DANGEROUS SPEECH PROJECT

Counterspeech: A Literature Review

Cathy Buerger[1]

June 2024

INTRODUCTION

Every day, some of the internet users who encounter hateful and dangerous speech online choose to respond directly, to refute or undermine it. We call that counterspeech. Many of those who have taken on this volunteer effort go about it alone, while others form groups to coordinate responses and support each other. Some executives of social platforms[2] have touted counterspeech as a method of reducing online hate, but like U.S. Supreme Court Justice Louis Brandeis, who famously opined that the remedy for bad speech is good speech,[3] they don't cite any basis for this assertion. This literature review is an effort to bridge the evidence gap by answering the question: what have scholars learned about the effectiveness of counterspeech?

This raises another question, namely what it means for counterspeech to be effective. An obvious answer is that it changes the beliefs or behavior of the person to whom it responds, persuading them to apologize or stop posting harmful messages. That's very difficult to achieve, and most counterspeakers we have interviewed say it is not their primary goal. Far more often, counterspeakers try to influence the audience — the hundreds or thousands of people who witness the exchanges. Thus in their view, and in ours, counterspeech is effective if it dissuades audience members from also spreading vitriol or if it galvanizes more counterspeech.

As far as we know, this is the first review of relevant literature. We've collected and summarized useful articles from a range of fields including political science, sociology, computational social science, and 'countering violent extremism' or CVE. These articles do not all use the term 'counterspeech,' and only a few studies have attempted to measure the effectiveness of counterspeech directly. They do, however, shed light on various features of effective counterspeech, such as qualities that make speakers/authors

[1] This is the third version of a review written by Cathy Buerger and Lucas Wright in 2019 and updated by Cathy Buerger in 2021.

[2] For example, in January 2016, speaking at the World Economic Forum in Davos, Switzerland, Facebook Chief Operating Officer Sheryl Sandberg said, "Counter-speech to the speech that is perpetuating hate we think by far is the best answer." https://www.theguardian.com/technology/2016/jan/20/facebook-davos-isis-sheryl-sandberg

[3] In his concurring opinion in Whitney v. California (1927), Justice Brandeis wrote, "If there be time to expose through discussion the falsehoods and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence."

more influential in online interactions or the extent to which pro- and antisocial behavior is contagious on the internet.

WHO DOES COUNTERSPEECH, AND WHY?

Understanding the motivations behind why individuals engage in counterspeech is crucial for comprehending the dynamics of online discourse and the efforts to combat online hatred. The following articles delve into the various factors that inspire people to participate in counterspeaking. These factors may be related to the individual doing the counterspeech or the hateful speech prompting the response and the context surrounding it. Multiple studies (Wachs et al. 2023; Ziegele, Naab, and Jost 2020) found that a person's confidence in their ability to write effective counterspeech was related to their willingness to engage. Other authors found that factors such as the number of bystanders, the perceived severity of the hateful speech and the group of people being targeted by the speech affected the willingness of people to respond. These studies collectively provide a comprehensive view of the motivations driving counterspeech and underscore the importance of group identity and perceived responsibility in fostering proactive online engagement against hate speech. Several of the articles also provide evidence for the role that training can do in increasing confidence and

- **Buerger, Catherine. 2020. "The Anti-Hate Brigade: how a group of thousands responds collectively to online vitriol." Dangerous Speech Project. https://dangerousspeech.org/anti-hate-brigade/**

  This is a detailed account of #jagärhar, one of the largest and best-organized collective efforts to respond directly to hatred online anywhere in the world. Founded in Sweden, it has been replicated in more than a dozen other countries. In interviews, #jagärhär members described how and why they do what they do. They reported being emboldened by the group to counterspeak more frequently and say they feel a sense of solidarity with other members — something that has likely helped sustain their efforts over time. The paper further describes how the group has carefully strategized to take advantage of Facebook's algorithms in their work, and to influence ideas and discourse norms among the general public rather than among the people whose hateful comments they counter online.

- **Kunst, Marlene, Pablo Porten-Cheé, Martin Emmer, and Christiane Eilders. 2021. "Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments." Journal of Information Technology & Politics 18(3): 258-273.**

Does a person's support for citizenship norms correlate with the likelihood that they will engage in counterspeech? And do different types of negative comments (disparaging vs. hateful, differences in the group attacked by the comment) motivate different reactions from readers? This study examines these questions. After taking a survey about their support for citizenship norms (adapted from the International Civic and Citizenship Education Study (ICCS)), the experimental inquiry (N = 337) exposed two group of participants to comments that disparaged members of two different groups (women and people on social welfare) – that either included or did not include hateful language. Participants were then asked how likely they were to report, dislike, and write a counterspeech response to the comment (what the authors call online civic intervention (OCI)). The study found that participants were more likely to flag comments containing hateful language on both issues, but only hateful language against women increased the likelihood that they would write a counterspeech response. People who indicated on the citizenship norms survey that they believe that taking care of others makes one a "good citizen" were more likely flag comments and respond with counterspeech – but only for comments that attacked women. The findings suggest "that the willingness to engage in OCI against hate comments depends on the attacked social group," (267).

- **Leonhard, Larissa, Christina Rueß, Magdalena Obermaier, and Carsten Reinemann. 2018. "Perceiving threat and feeling responsible how severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook." Studies in Communication and Media 7: 555–579.**

  How does the severity of hateful speech, the number of bystanders, and the presence of previous counterspeech impact an individual's intention to engage online? These were the guiding questions for an online experiment (n=304) in which participants were exposed to a variety of treatments (i.e. hateful comments of varying levels of severity, posts with high engagement and those seen by only a few people, and posts with critical comments and those without). After exposing participants to a treatment post, the authors asked them to rate how threatening the post was, whether they felt personally responsible for intervening, and the likelihood that they would have intervened if they had encountered the speech online. There was apparently no connection between the perceived severity of the speech and a person's willingness to counterspeak. The authors did find, however, that individuals expressed a lower intention of intervening if a high number of people had already seen the post. The presence of previous counterspeech comments did not impact participants' willingness to intervene.

- **Obermaier, Magdalena, Ursula Kristin Schmid, and Diana Rieger. 2023. "Too civil to care? How online hate speech against different social groups affects bystander intervention." European Journal of Criminology 20(3): 817-833.**

  In this work, researchers studied whether the identity of a group being targeted by hateful speech has an impact on bystanders' willingness to counterspeak. They conducted an online

experiment with 140 participants who were assigned to different treatment conditions containing hateful speech directed at migrants, women, and LGBTQIA+ people. Participants read a fictitious Facebook post and comment thread that appeared to be from the news outlet Spiegel Online and then answered survey questions about how uncivil they found the comments to be and their intention to intervene (or not). Participants categorized hateful speech aimed at all three groups as uncivil, but rated speech aimed at the LGBTQIA+ community slightly less uncivil than that aimed at women (perhaps a result of the participant sample, as most identified as heterosexual, cisgender women). The authors found that individuals were more willing to respond when they determined the speech to be more uncivil.

