Dangerous speech

Benesch, Susan

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Digital
Communication
Research.de

Recommended citation: Benesch, S. (2023). Dangerous speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 185–197). Digital Communication Research. https://doi.org/10.48541/dcr.v12.11

Abstract: The concept of "dangerous speech," which I proposed in the early 2010s, illuminates a key fact that is often missed: hate speech (and related categories like toxic and extreme speech) affects people gradually, cumulatively, and by dint of repetition. Dangerous speech is defined based on the specific harm it engenders (inspiring intergroup violence) rather than its content alone or the intent of those who spread it, allowing for a more consistent definition and broader consensus that it should be addressed. In this article, I explain why this concept is useful; describe the five aspects of speech that must be evaluated in order to determine dangerousness; share examples of projects that have been conducted to monitor, evaluate, and counteract dangerous speech; and suggest future avenues for research.

DOI 10.48541/dcr.v12.11

Susan Benesch

Dangerous Speech[1]

1 Dangerous speech as a practical tool for research

Two facts about hate speech (and related categories, including toxic, extreme, and dangerous speech) are vital for understanding its effects, and should inform research. The first is that hatred is not innate; it must be taught. Hate speech should be conceptualized as a tool for doing that, in order to learn how to prevent or reverse the teaching process.

The second fact is that hate speech affects people gradually, cumulatively, and by dint of repetition.[2] This point is often neglected by scholars and content moderation policymakers who work on hate speech, but it is well known to those who suffer from it. One of the latter, a witness in a trial at the United Nations Tribunal for Rwanda, after the 1994 genocide, described it brilliantly. Explaining how a radio station had groomed its listeners to commit and condone unthinkable violence, the witness said, "In fact, what RTLM [Radio Télévision Libre des Mille Collines] did was almost to pour petrol, to spread petrol throughout the country

1 The author is very grateful to her colleague Tonei Glavinic for contributing invaluable research and editing to this chapter.

2 See e.g., Hasher et al. (1977), which coined the term "illusory truth effect" to describe the phenomenon of people coming to believe a falsehood after hearing it repeatedly stated as fact.

little by little, so that one day it would be able to set fire to the whole country" (Prosecutor v. Nahimana, 2003, para. 436).

The concept of "dangerous speech" (Dangerous Speech Project [DSP], 2021), which I proposed, incorporates this phenomenon of gradual norm change, allowing for study that more clearly depicts human experience than hate speech and similar categories, and permitting more sensitive monitoring for increased risk of violence. Dangerousness, in this formulation, is the capacity of a speech act[3] – as disseminated—to inspire violence against members of another human group. Dangerous speech is defined by the specific harm[4] it engenders, not by its content alone, nor by the intent or motives of the people who produce and spread it. This makes for more consistent definition, and more consensus against this kind of speech, since it is very difficult for people to agree on which content is offensive, but easy for them to agree that mass intergroup violence should be prevented.

To capture the variable impact of speech acts, or "drops of petrol," dangerous speech is not a binary concept; dangerousness falls on a spectrum. Speech (including words, sounds, and images) can be more or less dangerous, depending on characteristics including the means by which it was disseminated, the speaker or source, the audience, the message itself, and the social and historical context in which the speaker and audience find themselves. The social context includes previous dangerous messages, which slowly shift people's states of mind so that they become more susceptible to the next such message, which is therefore more dangerous.

It is also important to note that repeated exposure to dangerous speech can convince members of an audience that such ideas are widely accepted by the people around them, even if they do not believe or accept the messages themselves. In other words, dangerous speech can shift norms, and people eagerly comply with norms to maintain good standing in a group (Leader Maynard, 2014).

3 In language theory a "speech act" is any form of 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 provoke a response in its audience. Dangerous speech is defined by such force: its capacity to inspire violence.

4 For more on the wide variety of harms speech can engender, and an argument that for robust research and policymaking, it is important to categorize speech by harms, not only by content, see Benesch (2020).

2 Dangerous speech and hate speech

Dangerous speech is a narrower and more precisely bounded category than hate speech, the most prevalent term in academic literature and common discourse (see also Sponholz and Frischlich in this volume). Although some hate speech is explicit and all too easy to identify as such, as a category it is large and contested, with blurry boundaries. We lack consensus on how to define it in law,[5] scholarly literature, common parlance, and even in the rules under which internet companies prohibit some content—and permit the rest.[6]

The term hate speech itself presents important questions that have not yet been consistently answered. First, must hate speech express hatred, promote hatred, or make someone feel hated? For example, if asked whether a drawing of the Prophet Mohammed constitutes hate speech, should one consider the intention of the person who made the drawing, or of someone else who disseminated it, or its effect on some or all of the people who see it or hear about it?[7] If it is the intention of the author that is definitive, the state of another person's mind is not always easy to discover, especially when its expression is found online.

