [Accessed 10 Oct. 2018].

Mitigating Dangerous Speech: Monitoring and Countering Dangerous Speech to Reduce Violence. (2017). Available at: http://www.nsrp-nigeria.org/wp-content/uploads/2017/12/NSRP-How-to-Guide-Mitigating-Hate-and-Dangerous-Speech.pdf. [Accessed 10 Oct. 2018].

Mugesera *v*. Canada (Minister of Citizenship and Immigration), [2005] 2 S.C.R. 100, 2005 SCC 40. Available at: https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/2273/index.do. [Accessed 10 Oct. 2018].

Netherlands Embassy. (2017) 29 November. Available at: https://twitter.com/NLintheUSA/status/935953115249086464?ref_src=twsrc%5Etfw. [Accessed 10 Oct. 2018].

The General Civil Penal Code (Act No. 10 of May 22, 1902, as last amended by Act No. 131, Dec. 21, 2005), University of Oslo Law Library Translated Norwegian Legistlation online database, https://app.uio.no/ub/ujur/oversatte-lover/data/lov-19020522-010-eng.pdf [Accessed 10 Oct. 2018].

Ogundipe, S. (2016). Kaduna lecturer detained for alleged 'hate speech' released. *Premium Times.* November 7. Available at: https://www.premiumtimesng.com/news/top-news/214706-breaking-kaduna-lecturer-detained-alleged-hate-speech-released.html [Accessed 25 Sept. 2018].

OHCHR (2018). "Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar." United Nations Human Rights Council. Available at: https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf [Accessed 9 Oct. 2018].

Ohlheiser, A. (2017). The work of monitoring violence online can cause real trauma. And Facebook is hiring., *The Washington Post*, May 4, 2017. Available at: https://www.washingtonpost.com/news/the-intersect/wp/2017/05/04/the-work-of-monitoring-violence-online-can-cause-real-trauma-and-facebook-is-hiring/?utm_term=.de4b78129afd. [Accessed 10 Oct. 2018].

Oltermann, P. (2018). Tough New German Law Puts Tech Firms and Free Speech in Spotlight. *The Guardian*. Available at: https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight [Accessed 9 Oct. 2018].

Osborne, S. (2017). Ratko Mladic guilty: 'Butcher of Bosnia' convicted of genocide, crimes against humanity and war crimes, *Independent*. November 22. Available at: https://www.independent.co.uk/news/world/europe/ratko-mladic-guilty-genocide-verdict-latest-bosinian-war-crimes-humanity-serbian-general-a8068986.html [Accessed 9 Oct. 2018].

Paul, K. (2014). Why Are Women in Turkey Laughing? Because a Politician Told Them Not To. *Mashable.* Available at: https://mashable.com/2014/07/30/turkey-women-laugh/#fkc9ecJ1Gsq4 [Accessed 9 Oct. 2018].

PeaceTech Lab (2017). "Social media and conflict in South Sudan: A lexicon of hate speech terms." Available at: https://www.peacetechlab.org/hate-speech-in-south-sudan/ [Accessed 10 Oct. 2018].

The Penal Code, c. 20. Norway. Available at: https://lovdata.no/dokument/NLE/lov/2005-05-20-28/KAPITTEL_2#KAPITTEL_2 [Accessed 10 Oct. 2018].

Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000, c. 1. South Africa. Available at: http://www.justice.gov.za/legislation/acts/2000-004.pdf [Accessed 10 Oct. 2018].

The Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza, Hassan Ngeze (Trial Judgment). (2003) ICTR-99-52-T, International Criminal Tribunal for Rwanda (ICTR) available at: http://unictr.irmct.org/sites/unictr.org/files/case-documents/ictr-99-52/trial-judgements/en/031203.pdf. [Accessed 10 Oct. 2018].

Rebelo, K., (2017). Child Kidnapping Rumours In India Being Spread With Syria Image, Pak Video. *BOOM Live*. Available at: https://www.boomlive.in/child-kidnapping-rumours-in-india-being-spread-with-syria-image-pak-video/. [Accessed 10 Oct. 2018].

Roberts, S T (2014). Behind the screen: the hidden labor of commercial content moderators, PhD dissertation, University of Illinois at Urbana-Champaign, Available at: http://hdl.handle.net/2142/50401. [Accessed 10 Oct. 2018].

Ronson, J. (2015). So You've Been Publicly Shamed. 1st ed. London: Penguin Books.

Russian Public Opinion Research Center (2018). Conspiracy Theory Against Russia. Available at: https://wciom.ru/index.php?id=236&uid=9259 [Accessed 14 Dec. 2018].

Saldanha, A. (2017) *2017 Deadliest Year For Cow-Related Hate Crime Since 2010, 86% Of Those Killed Muslim* Available at: http://www.indiaspend.com/2017-deadliest-year-for-cow-related-hate-crime-since-2010-86-of-those-killed-muslim-12662/. [Accessed 10 Oct. 2018].

Saldanha, A., Hazare, J. & Rajput, P. (2018). Child-Lifting Rumours: 33 Killed In 69 Mob Attacks Since Jan 2017. Before That Only 1 Attack In 2012. *IndiaSpend*. Available at: http://www.indiaspend.com/child-lifting-rumours-33-killed-in-69-mob-attacks-since-jan-2017-before-that-only-1-attack-in-2012-2012/. [Accessed 10 Oct. 2018].

Saleem, H.M., Dillon, K.P., Benesch, S., and Ruths, D. (2016). A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety.* Available at: http://www.ta-cos.org/sites/ta-cos.org/files/tacos2016_SaleemDillionBeneschRuths.pdf. [Accessed 10 Oct. 2018].