- **Obermaier, Magdalena, Desirée Schmuck, and Muniba Saleem. 2021. "I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene." New Media & Society. 25(9).**

  In this study, the researchers asked whether being a member of the group being attacked by online hate speech influences someone's desire to intervene with counterspeech. The study was an online experiment with 362 Muslim participants living in Germany. Participants were shown an Islamophobic statement either by itself or followed by counterspeech written by a non-Muslim or a Muslim (or they were shown a control statement without hate speech or counterspeech). Participants were then surveyed with questions about how they felt, and if they intended to counterspeak (and if so, what kind of response they would use). The authors found that regardless of the conditions (counterspeech or not, coming from in-group members or not), the Islamophobic speech always "led to a religious identity threat… and higher intentions to counter factually compared to the control group." Notably, however, the researchers found that when participants saw counterspeech (as opposed to an Islamophobic comment alone, without any previous response), they reported being less likely to respond hatefully (as opposed to factually).

- **Wachs, Sebastian, Norman Krause, Michelle F. Wright, and Manuel Gámez-Guadix. 2023. "Effects of the prevention program "HateLess. together against hatred" on adolescents' empathy, self-efficacy, and countering hate speech." Journal of Youth and Adolescence 52(6): 1115-1128.**

  The authors evaluated the "HateLess: Together against hatred" educational program to test its impact on empathy, self-efficacy, and countering hate speech. The German program, which is designed for students in grades 7-9, consists of five modules, designed to be delivered over the course of five days. The study sample included 820 12-16 year-olds from 11 schools across Germany. Data from a pre-test (one week before) and post-test (one month after) suggest that the program increased counterspeech, empathy, and the students' belief in their ability to counter hatred (what the authors call "self-efficacy"). The authors argue that increased empathy and self-efficacy both increase the likelihood that students will counterspeak.

- **Ziegele, Marc, Teresa K. Naab, and Pablo Jost. 2020. "Lonely together? Identifying the determinants of collective corrective action against uncivil comments." New Media & Society 22(5): 731-751.**

  In their study of #ichbinhier, the German branch of the international counterspeech network, #iamhere, the authors asked what factors affect whether members of a collective counterspeech group choose to write counterspeech comments or not (what the authors call "engaging in corrective action"). To do this, they posted a link to their survey in the group (which had over 39,000 members at the time). Their final sample contained 576 respondents. Consuming more political content on Facebook (as compared with those who did not use Facebook as a source of political information as much), feeling personal responsibility for improving online discourse, a confidence in one's comment writing ability, knowledge of the group's rules and structure, and a belief that their action might yield personal benefits (such as appreciation or a good reputation) were all positively correlated with a greater willingness to engage in corrective action. At the group level, perceived group efficacy was positively correlated with willingness to counterspeak.

WHO CAN COUNTERSPEECH REACH?

When people choose to respond to hateful speech instead of just ignoring it, they often have a variety of motivations, and similarly, their responses may reach a number of different audiences. Many counterspeakers say that their posts and comments primarily target those who read hateful speech (bystanders) rather than those who write it. Some hope to change the views or behavior of people in the "movable middle," who hold relatively moderate views and are therefore more easily persuaded, while others hope to motivate like-minded people to join the conversation as additional counterspeakers. There are also counterspeakers who try to persuade those posting hateful comments to stop. Changing the mind or behavior of the original speaker seems to be more difficult than influencing the audience, but it is not impossible. The following subsections include research studies that have examined the effectiveness of counterspeech on each of these different groups.

The Impact of Counterspeech on Bystanders

As noted in the introduction, counterspeech should be studied for its effect on the witnesses to an exchange, not only on the participants. While few studies have examined such an effect explicitly, researchers have studied how behavior spreads online by means of behavior modeling, imitation, and

descriptive norm adoption. These studies ask: Does exposure to pro- or antisocial posts make other internet users more likely to speak in a similar way? Social psychologists call this 'the contagion effect.'

Generally, this body of literature finds that the answer is yes — internet users do take cues from others, for good and for ill. Han and Brazeal (2015) found that people exposed to civil comments were more likely to write a civil comment themselves, but they did not find that exposure to incivility increased uncivil expressions (overall expressions of incivility were low in their study). Conversely, other studies (Cheng, Bernstein, Danescu-Niculescu-Mizil & Leskovec, 2017) found that exposure to anti-social or negative comments make a person more likely to post an anti-social comment. Two studies (Molina & Jennings, 2018; Han, Brazeal & Pennington, 2018) found that metacommunication comments (those that address the tone of a comment rather than its content, such as when a user scolds incivility rather than commenting on the opinions being expressed) don't increase civility but do engender additional metacommunication comments.

These findings have important ramifications for counterspeakers, as they demonstrate that the style and tone of responses can improve the quality of a discussion, and thus improve the likelihood of influencing the behavior of others. And because some research has found that antisocial behavior is also contagious, reducing exposure to hateful comments could limit the spread of similar behavior.

In many of these studies, it is difficult to distinguish whether the effect on the quality of the conversation is due to changes in the quality of the contributions or to changes in who participates. In other words, is the effect of behavioral contagion to encourage more like-minded people to join the conversation, or does it actually alter the content of what participants would otherwise have posted? Berry and Taylor (2017) analyzed historical data of participants to answer this question and found that the change in discussion quality they detected was due to changes in behavior, not changes in who participates. More research is needed on this question, especially on why people choose not to participate — a type of behavior that isn't visible and is therefore more difficult to measure and study.

- **Álvarez-Benjumea, Amalia, and Fabian Winter. 2020. "The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks." Proceedings of the National Academy of Sciences 117(37). 22800-22804.**

  How does the expression of xenophobic attitudes change after a terrorist attack? And does counterspeech play any role in what members of the audience choose to write in response? These are the questions asked by the authors. This study is unique in that two terrorist attacks occurred in Germany in the middle of their research, which allowed them to do both a natural and a lab experiment. In the study, the authors recruited 274 participants (139 before the terrorist attacks and 135 after) and presented them with eight discussion forum pages with instructions to leave a comment on each page. There were three experiment conditions: a no-norm (with a mixture of positive and hateful

comments), a weak-norm (where researchers removed the hostile comments), and a strong-norm (where only comments that were strongly positive toward the respective group being discussed remained). Their findings revealed that 1) In the no-norm treatment, there were more hateful comments directed toward refugees after the terrorist attacks, suggesting that prejudice increases if an anti-hatred norm is not maintained, 2) when anti-hatred descriptive norms are present, they work to discourage members of the audience from posting hateful content – especially more extreme comments. In fact, in this study, when "confronted with weak-norm and strong-norm comments only, the amount of prejudice they express is statistically indistinguishable before and after the [terrorist] attacks."