Moreover, if hate speech is related to hatred, what exactly is that? How strong or how durable must emotion be to count as hatred?

One point that is clear, paradoxically, is that "I hate you," no matter how vehemently or sincerely expressed, is generally not hate speech (European Commission, 2018, p. 2), since a common thread among definitions is that hate speech denigrates or attacks a person or people *due to some characteristic or identity that they share* with other people, such as race, religion, nationality, sexual orientation, gender, age, caste, immigrant status, or disability. Most definitions list some but not all of these characteristics, which has generated disagreement over which kinds of groups *count* as targets of hate speech. The United Nations wisely avoided

5 For details on the variety of definitions for hate speech, see Benesch (2014, p. 20); also Herz & Molnar (2012, p. 81).

6 See e.g., Facebook (2021); Google (2021); Twitter (2021).

7 For key relevant ideas, e.g., on the distinction between giving offense and taking offense, see George (2016). For description of the overlooked role of 'malevolent bridge figures,' or people who transmit content from one normative community in which it is offensive or controversial, to another in which it is highly inflammatory, see Benesch (2015).

this problem in a new definition of hate speech that it introduced in May 2019, by giving a non-exhaustive list of group characteristics—"religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (2019, p. 2). Unfortunately the same definition is vague and overbroad in another way, by describing hate speech as "pejorative" language with no explanation or limitation of that term. The full UN definition is this:

> Any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or group on the basis or who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor (United Nations, 2019, p. 2).

Definitional problems arise with other categories of speech as well. What may be called extremist content, for example, can be thus identified because it was produced by extremists, or because it depicts and endorses gore, or because—to the contrary—it is designed to recruit for extremist groups, by falsely describing a safe and satisfying life within them or simply by criticizing life outside those groups in ways that are compelling and convincing to certain audiences, such as lonely, frustrated youth. Though it might not be wrong to label all such content "extremist," it would be a mistake to use the same method to try to identify, study, or protect people from all of it.

3 Defining and identifying dangerous speech

I coined the term "dangerous speech" after noticing patterns in public speech during the months and years before mass violence happened in many parts of the world and in many historical periods. Political, religious, and cultural leaders' language tends to be similar during such times, from case to case, country to country, culture to culture, religion to religion, even from one historical period to another. If people could learn to identify the hallmarks of speech that seems to increase the risk of violence, then perhaps one could decrease that kind of violence. That is the most powerful reason for doing research on this kind of speech, and in my view it is reason enough.

Rhetoric alone cannot make speech dangerous, though; the context in which it is communicated is equally important. One can systematically capture many

features of that context, and analyze speech for dangerousness, by asking about five aspects of the speech as it was disseminated. This analysis can capture the cumulative effect of speech mentioned above, and the fact that dangerousness is relative; speech can be more or less dangerous. Here we mention the five aspects, with examples of the questions one would ask regarding each of them:

- Speaker: Did the message come from an influential speaker? What is the source of that influence: for example cultural, social, or religious status; public office; access to troops or another means of threatening force; charisma?
- Audience: Was the audience susceptible to inflammatory messages, e.g., because they were already fearful? Have they already been exposed to similar messages over time, like the drops of petrol described by the Rwandan witness?
- Message: Does the speech carry hallmarks of dangerous speech? These are the rhetorical patterns that my colleagues and I have identified in many examples of public speech before outbreaks of violence (DSP, 2021). A hallmark is not sufficient to identify dangerous speech on its own, since speech cannot be dangerous if an audience is not moved by it, and some audiences are resistant, fortunately. The hallmarks we have identified thus far include (DSP, 2021, pp. 12-19):
    - *Dehumanization*. Describing other people in ways that deny or diminish their humanity, for example by comparing them to disgusting or deadly animals, insects, bacteria, or demons. This makes violence seem acceptable.
    - *"Accusation in a mirror."* Asserting that the audience faces serious and often mortal threats from the target group—in other words, reversing reality by suggesting that the victims of, e.g., a genocide will instead commit it. The term "accusation in a mirror" was found in a mimeographed guide for making propaganda, discovered in Rwanda after the 1994 genocide (Des Forges, 1999, pp. 65-66). Accusation in a mirror makes violence seem necessary by convincing people that they face a mortal threat, which they can fend off only with violence. This is a very powerful rhetorical move since it is the collective analogue of the one ironclad defense to murder: self-defense. If people feel violence is necessary for defending themselves, their group, and especially their children, it seems not only justified but virtuous.

