DANGEROUS SPEECH PROJECT

Counterspeech: A Literature Review

Cathy Buerger[1]

June 2021

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 more influential in

---

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

[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."

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

The review is divided into five sections that each cover a body of relevant literature:

1) Direct Responses
2) Contagion
3) Counterspeakers
4) Descriptive Studies
5) Bystander Interventions

DIRECT RESPONSES

(Can counterspeech change the behavior of hateful speakers?)

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.

- **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).

- **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.

- **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.

- **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.

- **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.

THE CONTAGION EFFECT

(The impact of counterspeech on the audience)

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 Winter, Fabian. (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, & Taylor, Sean. (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. "Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions." Political Communication (2020): 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.

- **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).

- **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.

- **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).

COUNTERSPEAKERS

Adding nuance, some studies found evidence that certain specific variables pertaining to the counterspeaker, such as their race or level of influence, were important in determining whether the counterspeech was effective. Munger (2017) found that a speaker's perceived race and number of followers had an impact on the person's ability to persuade others to change their behavior. He found that "subjects who were sanctioned by a [bot representing itself as a] high-follower white male significantly reduced their use of a racist slur" (629). Seering et al. (2017) similarly found that messages coming from authoritative users on Twitch (moderators and paid subscribers) were imitated more frequently than those coming from less authoritative users.

- **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**

  Section 6.2 of this report is focused on 'credible messengers': survivors, former extremists, and others who have authority with the target audience. The authors argue that although these speakers are essential for effective counter-messaging, they often lack the capacity or networks to reach a large audience. Therefore civil society and governments should focus their efforts on helping the speech of credible messengers reach the target audience.

- **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.

- **Seering, Joseph, Robert Kraut, and Laura Dabbish. (2017). "Shaping pro and anti-social behavior on Twitch through moderation and example-setting." In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pp. 111-125.**
  https://dl.acm.org/doi/10.1145/2998181.2998277

  This study used an interrupted time series model to study behavior imitation (which we refer to in this review as 'contagion') on Twitch. The authors found that messages coming from authoritative users (moderators and paid subscribers) were imitated more frequently than those from less authoritative users.

DESCRIPTIVE 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.'

- **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.

- **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.

- **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.

- **Mathew, Binny; Punyajoy Saha; Hardik Tharad; Subham Rajgaria; Prajwal Singhania; Suman Kalyan Maity; Pawan Goyal; and Animesh Mukherjee. (2019) "Thou shalt not hate: Countering online hate speech." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, no. 01, pp. 369-380. https://arxiv.org/abs/1808.04409**

  The authors used multi-level annotation on a dataset of counterspeech comments from YouTube (n=13,924). They used the dataset to categorize counterspeech interactions and to support various insights about counterspeech; for example, counterspeech comments receive more likes and replies than non-counterspeech comments. The dataset is available to readers via a link in this article.

- **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.

- **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

although "hate is contagious," "counterhate messages can discourage users from turning hateful in the first place" (1).

BYSTANDER INTERVENTION

Bystander intervention research predates the internet, and the digital age has seen a wealth of research on "cyber-bystander intervention." This body of research asks why people choose to intervene — or not — against online bullying and harassment, and what the effects are on the harasser and the target of the harassment.

Bystander intervention is not the same as counterspeech. But we believe this literature may have useful lessons for counterspeakers, so we have included a selection of articles from it.

- **Allison, Kimberly R. and Kay Bussey. (2016). "Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying."** ***Children and Youth Services Review*****, *65*, 183–194. https://doi.org/10.1016/j.childyouth.2016.03.026**

  This is a literature review of studies on cyberbullying bystander behavior, covering a range of research to understand why some bystanders choose to get involved and others do not. The authors compare "the ability of two theoretical frameworks (the bystander effect and social cognitive theory) to account for" the findings of the study, and argue that "although the bystander effect is the dominant paradigm for explaining bystander inaction in many contexts, social cognitive theory may be better able to capture the complex and contextually dependent nature of cyberbullying situations" (183).

- **Dillon, Kelly P., and Bushman, Brad J. (2015). "Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context."** ***Computers in Human Behavior*****, *45*, 144–150. https://doi.org/10.1016/j.chb.2014.12.009**

  The authors of this paper conducted an experiment to test whether offline theories of bystander intervention apply to online environments — namely whether noticing a cyberbullying incident predicts intervention. They find that it does, although the majority of interventions (68%) are indirect and come after the threat has passed. This suggests that visibility of online harms is an important factor in determining whether people intervene.

- **Markey, Patrick M. (2000). "Bystander intervention in computer-mediated communication."** ***Computers in Human Behavior*****, *16*(2), 183–188.** https://doi.org/10.1016/S0747-5632(99)00056-4
  An early example of cyber-bystander research, this study established that the number of people present in a chat group was inversely related to the amount of time it took until one of them provided help to a target of harassment. This 'bystander effect' was eliminated when a target asked a specific person for help, addressing them by name.

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.