- **Álvarez-Benjumea, Amalia., and Fabian Winter. 2018. "Normative change and culture of hate: An experiment in online environments." European Sociological Review, 34(3): 223-237. https://doi.org/10.1093/esr/jcy005**

  The authors tested whether two interventions — counterspeech, which they call "informal verbal sanctions," and deleting hateful content from online forums — had an impact on the subsequent comments in the same spaces. Their experiment presented each research participant (n=180) with one of four variations of a discussion thread: one with hateful comments, one with hateful comments and counterspeech, and two where the hateful comments had been removed (one condition called 'censored' and the other 'extremely censored'). The researchers asked participants to read the thread and then contribute their own comment. They found that "[P]articipants were less likely to make use of hostile speech when they were presented with an environment in which previous extreme hate content had been censored" (233). The counterspeaking treatment showed no significant effect. The study was limited, however, due to the static nature of the thread, which prevented back-and-forth conversations (232). It also cannot shed light on a person's behavior after being censored or being the target of counterspeech, so long-term implications are unknown.

- **Berry, George and Sean Taylor. 2017. "Discussion quality diffuses in the digital public square." Proceedings of the 26th International Conference on World Wide Web (pp. 1371-1380). International World Wide Web Conferences Steering Committee. https://arxiv.org/abs/1702.06677**

  The researchers behind this study (which was part of a product test at Facebook) conducted a within-subject experiment to determine the effect of the order of comments (chronological or by engagement) on the quality of comments shown to users and the quality of user comments in response. The sample consisted of 100,000 comments drawn from the 5,000 largest English-language Facebook pages. On average, social treatment ranking resulted in high quality visible comments and, among the users who choose to contribute to the discussion, seeing those

higher quality comments increased the quality of their subsequent contributions. The authors attribute this effect to the adoption of descriptive norms — social rules based on perceptions of how others are behaving.

- **Cheng, Justin, Michael Bernstein, Christian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions." Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17, 1217–1230. https://dl.acm.org/doi/10.1145/2998181.2998213**

  In this online experiment, researchers exposed participants to either a positive or negative stimulus (being told that their answers to a short quiz were good and above average, or poor, both absolutely and in relation to other participants). Afterwards, participants were asked to read an article with a comment section that was either benign or 'troll-like' — and then write their own comment. The authors found that both negative mood (exposure to the negative stimulus) and exposure to a troll-like discussion increased the likelihood that a participant would write a trolling comment, doubly so when both conditions were combined. In fact, the authors claim that their "predictive model of mood and discussion context together can explain trolling behavior better than an individual's history of trolling" (1217). They corroborate their experimental findings through the analysis of "large-scale and longitudinal observational data" (1223).

- **Friess, Dennis, Marc Ziegele, and Dominique Heinbach. 2020. "Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions." Political Communication. 1-23. https://doi.org/10.1080/10584609.2020.1830322**

  For this study, the authors evaluated the effectiveness of a German collective counterspeech effort called #ichbinhier ("I am here"). They used a dataset of comment threads to which #ichbinhier members had contributed between November 01, 2017 and January 31, 2018 to answer two questions: whether comments made by #ichbinhier members were more 'deliberative' than those posted by non-members (researchers coded for rationality, constructiveness, politeness, civility, and reciprocity), and whether deliberative top-level comments were associated with more deliberative second-level comments. They found the answer to both questions to be 'yes,' suggesting that discourse norms established or reaffirmed by members of a group can have an impact on the quality of online discourse (15). The study was somewhat limited by its small sample size and also because it investigated only the relationship between top-level comments and direct replies to them (17) rather than looking at the impact of counterspeech on the overall discourse in the thread.

- **Han, Soo-Hye, and LeAnn M. Brazeal. 2015. "Playing Nice: Modeling Civility in Online Political Discussions." Communication Research Reports, 32(1): 20–28. https://www.doi.org/10.1080/08824096.2014.989971**

This online experiment found that exposure to civility increased willingness to participate and heightened civility in participants' comments. Exposure to incivil comments did not affect the participants' comments, but the participants in this study exhibited low levels of incivility generally, so this could be a feature of the sample.

- **Han, Soo-Hye, LeAnn Brazeal, and Natalie Pennington. 2018. "Is Civility Contagious? Examining the Impact of Modeling in Online Political Discussions." Social Media + Society, 4(3). https://doi.org/10.1177/2056305118793404**

  In another online experiment on the effect of exposure to civility on participation, researchers found that participants in the civil condition were more likely to write a civil comment, less likely to go off-topic, and more likely to "offer a fresh perspective" (7). Exposure to metacommunication (comments that scold incivility and encourage civility) in an uncivil discussion did not increase comment civility, but it did increase metacommunication in participant comments.

- **Miškolci, Jozef, Lucia Kováčová, and Edita Rigová. 2018 "Countering hate speech on Facebook: The case of the Roma minority in Slovakia." Social Science Computer Review. https://doi.org/10.1177/0894439318791786**

  Drawing on over 7,500 Facebook comments, this study used qualitative content analysis to identify particular themes and terms used on Facebook to describe Roma in Slovakia. It also tested the effectiveness of counterspeech to respond to these generally negative portrayals. The study found that counterspeech was not effective for changing the behavior of the user who posted negative comments about Roma people. It was, however, followed by an increase in the number of pro-Roma comments within the same comment thread.

- **Molina, Rocío Galarza, and Freddie Jennings. 2018. "The Role of Civility and Metacommunication in Facebook Discussions." Communication Studies, 69(1): 42–66. https://doi.org/10.1080/10510974.2017.1397038**

  This study used an online experiment to measure how discussion civility affects participant commenting behavior. Participants viewed a Facebook post about genetically modified organisms and a comment section in one of the following conditions: civil discussion, uncivil discussion, uncivil discussion with metacommunication (comments that scold incivility and encourage civility), and a control group with no comments. Results showed that exposure to civility and metacommunication increased participants' willingness to write a comment and that their comments were most likely to be modeled on the condition comments (i.e. civility begets civility, comments with metacommunication beget comments with metacommunication).