- *Assertion of attack on women/girls.* Suggesting that women or girls of the audience's group have been – or will be – threatened, harassed, or defiled by members of a target group. In many cases, the purity of a group's women is symbolic of the purity of the group itself, or of its identity or way of life.
- *Coded language.* Including phrases and words that have a special meaning, shared by the speaker and audience. The speaker is therefore capable of communicating two messages, one understood by those with knowledge of the coded language and one understood by everyone else. This can make the speech more dangerous in a few ways. For example, the coded language could be deeply rooted in the audience members' sense of identity or shared history and therefore evoke disdain for an opposing group. It can also make the people who use the term feel that they are more strongly bound to the group of other people who use the code, like a password or a special handshake. Finally, coded speech can be harder to identify and counter for those who are not familiar with it, including social media company staff. One example of coded language – or symbols – is the use of a pineapple to mock and denigrate Jews in France (Nelson, 2015). Another is the name of the mobile phone company "MTN" which has been used as a powerful slur against Dinka people in South Sudan because of the company's slogan "Everywhere you go," understood as a coded reference to the claim that the Dinka invaded the lands of other groups (Patinkin, 2017).
- *Impurity/contamination.* Giving the impression that one or more members of a target group might damage the purity or integrity or cleanliness of the audience group. Members of target groups have been compared to rotten apples that can spoil a whole barrel of good apples, weeds that threaten crops, or stains on a dress, for example.

- Context: Is there a social or historical context that has lowered the barriers to violence or made it more acceptable? Examples of this are competition between groups for resources and previous episodes of violence between the relevant groups.
- Medium: How influential is the medium by which the message is delivered? For example, is it the only or primary source of news for the relevant audience?

All five conditions need not be relevant for speech to be dangerous. For example, a message can be dangerous even when the speaker is anonymous. Only two conditions are necessary: the message must be inflammatory, and the audience must be susceptible.

4 Studying dangerous speech and efforts to counter it

The idea of dangerous speech can be useful for research of several kinds, of which the first is to collect and analyze examples of it, as a way of understanding whether and where violence may occur, and to inform violence prevention efforts.

This type of monitoring began in Kenya, leading up to that country's 2013 national elections—a tense moment since the previous national election campaign brought months of dangerous speech and was followed by mass violence in 2007, in which more than 1,200 people were killed and more than 600,000 were displaced from their homes. I worked with Ushahidi and iHub Research on their Umati project[8], in which teams of full-time monitors manually collected examples of vitriolic speech in six different languages. Their codebook built upon the contextual factors of dangerous speech outlined above by directing coders to consider each of the five factors for each speech act and then classify it as offensive, moderately dangerous, or (very) dangerous (Awori, 2013; DSP, 2016). The Umati codebook distilled that process for coders, and guided them through it, by asking them two questions about the speaker and about the content itself: "On a scale of 1 to 3 with 1 being little influence and 3 being a lot of influence, how much influence does the speaker have on the audience?" and "On a scale of 1 to 3, with 1 being barely inflammatory and 3 being extremely inflammatory, how inflammatory is the content of the text?" This method was inventive, and it may have increase inter-rater reliability, but it was of limited use when the speaker was unknown, which is quite often the case for online speech.

Coding questions arose frequently, and the team held regular meetings to consider and resolve them. The meetings were lessons in how varied hateful and inflammatory speech is, and how important context can be, for understanding it. In one example, a coder identified the sentence "I hate Raila" (Odinga, one of the

8 Umati means "crowd" in the Kenyan national language of Kiswahili.

leading presidential candidates) as dangerous speech. I said this was neither hate speech nor dangerous speech, since it was directed only at an individual, without reference to any group. The coder replied unequivocally that in the Kenyan context of that time, to say "I hate Raila" was also to say that the speaker hated Luos, the ethnic group of which Odinga was a leader.

In 2015, the Nigerian Centre for Information Technology and Development (CITAD, 2016) monitored online speech during and after Nigeria's 2015 election campaign, building on the Umati model.