Sanghvi, V. (2018). India's Lynching App: Who is Using WhatsApp as a Murder Weapon? *SCMP: This Week in Asia.* July 9. Available at: https://www.scmp.com/week-asia/society/article/2154436/indias-lynching-app-who-using-whatsapp-murder-weapon [Accessed 28 Sept. 2018].

Savage, D.G. (2011). U.S. official cites misconduct in Japanese American internment cases. *Los Angeles Times*. Available at: http://articles.latimes.com/print/2011/may/24/nation/la-na-japanese-americans-20110525 [Accessed 25 Sept. 2018].

Shachtman, N. & Beckhusen, R. (2012). Anti-Islam Filmmaker Went by 'P.J. Tobacco' and 13 Other Names. *Wired*. Available at: https://www.wired.com/2012/09/anti-islam-flick/. [Accessed 10 Oct. 2018].

Shahani, A. (2016). With 'Napalm Girl,' Facebook Humans (Not Algorithms) Struggle To Be Editor. *National Public Radio.* September 10. Available at: https://www.npr.org/sections/alltechconsidered/2016/09/10/493454256/with-napalm-girl-facebook-humans-not-algorithms-struggle-to-be-editor [Accessed 9 Oct. 2018].

Shen, C., Kasra, M., Pan, W., Bassett, G.A., Malloch, Y., and O'Brien, J.F. (2018). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media and Society*. Available at: doi:10.1177/1461444818799526. [Accessed 10 Oct. 2018].

Sémelin, J. (2014). *Purify and destroy: the political uses of massacre and genocide*, London: Hurst & Company.

Simões, M. (2018). Brazil's polarizing new president, Jair Bolsonaro, in his own words. *The New York Times.* Available at: https://www.nytimes.com/2018/10/28/world/americas/brazil-president-jair-bolsonaro-quotes.html [Accessed 17 Dec. 2018].

Smith, D.L. (2011). *Less than human: Why we demean, enslave, and exterminate others*. New York City: St. Martin's Press.

Smith, D.L. (2018). Donald Trump, Dangerous Speech, and the Legacy of White Supremacist Terrorism. *Dangerous Speech Project*. Available at: https://dangerousspeech.org/donald-trump-dangerous-speech-and-the-legacy-of-white-supremacist-terrorism/. [Accessed 10 Oct. 2018].

Specia, M. and Mozur, P. (2017). A War of Words Puts Facebook at the Center of Myanmar's Rohingya Crisis. *The New York Times,* October, 27. Available at: https://www.nytimes.com/2017/10/27/world/asia/myanmar-government-facebook-rohingya.html. [Accessed 10 Oct. 2018].

Stecklow, S. (2018). Why Facebook is losing the war on hate speech in Myanmar. *Reuters.* August 15. Available at: https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/. [Accessed 10 Oct. 2018].

Straus, S. (2007). What Is the Relationship between Hate Radio and Violence? Rethinking Rwanda's "Radio Machete". *Politics & Society 35*(4), pp. 609-637. Available at: doi:10.1177/0032329207308181.

Straus, S. (n.d.) African Presidential Speeches Database. University of Wisconsin. Available at: https://faculty.polisci.wisc.edu/sstraus/african-presidential-speeches-database/ [Accessed 10 Oct. 2018].

Streicher, J. (1934). Jüdischer Mordplan. *Der Stürmer*.

Thomas, J. (2001). Behind a Book That Inspired McVeigh. *The New York Times*. Available at: https://www.nytimes.com/2001/06/09/us/behind-a-book-that-inspired-mcveigh.html. [Accessed 25 Sept. 2018].

Thuku, W. (2014). ICC Witness: William Ruto never said 'madoadoa.' *Standard Digital.* Available at: https://www.standardmedia.co.ke/article/2000105550/icc-witness-william-ruto-never-said-madoadoa. [Accessed 10 Oct. 2018].

True, E. (2014). The gaming journalist who tells on her internet trolls – to their mothers. The Guardian. Available at: https://www.theguardian.com/culture/australia-culture-blog/2014/nov/28/alanah-pearce-tells-on-her-internet-trolls-to-their-mothers [Accessed 9 Oct. 2018].

Twitter, Inc. (2018). *The Twitter Rules*. Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules [Accessed 9 Oct. 2018].

Weitz, E. D. (2015) *A Century of Genocide: Utopias of Race and Nation* - Updated Edition. 2nd edn. Princeton: Princeton University Press.

WhatsApp. (2018). More Changes to Forwarding. WhatsApp Blog. Available at: https://blog.whatsapp.com/10000647/More-changes-to-forwarding. [Accessed 10 Oct. 2018].

Wilson, J. (2018). 'Dripping with poison of antisemitism': the demonization of George Soros. *The Guardian.* Available at: https://www.theguardian.com/us-news/2018/oct/24/george-soros-antisemitism-bomb-attacks [Accessed 17 Dec. 2018].

Wright, L., Ruths, D., Dillon, K.P., Saleem, H.M., and Benesch, S.. (2017). Vectors for Counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pp. 57-62.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics*, *129*(4), pp.1947-1994.

Zephoria Digital Marketing. (2018) *The Top 20 Valuable Facebook Statistics – Updated September 2018*. Available at: https://zephoria.com/top-15-valuable-facebook-statistics/ [Accessed 9 Oct. 2018].

Zikri, A.B. (2018). Netanyahu defends Egypt border fence: Influx of African migrants more dangerous than terrorism. *Haaretz*. Available at: https://www.haaretz.com/israel-news/.premium-netanyahu-danger-posed-by-african-migrants-is-greater-than-terrorism-1.5930984. [Accessed 10 Oct. 2018].