- **Rösner, Leonie, Stephen Winter, and Nicole Krämer. 2016. "Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior." Computers in Human Behavior, 58: 461–470. https://doi.org/10.1016/j.chb.2016.01.022**

  Researchers exposed treatment groups to online conversations with varying proportions of uncivil comments. Each group was exposed to only one treatment condition. The authors did not find a relationship between exposure to incivility and incivility in participant comments, but they did find that exposure to incivility increased participants' aggressive reactions to a subsequent (unrelated) story completion task.

- **Seering, Joseph, Robert Kraut, and Laura Dabbish. 2017. "Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting." Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17, 111–125. https://dl.acm.org/doi/10.1145/2998181.2998277**

  This observational study of Twitch live chatrooms used an interrupted time series model, in which data is collected at several, equally-spaced points in time, to measure imitation effects for prosocial behavior, anti-social behavior, and questions. Results showed that all three types of behavior resulted in an increase in that same behavior within the next ten messages compared to the previous ten messages in the chat. This effect was stronger when the message originated from high influence users (moderators or paid subscribers to the channel).

- **Ziegele, Marc and Pablo B. Jost. 2020. "Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments." Communication research 47(6): 891-920.**

  Ziegele and Jost ask whether interactive moderation (moderation that is conversational and visible to the others in an online space), has an impact on the willingness of others to participate in a comment thread discussion. To test it, they conducted an online experiment where participants (n=811) were shown one of six sets of three consecutive screenshots of a news story and comments (with or without moderation). On some examples, moderators responded to uncivil comments with a calm, factual response. In others, they used a sarcastic tone. They also varied the type of news story with one half being on "lighter" topics and the other half being on more controversial ones. After seeing the screen shots, participants were then asked if they would like to participate by writing a comment, and they rated the respectfulness of the discussion atmosphere and the credibility of the news source. The researchers found that factual interactive moderation increased participants' positive perception of the discussion atmosphere and their willingness to participate. Sarcastic responses decreased users' perception of the credibility of the news source – especially when sarcasm was used to respond to incivility in comments on more controversial news stories.

The Impact of Counterspeech on the Hateful Speaker

The studies in this section all attempt to gauge the effectiveness of counterspeech in cases where an internet user (a 'counterspeaker') directly addresses someone who posted a hateful or dangerous message in an effort to change that person's opinion or behavior. They have produced limited **and** varied findings. Miškolci et al. (2018) found that responding directly was not effective at stopping the behavior (i.e. posting hateful content) of the original speaker, but it was a useful way to reach a larger audience and provoke more counterspeech. Schieb and Preuss (2016), however, concluded that counterspeech can influence the original speaker, although the effectiveness of a counterspeech interaction depends on the proportionate size of the group of hateful speakers in a particular online space. In their study, a message was more effective when counterspeakers greatly outnumbered those sharing hateful messages. They also found that a small group of counterspeakers could still be effective, as long as the other users within an online space held relatively moderate (rather than extreme) views.

Other factors are important as well. Some studies (Bartlett and Krasodomski-Jones, 2015; Frenett and Dow, 2015), found that the tone of a counterspeech message affects whether the interaction has a measurable impact. These studies also demonstrated that specific variables of the interaction (how many people are speaking, what is being said, and who is listening) influence the effectiveness of counterspeech.

- **Abdelkader, Engy. 2014. "Savagery in the Subways: Anti-Muslim Ads, the First Amendment, and the Efficacy of Counterspeech." Asian Am. LJ 21: 43. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2264791**

  The author documented responses to anti-Muslim ads placed in the public transportation systems of three cities: New York, Detroit, and Washington, D.C. In Abdelkader's view, counterspeech that focuses on understanding and tolerance educates the public, allowing for anti-hatred coalitions to form within communities, and therefore should be viewed as a positive remedy for harmful speech. The author did note that in communities where a majority of the people support the hateful speech, counterspeech may fail.

- **Bartlett, Jamie and Alex Krasodomski-Jones. 2015. "Counter-speech: Examining content that challenges extremism online." Demos. https://www.demos.co.uk/wp-`content/uploads/2015/10/Counter-speech.pdf**

  This report, commissioned by Facebook, examines how counterspeech that challenged far-right political Facebook pages in France, Italy, and the UK was produced and shared. For the purposes of the report, the authors used interaction data (comments, likes, and shares) to determine a post's effectiveness (as that gives a sense of the reach of the content). The authors also

analyzed comments and interactions on counterspeech and populist right wing pages. They found that form and tone mattered. For example, counterspeech posts including questions generated the most interaction (likes and comments) among forms of content, and 'funny or satirical' counterspeech posts received the most interaction among all the tones studied. Additionally, the data suggest that "counter-speech pages are not as active as populist right wing pages," so the authors logically suggest that "if counter-speech page administrators and users were more active, and changed their content slightly, it could dramatically increase the reach of their messages" (14). The authors also recommend that counterspeakers write more "'constructive counter-speech' compared to nonconstructive counter-speech; and more comments about specific policy issues," (14).

- **Eschmann, Rob, Jacob Groshek, Senhao Li, Noor Toraif, and Julian G. Thompson. 2021. "Bigger than sports: Identity politics, Colin Kaepernick, and concession making in# BoycottNike." Computers in Human Behavior 114: 1-11.**

  In 2016, Colin Kaepernick, who was an NFL quarterback, knelt during the national anthem before a game. His actions, done in protest of the treatment of black people in the US, sparked debate, and when Nike featured his image in an ad along with the words "believe in something" in 2018, some proposed a boycott of the company. Using a mixed methods approach that combined quantitative and qualitative analysis, the authors explore the topic of "concession making," which they define as "users demonstrating a willingness to adjust their position on a contentious political issue." They analyzed 958,738 tweets that mentioned the hashtag #justdoit (Nike's motto), 203,021 tweets that mentioned the hashtag #nikeboycott, and 844,121 tweets that mentioned the user @kaepernick7 (Kaepernick's username on Twitter) between September 4, 2018, to September 11, 2018. From this, they narrowed their sample down to tweets coming from 48 users (chosen because they were highly networked users with fewer than 1,000 followers). From this sample, 10.4% were found to have made concessions. The authors found that several factors increased the likelihood that a user would make concessions:
    - Users with fewer followers were more likely to make concessions than those with more followers.
    - Users who had more relevant tweets on a topic and received more replies were more likely to make concessions.
    - Users who began anti-Kaepernick were more likely to make concessions than those who started off pro-Kaepernick

- **Frenett, Ross and Dow, Moli. 2015. "One to one online interventions: A pilot CVE methodology." Institute for Strategic Dialogue. https://www.isdglobal.org/isd-publications/one-to-one-online-interventions-a-pilot-cve-methodology/**

Frenett and Dow conducted a pilot study of far-right and Jihadist Facebook users "at risk of falling into the orbit of extremist groups"(7). This report describes the types of messages that were most effective at drawing 'reactions.' The authors defined reactions broadly, including sending a response message to the counterspeaker and blocking the counterspeaker. Most useful is their analysis of the messages that were successful in prompting a 'sustained engagement' (five or more messages exchanged). They found that the tone of the message was highly correlated with response rate. Antagonistic messages, for example, never got responses. Casual or sentimental messages, however, prompted 83% of people to respond. Similarly, offers of assistance or personal stories were much more likely to prompt a sustained engagement than calling attention to the negative consequences of someone's hateful speech.

- **Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci et al. 2021. "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment." Proceedings of the National Academy of Sciences. 118(50).**

  What impact does the content of counterspeech messages have on the behavior of Twitter users who have posted xenophobic or racist messages? To answer the question, the authors designed a field experiment where users (n=1,350) were randomly assigned to receive one of three counterspeech messages (using empathy or humor or warning of consequences) or to a control group. After the intervention, they tested whether the user deleted past xenophobic tweets and whether they created new xenophobic tweets in the following four weeks, in addition to measuring the "negative sentiment of all tweets" in the four-week follow-up period. They found that "users assigned to the empathy treatment sent, on average, 1.3 fewer xenophobic tweets and 91.6 fewer total tweets, and were 8.4 percentage points more likely to delete the original xenophobic tweet." The other treatments produced no significant effects.

- **Munger, Kevin. 2017. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." Political Behavior. 39(3): 629-649. https://doi.org/10.1007/s11109-016-9373-5**

  Munger tested the impact of identity and social status on successful group norm promotion. He rebuked accounts that had used anti-black slurs on Twitter, using bots variously identified as black or white and as high- or low-status (many vs. few followers), documenting the difference in reaction. White men who had used racist slurs were more likely to change their behavior when confronted by a bot masquerading as a white counterspeaker with many followers, than when called out by what appeared to be a black counterspeaker or a white counterspeaker with fewer followers.

- **Saltman, Erin, Farshad Kooti, and Karly Vockery. 2021. "New models for deploying counterspeech: Measuring behavioral change and sentiment analysis." Studies in Conflict & Terrorism. pp.1-24.**

  The authors used a two-part study to investigate "whether exposure to counterspeech content has an effect on a target audience's continued engagement and/or consumption of violent extremist content." In the first part, they used A/B testing to determine whether counterspeech (either preventative alternative messaging or direct counter-messaging) influenced user behavior. After exposing groups to one of the forms of counterspeech, they compared the rate of terms of service violations between the treatment and control groups for 90 days. The counterspeech messages (which appeared as pop-up ads), were shown to individuals who had engaged with content that sympathized with global Islamist extremist terrorist movements. There was no noticeable impact of the counterspeech on the wider audience, but among individuals who had "higher hard signals of views, shares and/or direct engagement with violent extremist content," some sought out less with such content.

  In the second part of the study, the Redirect Method (which, when people search for harmful content, populates their search results with alternative links) was used to nudge users at risk of becoming more involved with white supremacy and neo-Nazi groups toward Life After Hate, an NGO that helps people leave far-right extremist groups. The study showed that the intervention did successfully drive users to engage with the organization (as measured through sustained click through rates and website engagement).

  In addition to its findings, the article is notable as it was a project taken on by employees at Facebook, who therefore had access to data and could build the experiment into a user's experience on the platform. As the authors note, the project was "led by the Facebook Counterterrorism and Dangerous Organizations Policy Team in coordination with internal data scientists, community integrity engineers, market teams and safety researchers with privacy and legal reviews of both methodologies before their deployment."

- **Schieb, Carla, and Mike Preuss. 2016. "Governing hate speech by means of counterspeech on Facebook." 66th ICA Annual Conference, at Fukuoka, Japan, pp. 1-23. https://www.researchgate.net/publication/303497937_Governing_hate_speech_by_means_of_counterspeech_on_Facebook**

  The authors used a computational simulation model to determine factors that impact the effectiveness of counterspeech on Facebook. Not surprisingly, they found that the proportion of counterspeakers to hateful speakers and the intensity of opinion held by the hateful speakers are both important determinants of success.

AUTOMATED COUNTERSPEECH

Recent advances in generative Artificial Intelligence (AI) have inspired interest in using it to counter online hatred. Until now, most of the research being done in this area has focused on developing datasets of hateful speech and counterspeech pairs (or, as in the case of Bonaldi et al., (2022), "multi-turn" dialogues where there is a back-and-forth conversation between two users), testing different processes for generating automated responses, and seeing which ones produce the highest quality responses. Although the research on automated counterspeech has taken large steps forward in the past few years, there are still significant gaps. One of the most glaring is the question of whether automated counterspeech affects the target audience in the same way that human-generated messages do. In most of the work described below, the "quality" or "effectiveness" of a response was defined and evaluated by a research team, often based on qualities such as inoffensiveness and informativeness. A notable exception to this is Bilewicz et al (2021), in which the authors used a bot on Reddit to perform actual interventions, and then analyzed the speech of the users targeted by the intervention in the 60 days before and after they received a counterspeech message, to document any changes in behavior. As AI technology continues to improve, and as the body of research on automated counterspeech grows, we hope to see more studies testing the effects of automated counterspeech on people and their behavior.

- **Ashida, Mana, and Mamoru Komachi. 2022. "Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions." In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH): 11-23.**

  The authors conducted an evaluation of automated counternarratives against microaggressions, which they define as "unintentional offensive remarks." They compared the text of these "machine-generated microinterventions," produced using three different language models, GPT-2, textscGPT-Neo, and textscGPT-3. Participants recruited through Amazon Mechanical Turk rated each response on its offensiveness, stance (whether it agreed or disagreed with the original microaggression), and informativeness. Responses received the highest ratings when they were specific, informative, deemed to be inoffensive, and opposed to the microaggression. (The authors do not note whether this definition of an "effective counternarrative" is based on empirical research or their own thinking). GPT-3 performed the best in their evaluation, although they recommend an additional fact-checking stage to ensure that the interventions do not contain misinformation.

- **Bilewicz, Michał, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, and Michał Wroczyński. 2021. "Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment." Aggressive behavior 47(3): 260-266.**

Using a bot on the social news website Reddit, the authors conducted an intervention study in which the bot monitored r/MensRights and r/TooAfraidToAsk and responded with one of three interventions whenever it detected a personal attack. The bot either intervened with a disapproving message (example: “Hi:) I somewhat commiserate with what you're feelin’, but let's try to express our points without hurtful language,”), a reference to abstract norms (example: “Good day sir, have you ever thought about how this discussion could be more enjoyable for all if we would treat each other with respect?”), or an empathetic response (example, “Some behaviors might be hard to get for some people but let's keep in mind there are people of flesh and blood on the other side of the screen,”). In total the bot conducted 454 interventions (with an additional 437 users from other subreddits selected for a control group). The authors made efforts to disguise the fact that the bot was a bot, for example by giving it a history and having it create posts that were not part of the intervention, during the 6 months of the experiment. The authors calculated the proportion of comments that took the form of personal attack during the 60 days before and after the intervention. They found all three methods of response effective at diminishing the proportion of personal attacks with no significant difference between them.

- **Bonaldi, Helena, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. “Human-machine collaboration approaches to build a dialogue dataset for hate speech countering.” In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. pp. 8031–8049**

  Datasets that could be used for counterspeech generation have been produced before, but they have all been single turn, meaning one hate comment paired with one counterspeech response. DIALOCONAN is the first dataset produced containing multi-turn hate speech and counterspeech dialogues. The authors of this article posited that most counterspeech interactions contain some back and forth between a person posting hateful speech and someone who counters it (although it is unclear how they came to this conclusion). The article describes their novel approach to assembling the dataset (which contains over 3,000 fictitious dialogues). They used a 3-stage hybrid approach, and the article’s focus is on describing this approach in detail. In stage one, two human experts (NGO workers who regularly write online counternarrative posts) wrote out sample dialogues that had 4, 6, or 8 turns. Stage two consisted of increasing the available data by paraphrasing the counterspeech samples produced in stage one. In stage three, a large language model (LLM) trained on the data from stages one and two generated additional examples.

- **Chaudhary, Mudit, Chandni Saxena, and Helen Meng. 2021. "Countering online hate speech: An nlp perspective." arXiv preprint arXiv:2109.02941.**

  This paper provides a conceptual framework for countering hateful speech using Natural Language Processing (NLP), dividing methods into three categories: detection, prevention, and intervention. The authors note that there is a dearth of research on using NLP to prevent and intervene against online hateful speech, although researchers have made significant strides in exploring using NLP for detection. The authors discuss three types of reactive countering

methods (responding to hateful speech that has already been posted): 1) AI generated counterspeech, 2) neural style transfer (where the offensive parts of a comment would be redacted or modified to make the comments inoffensive), and 3) automated and semi-automated moderation (where a NLP-trained classifier would detect hateful speech, automatically moderate comments flagged with high-confidence, and flag those detected with lower-confidence for review by a human moderator. The authors also discuss the proactive method of "preemptive moderation" that uses NLP tools to predict which comment threads will turn severely hateful so they can be targeted for intervention and moderation. For each method, the authors discuss existing studies on the method as well as ethical and legal concerns.

- **Ching, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. "Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2819–2829.**

  With the hope of one day producing tools that can automatically generate counterspeech, the authors assembled a multilingual dataset of Islamophobic hate speech/counternarrative (called CONAN). The dataset contains 4078 hate speech/counter-narrative pairs; (1288 pairs in English, 1719 pairs in French, and 1071 in Italian). The authors produced the dataset with the assistance of volunteers from three different NGOs (one in France, one in Great Britian, and one in Italy). Two volunteers per language were asked to write around 50 examples of Islamophobic hateful speech reflective of the "typical arguments" used against Islam in their context. NGO volunteers (called "operators" by the authors) then participated in three-hour sessions to craft responses to the hateful speech, using "fact-bound information and non-offensive language." In total, 111 operators took part in nine data collection sessions. To increase the number of examples in the dataset, the authors also paraphrased the examples produced by the NGO volunteers. The authors hope to use the data to develop tools in the future to automatically generate counterspeech.

- **Cypris, Niklas Felix, Severin Engelmann, Julia Sasse, Jens Grossklags, and Anna Baumert. 2022. "Intervening against online hate speech: A case for automated Counterspeech." IEAI Research Brief. pp. 1-8.**

  The authors propose using automated counterspeech along with deletion to counter online hate speech. The research report begins by making a case for the efficacy of counterspeech, citing relevant literature. There are two primary psychological mechanisms that allow counterspeech to be effective, they argue. The first is that counterspeech comments can make a specific "mode and tonality" more visible to the audience. Second, members of the audience may adopt the social norms expressed by the counterspeaker, if they feel some sort of social connection with them. Based on scant evidence, the authors posit that automated counterspeech should

function similarly, making it a viable solution to online hatred. The article also discusses various ethical concerns with using, and conducting research on, automated counterspeech. The authors never discuss details of the proposed automated counterspeech bot (e.g. would it identify itself as a bot? To what content would it respond? What kind of strategies would it use?) nor do they discuss potential differences in human-generated and AI-generated counterspeech, both of which are significant weaknesses of the report.

- **Fanton, Margherita. Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini.ˇ "Human-in-the loop for data collection: a multi-target counter narrative dataset to fight online hate speech." 2021. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp.3226–3240.**

  This article is a detailed description of a new "human-in-the-loop" methodology for developing a generative language model for counternarratives. The authors began by asking an NGO expert to assemble a list of prototypical hateful messages targeting a predetermined list of groups. They then created an online form for NGO volunteers to enter 1) counternarrative responses to the hateful messages, and 2) new pairs of hateful speech and counternarratives from NGO volunteers. After assembling the initial set of pairs, they used an iterative process in which ChatGPT-2 generated additional pairs, which were refined by human annotators to improve quality. They found that their methodology is "scalable and facilitates diverse, novel, and cost-effective data collection," although it still faces challenges in terms of quality, as ChatGPT-2 sometimes produced counternarratives that were logical, but hateful in their own way or factually inaccurate. As there have been technological improvements to generative AI since this article was published, some of these challenges may have been mitigated.

- **Fraser, Kathleen C., Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. "What Makes a Good Counter-Stereotype? Evaluating Strategies for Automated Responses to Stereotypical Text." In Proceedings of the First Workshop on Social Influence in Conversations. pp25-38.**

  In this article, the authors investigate the feasibility of using ChatGPT to effectively counter harmful stereotypes online. Based on a review of the literature, they identify 11 possible strategies that can be used to counter stereotypes: denouncing stereotypes, presenting counter-facts, sharing counter-examples / contradictions, using humor, warning of consequences, showing empathy for the speaker, asking critical questions, providing examples of individuals from outside the target group who also have the stereotypical trait (what they call, "broadening exceptions," broadening universals (stating that anyone could have the stereotypical trait), emphasizing positive qualities of the target group, and asking the speaker to consider how they would feel if they were part of the target group. They then use ChatGPT to

generate responses to a set of stereotypes they compiled, and then they analyzed them in two ways: "(1) Technical: Is ChatGPT capable of generating counterstereotypes that are believable, inoffensive, and use the requested strategy? (2) Social: Do annotators believe that the generated response will be effective from a bystander's perspective?"

The authors found that ChatGPT struggled to produce some types of strategies – specifically the strategy they call "broadening exceptions" and using humor (the authors deemed over a third of the generated humorous response potentially offensive). As has been widely reported, generative AI also struggles with truthfulness (in this study, the authors found approximately 40% of the counter-facts produced were either inaccurate or unverifiable). Because of this, the authors recommend conducting more research on the strategies that showed the most promise: denouncing, warning of consequences, and using an empathetic tone. These are useful findings, but there are clear weaknesses. First, the study's measure of effectiveness is based only on the authors' assessment of whether they believed the statement would be effective on an audience of bystanders. They do not define what it would mean for it to be effective (e.g. change someone's mind? Encourage them to also speak out against the stereotype?), nor did they test it on people likely to be in the imagined audience.

- **Mun, Jimin, Cathy Buerger, Jenny T. Liang, Joshua Garland, and Maarten Sap. 2024. "Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate." In Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1-22.**

  Although AI has been proposed to help scale up counterspeech efforts, many questions remain about how exactly AI could assist in this process, since counterspeech is a deeply empathetic and agentic process for those involved. In this work, the authors aim to answer this question. They conducted in-depth interviews with 10 extensively experienced counterspeakers and a large-scale public survey with 342 everyday social media users. In participant responses, they identified four main types of barriers and AI needs related to resources, training, impact, and personal harms. The results also revealed overarching concerns surrounding the themes of authenticity, agency, and functionality in using AI tools for counterspeech. The article concludes with considerations for designing AI assistants that lower barriers to counterspeaking without jeopardizing its meaning and purpose.

- **Qian, Jing , Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. "A benchmark dataset for learning to intervene in online hate speech." arXiv preprint arXiv:1909.04251.**

  In contrast to many other hate speech datasets, the two introduced in this article (one collected from Gab and one from Reddit) include conversational segments and therefore offer annotators conversational context. One especially useful contribution of the article is a table in which the

authors compare various attributes of their datasets with nine others produced by other researchers. Their Reddit database contains 5,020 “conversations” (22,324 comments) and the Gab database contains 11,825 “conversations (33,776 posts). After collecting the data, 926 Mechanical Turk workers were shown five conversations each, asked to respond whether they contained hate speech (using Facebook’s definition of hate speech as a guide). If so, they were asked to write a response to it. The authors identified four common types among those:

- Identifying hate keywords (Example: “The C word and language attacking gender is unacceptable. Please refrain from future use.”)
- Categorizing hate speech (Example: “The term ””fa**ot”” comprises homophobic hate, and as such is not permitted here.”)
- Using a positive tone followed by transitions: (Example: “I understand your frustration, but the term you have used is offensive towards the disabled community. Please be more aware of your words.”)
- Suggesting proper actions: (“I think that you should do more research on how resources are allocated in this country.”)

The authors then used the collection of responses, along with the two hate speech datasets, to train and test the ability of three different generative LLM models to automatically generate interventions. They conclude the article by discussing the strengths and weaknesses of each model as well as their datasets.

- **Zhu, Wanzheng, and Suma Bhat. 2021. "Generate, prune, select: A pipeline for counterspeech generation against online hate speech." arXiv preprint arXiv:2106.01625.**

This article is a detailed description of a 3-step counterspeech generation system designed by the authors to produce “diverse and relevant” responses to hateful speech. The authors state that they set this goal because they believe other natural language generation (NLG) methods produce “commonplace, repetitive and safe responses regardless of the hate speech (e.g., “Please refrain from using such language.”) or irrelevant responses, making them ineffective for de-escalating hateful conversations.” Their “Generate, Prune, Select” (GPS) pipeline uses generative models to develop a large set of responses (generate), then uses a BERT model (a tool to better understand the context around language) to filter out the ungrammatical responses (prune), before using another automated process to select for those most relevant to the initial hateful speech (select). The authors do not attempt to evaluate effectiveness, but find that their methodology succeeds in producing diverse and relevant counterspeech responses.

CASE STUDIES

Though we focus here on research that can help determine whether counterspeech is effective, we also summarize a rich body of descriptive literature. These studies illuminate many types of interactions that fall under the term "counterspeech," and use different ways of defining success in counterspeech interactions. Some describe a few detailed case studies (Stroud and Cox, 2018). Others propose classification models (Mathew et al., 2019) or typologies of counterspeech interactions (Wright et al., 2017; Benesch et al., 2016). Wright et al. (2017), for example, categorize counterspeech interactions by the number of people involved, describing four different 'vectors': one-to-one, one-to-many, many-to-one, and many-to-many. Other articles (Benesch et al., 2016; Briggs and Feve, 2013) classify counterspeech interactions by the strategies used (humor, shaming, etc). We have also included two articles about offline counterspeech (Abdelkader, 2014; Richards and Calvert, 2000) to illustrate the wide breadth of speech that scholars have called 'counterspeech.'

- **Benesch, Susan; Derek Ruths, Kelly P. Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. "Counterspeech on Twitter: A Field Study." Dangerous Speech Project.** https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/

  A qualitative analysis of counterspeech interactions found on Twitter, this paper classifies counterspeech conversations by the number of people involved, describing four 'vectors': one-to-one, one-to-many, many-to-one, and many-to-many. The authors also describe a variety of counterspeech strategies that they observed in their data such as pointing out hypocrisy, providing facts to correct misstatements, and denouncing hateful or dangerous speech.

- **Briggs, Rachel, and Sebastian Feve. 2013. "Review of programs to counter narratives of violent extremism." Institute of Strategic Dialogue.** https://www.dmeforpeace.org/peacexchange/wp-content/uploads/2018/10/Review-of-Programs-to-Counter-Narratives-of-Violent-Extremism.pdf

  Largely focused on what governments might do to counter extremist messaging, this report divides what the authors call 'counter-messaging' into three categories: government strategic communication, alternative narratives, and counter-narratives. This distinction is useful for thinking about the different goals and audiences of counterspeech interactions. The article concludes with an appendix of 18 case studies illustrating each form of counter-messaging, including online and offline examples, primarily from Europe and the United States. The authors conclude that "Counter-messaging strategies should be multi-layered, integrating the use of messages that erode the intellectual framework of violent extremist ideologies, combined with more constructive approaches aimed at providing credible alternatives to those susceptible to such messaging" (25).

- **Brisson-Boivin, Kara. 2019. "Young Canadians Pushing Back Against Hate Online." MediaSmarts. Ottawa. https://mediasmarts.ca/research-policy/young-canadians-pushing-back-against-hate-online**

  How do Canadian adolescents experience casual prejudice online, and how do they decide whether to counterspeak against it? This report attempts to answer these questions. Between October and December 2018, researchers surveyed 1,000 Canadians between the ages of 12 and 16 about their experiences with hate online, gathering data regarding their internet usage and their opinions on casual prejudice. Most relevant for this literature review are the sections on "enabling factors for pushing back" and "barriers to pushing back." Enabling factors included hearing that the content was hurtful to someone else, having clear paths to report the content on social media platforms, and thinking that friends agreed that the speech was prejudiced. Factors that participants cited as barriers to responding to prejudice included being afraid their reaction might make things worse, not knowing what to say, not being sure if a response was needed, and not knowing how others would react to a response.

- **Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. "Countering hate on social media: Large scale classification of hate and counter speech." Association for Computational Linguistics. pp 102-112. https://www.aclweb.org/anthology/2020.alw-1.13/**

  The authors collected over 9 million tweets originating from two competing online groups: Reconquista Germanica (RG) and Reconquista Internet (RI). RG is "a highly-organized hate group which aimed to disrupt political discussions and promote the right-wing populist, nationalist party Alternative für Deutschland (AfD)" (103). RI is a group formed with the goal of countering RG's messaging. At their peaks, RG had between 1,500 and 3,000 active members and RI had about 62,000 registered members, of whom over 4,000 were active (103). The authors used the tweets, each posted from an account associated with one of the two groups, to create an automated classifier that could recognize hateful speech and counterspeech. They then used the classifier to identify these two types of discourse in 135,000 "fully-resolved Twitter conversations" that took place between 2013 and 2018 in order to study the frequency of hateful speech and counterspeech as well as the interaction between the two forms of discourse. After RI formed, the intensity and proportion of hateful speech apparently decreased. The authors note that "this result suggests that organized counter speech might have helped in balancing polarized and hateful discourse, although causality is difficult to establish given the complex web of online and offline events and process in the broader society throughout that time" (109). The study is notable not only for its findings, but also for its method; it produced the first automated classifier for counterspeech.

- **Richards, Robert D., and Clay Calvert. 2000. "Counterspeech 2000: A New Look at the Old Remedy for Bad Speech." BYU L. Rev.**
  https://digitalcommons.law.byu.edu/lawreview/vol2000/iss2/2

  Using five case studies of offline counterspeech, Richards and Calvert examined whether, and under what conditions, it might serve as an effective remedy to harmful speech. The case studies were counterspeech campaigns organized by groups or companies, and the authors defined harmful speech broadly, ranging from speech supportive of the Klu Klux Klan to messages that damaged the reputation of a business. They conclude that large-scale counterspeech campaigns are most effective when they are able to leverage media connections in order to increase their audience (556).

- **Stroud, Scott R. and William Cox. 2018. "The varieties of feminist counterspeech in the misogynistic online world." In Mediating Misogyny, pp. 293-310. Palgrave Macmillan, Cham.**
  https://doi.org/10.1007/978-3-319-72917-6_15

  The authors used two case studies to outline a "spectrum of force" of feminist counterspeech, ranging from efforts to "negatively [...] affect [the] psychological or physical well-being" of the original misogynistic speaker ("targeted negative counterspeech") to efforts to create a support network for the target of the misogyny that mostly disregard the misogynist ("directed positive counterspeech") (302). The article also discusses ethical issues related to feminist counterspeech.

- **Thant Sin Oo, Zaw Myo Min, and Matt Schissler. 2020. "Structures and vocabularies of counter-speech on Facebook in Myanmar." Article 19 Working Paper, Bangkok.**
  https://www.article19.org/wp-content/uploads/2020/11/2020.11.27-Thant-Sin-Oo-et-al-A19-Counter-speech-working-paper-FINAL.pdf

  This article is a qualitative study of hateful speech and counter responses in Myanmar. The authors conducted 10 interviews and eight focus groups with "people targeted for their status as members of religious, ethnic, or LGBTQ communities," then analyzed 17 Facebook posts with their associated comment threads. They use a broad definition of counterspeech, which includes both direct responses to hatred and proactive positive messaging that isn't a response to any specific message. The article is organized around a collection of the authors' observations, with each one including sections for interpretation and implications. The authors found that comments on posts containing negative language about a minority group usually "reverberate the negativity" and that posts that portrayed a minority group positively were likewise primarily followed by positive comments. The article also includes a section on the "vocabularies" of counterspeech in Myanmar, describing several common rhetorical themes observed in

counterspeech comments. These include describing positive human qualities, references to the religious notion of karma, and criticizing extremism. The article concludes with an extensive appendix of examples illustrating the different counterspeech vocabularies (written in the original Burmese).

- **Wright, Lucas, Derek Ruths, Kelly P. Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. "Vectors for counterspeech on Twitter." In Proceedings of the First Workshop on Abusive Language Online, pp. 57-62. https://dangerousspeech.org/vectors-for-counterspeech-on-twitter/**

  A condensed version of the aforementioned Benesch et al. (2016) paper, this essay categorizes counterspeech conversations based on the number of people taking part in the interaction: one-to-one, one-to-many, many-to-one, and many-to-many. The authors argue that the success of counterspeech — its potential to have "a favorable effect on people to whom it responded" (57) varies, at least in part, according to the number of people who take part. The article describes each of the four vectors in detail, explaining factors that might influence the effectiveness of each vector of counterspeech.

- **Ziems, Caleb, Bing He, Sandeep Soni, and Srijan Kumar. 2020. "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis." arXiv preprint arXiv:2005.12423. https://arxiv.org/pdf/2005.12423.pdf**

  Hate speech and physical attacks on Asians during the COVID-19 pandemic have been widely documented in the United States and abroad. This article presents one of the few studies to examine efforts to counter anti-Asian speech. Researchers created a publicly-available dataset of over 30 million COVID-19 related tweets posted between January 15 and April 17, 2020. To create their classifier, they hand annotated 2,400 tweets, tagging each as 1) containing anti-Asian sentiment, 2) containing counterspeech to such messages, or 3) neutral (containing COVID-19 related speech that is not related to Asians). They then used the classifier to analyze the entire dataset, finding that 2.8% were hateful, 0.65% were counterspeech, and 86.77% were neutral. The analysis produced several findings, including that users who posted hateful tweets, but not counterspeech (people the authors call 'hateful users') followed more people and had more followers than those who had posted counterspeech, and that after 'hateful users' posted their first hateful tweet, they participated more frequently in COVID-related discussions than counterspeakers did after posting their first counterspeech tweet. Their analysis also revealed that

Contribute to this Literature Review

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

Please send ideas and inquiries to Cathy@DangerousSpeech.org

Dangerous Speech Project

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