Another promising body of research related to dangerous speech is studies on efforts to counter its harmful effects, focusing on whether and how they succeed. This is nascent, since only a few such projects have been carried out, so far without being rigorously studied. For example, Umati led to two efforts to counter dangerous speech during Kenya's national electoral campaign of 2013, by "inoculating" the public against such speech, i.e., teaching people that it is a tool used by unscrupulous leaders to manipulate them. One of those was studied. The first effort was called Nipe Ukweli, or "gimme truth" in Kenyan slang—a name reflecting the fact that much dangerous speech is also disinformation (DSP, 2016). This project consisted of flyers and community meetings that explained dangerous speech and encouraged people to report it to the Umati team. In the second effort, four episodes of a legendary, well-known Kenyan television courtroom drama called Vioja Mahakamani[9] focused on dangerous speech with a similar goal: to teach the audience to recognize it and to be skeptical of it. The Vioja Mahakamani project was independently evaluated for impact on audiences, by researchers from the University of Pennsylvania, with encouraging results. Focus groups conducted with young Kenyans from various ethnic backgrounds revealed that those who watched the dangerous speech episodes demonstrated a greater understanding of the origins, motivations, and consequences of incitement compared to those in control groups who watched unrelated episodes of the show (Kogen, 2013). More such studies are clearly needed, though they can be difficult to conduct, either because there are confounding third factors that make robust evaluation difficult, or because researchers cannot get sufficient data.

9 Vioja Mahakamani means "events in the courtroom." The four episodes were collectively designed by the actors who made them, after they all attended a workshop on dangerous speech.

Technology companies are one important type of stakeholder for such research, of course, since they have enormous power and capacity to experiment with methods to improve online behavior and norms. In response to public pressure and legal requirements, especially those with social media platforms, tech companies are increasingly trying out new techniques to try to more effectively identify and deal with hate speech, dangerous speech, and other harmful content on their online turf. While removing such content is the most visible remedy, it is a heavy-handed approach, and there are many alternatives that better protect freedom of expression but need to be better understood, including downranking content (reducing its algorithmic amplification), "nudging" users to reconsider their words before they are posted (Diaz, 2021), and proactively reminding users of the rules they must follow (Benesch & Matias, 2018). Yet it is rare that these efforts are A/B tested to see which is more effective—and virtually all research at companies is under nondisclosure agreements. Conducting independent, ethical, transparent, privacy-protecting research—either in cooperation with companies or in spite of them—and publishing it in reputable peer-reviewed journals would be a major step toward greater understanding of how to meaningfully address harmful online content.

There are also significant efforts in civil society to tackle hate speech and dangerous speech online, presenting many research opportunities that have not yet been seized. Anthropologist Cathy Buerger (2020) published the first in-depth, ethnographic exploration of #jagärhär, a thousands-strong network of volunteers in Sweden who launch coordinated responses to hate speech and dangerous speech on Facebook. That project is unusual not only for its large size (more than 70,000 people are members of the Swedish group alone) but for the fact that it is still going strong several years after its founding, and also for its replication in other countries (there are 16 groups operating in various countries, at this writing). Buerger (2020) interviewed 25 of the most active Swedish participants, many of whom said they observed favorable shifts in discourse norms in the spaces where they have intervened online.

We have also identified dozens of smaller anti-hate efforts in many countries. Activists, journalists, clergy, lawyers and others have been experimenting with quite a variety of methods, including some that deliberately amplify hateful or offensive content to force members of a society to accept that it is there and reckon with the racism and hatred it expresses. Of course technology plays a role in many of these efforts: just as new communications technologies are being used to amplify

inflammatory hate speech, they can also be marshalled to prevent and counter it. New technologies are also being employed to detect where dangerous speech may signal an increased risk of mass violence, and social media companies sometimes delete such content, or downrank it, as noted above (Facebook, 2020, p. 7).

Companies have so far missed other opportunities to detect dangerous speech, such as observing the way in which members of the public respond—in open online spaces—to the posts of unscrupulous leaders. This could give invaluable clues in many cases, without violating privacy or causing other harms. For example, in mid-December 2020, Donald Trump invited his followers to come to Washington, D.C. for a rally, on January 6, 2021. Though he wrote only that the rally "will be wild," many of his followers understood his ambiguous language as a call to violence, by telling each other that he wanted them to come with firearms, ready to use them. I have developed this idea in another article (Benesch, 2021).

In sum, dangerous speech is worth special attention from researchers for several reasons. First, it seems to be linked to intergroup violence, and therefore it may serve as a good early warning signal. Perhaps violence can be prevented, at least in part, if dangerous speech can be defanged or diminished without causing other harms (like infringing on freedom of expression). Second, dangerous speech is a more precise and less contested category than others like hate speech, so it should be possible to build comparable datasets of it from a variety of places or social groups. Transnational study is exceedingly rare in the literature on hate speech, and would be of great interest. Also, though the dangerousness of speech depends greatly on context, which cannot be detected and evaluated automatically, it may be possible to build classifiers for dangerous speech that operate by detecting similarities and patterns in it.

Finally, the concept of dangerous speech accommodates the fact that inciting language has a cumulative effect on people. This is key to understanding the capacity of speech to inspire behavior, but it has so far received scant attention. I hope the literature will soon grow in these areas.

Susan Benesch is founder and director of the Dangerous Speech Project, Faculty Associate of the Berkman Klein Center for Internet & Society at Harvard University, and Adjunct Associate Professor at American University's School of International Service, USA.

References

Austin, J. L. (1962). *How to do things with words.* Oxford University Press.

Awori, K. (2013, June). *Umati final report*. iHub Research. https://dangerousspeech.org/wp-content/uploads/2017/05/Umati-Final-Report.pdf

Benesch, S. (2015). Charlie the freethinker: Religion, blasphemy, and decent controversy. *Religion & Human Rights, 10*(3), 244–254. https://doi.org/10.1163/18710328-12341291

Benesch, S. (2020, June). *Proposals for improved regulation of harmful online content*. Dangerous Speech Project. https://dangerousspeech.org/wp-content/uploads/2020/06/Proposals-for-Improved-Regulation-of-Harmful-Online-Content-Formatted-v5.2.1.pdf

Benesch, S. (2021, March 4). *The insidious creep of harmful rhetoric*. Noēma. https://www.noemamag.com/the-insidious-creep-of-violent-rhetoric/

Benesch, S., & Matias, J. N. (2018, April 6). *Launching today: new collaborative study to diminish abuse on Twitter*. Medium. https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc

Buerger, C. (2020, December 14). *The anti-hate brigade: How a group of thousands responds collectively to online vitriol*. Dangerous Speech Project. https://dangerousspeech.org/anti-hate-brigade/

Centre for Information Technology and Development. (2016). Traders of hate in search of votes: Tracking dangerous speech in Nigeria's 2015 election campaign. http://www.citad.org/download/traders-of-hate-in-search-of-votes/?wpdmdl=2493

Dangerous Speech Project. (2016, November 1). *Monitoring and evaluating inflammatory speech in Kenya.* https://dangerousspeech.org/kenya/

Dangerous Speech Project. (2021, April 20). *Dangerous speech: A practical guide*. Dangerous Speech Project. https://dangerousspeech.org/guide

Des Forges, A. (1999). *Leave none to tell the story: Genocide in Rwanda*. Human Rights Watch.

Diaz, J. (2021, May 6). *Want to send a mean tweet? Twitter's new feature wants you to think again*. National Public Radio. https://www.npr.org/2021/05/06/994138707/want-to-send-a-mean-tweet-twitters-new-feature-wants-you-to-think-again

European Commission. (2018, January 19). *Countering illegal hate speech online* [Fact sheet]. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/memo_18_262/MEMO_18_262_EN.pdf

Facebook (2020, May 12). *Facebook response: Sri Lanka human rights assessment*. https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Sri-Lanka-HRIA.pdf

Facebook (2021). Community Standards. https://www.facebook.com/communitystandards/

George, C. (2016). *Hate spin: The manufacture of religious offense and its threat to democracy*. MIT Press.

Google. (2021). YouTube policies: Hate speech policy. https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=2803176

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Herz, M., & Molnar, P. (Eds.). (2012). Interview with Kenan Malik. In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses* (pp. 81–91). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.008

Kogen, L. (2013, December 13). Testing a media intervention in Kenya: Vioja Mahakamani, dangerous speech, and the Benesch guidelines. Center for Global Communications Studies, University of Pennsylvania. https://dangerousspeech.org/testing-a-media-intervention-in-kenya-vioja-mahakamani-dangerous-speech-and-the-benesch-guidelines/

Leader Maynard, J. (2014). Rethinking the role of ideology in mass atrocities. *Terrorism and Political Violence, 26*(5), 821–841. https://doi.org/10.1080/09546553.2013.796934

Nelson, L. (2015, January 14). The quenelle: France's notorious anti-Semitic hand gesture, explained. *Vox*. https://www.vox.com/2015/1/14/7548289/quenelle-dieudonne-antisemitism-france

Patinkin, J. (2017, January 16). *How to use Facebook and fake news to get people to murder each other*. Buzzfeed News. https://www.buzzfeednews.com/article/jasonpatinkin/how-to-get-people-to-murder-each-other-through-fake-news-and

The Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza, Hassan Ngeze (Trial Judgment). (2003) ICTR-99-52-T, International Criminal Tribunal for Rwanda (ICTR). https://ucr.irmct.org/LegalRef/CMSDocStore/Public/English/Judgement/NotIndexable/ICTR-99-52/MSC26797R0000541998.pdf

Twitter (2021). *The Twitter rules*. https://help.twitter.com/en/rules-and-policies/twitter-rules

United Nations (2019). *United Nations strategy and plan of action on hate speech*. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml