Counterspeech: A Literature Review

Cathy Buerger and Lucas Wright

November 2019

Introduction

Every day, internet users encounter hateful and dangerous speech online, and some of them choose to respond directly in order to refute or undermine it. We call this counterspeech. Many of those who have taken on this volunteer effort go about it alone, while others organize into groups to coordinate responses and support each other. Some staff at social platforms[1] have touted counterspeech as a method of reducing online hate, but, like Justice Louis Brandeis, who also opined that the remedy for bad speech is good speech,[2] they don't cite any evidence for their assertions. This is an effort to bridge that gap by answering what should be a prominent question: what does the scholarly literature have to say about the effectiveness of counterspeech?

That question begs another, of course, namely what it means for counterspeech to be effective, or successful. If one asks what it means to say counterspeech "works," one obvious answer is that it changes the beliefs or behavior of the original speaker, for example, persuading them to apologize or stop posting hateful messages. This is difficult, and most counterspeakers we interviewed say it is not, in any case, their primary goal. Far more often, counterspeakers define success as having a positive effect on the audience of their online exchanges - those that witnesses the interactions, regardless of whether the original speaker is affected. For example, counterspeech might be successful if it dissuades audience members from also speaking hatefully or galvanizes more counterspeech.

Only a few studies have attempted to measure the effectiveness of counterspeech directly, and as far as we know, this is the first review of relevant literature. We've collected and reviewed related articles from a range of fields including political science, sociology, countering violent extremism, and computational social science. These articles do not all use the term "counterspeech," but they shed light on various features of successful counterspeech, for example, qualities that make speakers/authors more influential in online interactions or the extent to which pro- and anti-social 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) The Counterspeaker

[1] In January 2016, speaking at the World Economic Forum, Sheryl Sandberg stated that "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

[2] In his concurring opinion in Whitney v. California (1927), Justice Louis 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."

4) Descriptive Studies
5) Bystander Interventions

Direct Responses (Can counterspeech change the behavior of hateful speakers?)

The studies in this section are all meant to gauge the effectiveness of direct counterspeech, where an internet user (a "counterspeaker") directly addresses another user in an effort to change their opinion or behavior. Very little research directly examines this topic. The studies that do exist have produced limited and somewhat contradictory findings. Miškolci et al. (2018), for example, found that although responding directly was not effective at stopping the behavior (i.e. posting hateful content) of the original speaker, it was a useful way to reach a larger audience and draw out more counterspeech. Schieb and Preuss (2016), however, argue that counterspeech can effectively reach the original speaker, but that the effectiveness of a counterspeech interaction is dependent on the size of the group of hateful speakers within a particular online space. For example, 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), for example, argue that the tone of a counterspeech message affects whether the interaction has a measurable impact. These studies demonstrate that when asking whether counterspeech is effective, one must pay special attention to the specific variables of the interaction (how many people are speaking, what is being said, and who is listening).

- **Bartlett, Jamie and Alex Krasodomski-Jones (2015). "Counter-speech: Examining content that challenges extremism online."** ***Demos*.**
  This report, commissioned by Facebook, examines how counterspeech that challenges far-right political Facebook pages in Europe is 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 it gives a sense of the reach of the content). The authors found that form matters; for example, of the different kinds of content posted, photos generate the most interaction (likes and comments). The researchers also found patterns related to tone, with "funny or satirical" counterspeech posts receiving the most interaction.

- **Frenett, Ross and Dow, Moli (2015). "One to one online interventions: A pilot CVE methodology."** ***Institute for Strategic Dialogue.***
  Based on a pilot study of far-right and Jihadist Facebook users determined by researchers to be at risk of becoming violent, 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, had a 100% no-response rate. 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.

- **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.*** Drawing on over 7,500 Facebook comments, this study uses qualitative content analysis to identify particular themes and terms used on Facebook to describe the Roma in Slovakia. It also tested the effectiveness of counterspeech to respond to these generally negative portrayals. The study found that counterspeech was not an effective method of changing the behavior of the user who posted negative comments about the Roma. It was, however, followed by an increase in the number of pro-Roma comments within a particular comment thread.
- **Schieb, Carla, and Mike Preuss. (2016) "Governing hate speech by means of counterspeech on Facebook." In 66th ICA Annual Conference, at Fukuoka, Japan, pp. 1-23.**

  The authors use a computational simulation model to determine factors that impact the effectiveness of counterspeech on Facebook. Their findings suggest that the proportion of counterspeakers to hateful speakers as well as the intensity of opinion held by the hateful speakers are both important determinants of success.

The Contagion Effect (What impact does counterspeech have on the larger audience?)

Counterspeech can also be studied for its effect on a wider audience. While few studies have examined such an effect explicitly, researchers have studied how behavior spreads online through the lens of behavior modeling, imitation, and descriptive norm adoption. These studies ask: "Does exposure to pro- (or anti-) social posts make other internet users more likely to speak in a similar way?" Social psychologists call this effect "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; Kwon & Gruzd 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 ("scolding incivility and encouraging civility") doesn't increase civility but does engender more metacommunication.

These findings have important ramifications for counterspeakers, as they demonstrate that the style and tone of responses can influence the behavior of others and potentially improve the quality of a discussion.

And because some research finds that anti-social 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 who participates or changes in the quality of the contributions. 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 have otherwise posted? Berry and Taylor (2017) analyzed historical data of participants to answer this question and found that the change in discussion quality they detect is 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 study.

- **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.**
  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 (ranked chronologically or by engagement) on the quality of comments shown to users and the quality of user comments in response. The sample included 100,000 comments drawn from the 5,000 largest English-language Facebook pages. The authors found that, on average, social treatment ranking results in high quality visible comments and, among the users who choose to participate, viewing higher quality comments increases the quality of subsequent contributions. They 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.**
  In this online experiment, researchers exposed participants to either a positive or negative stimulus prior to asking them to read an article with a comment section discussion that was either benign or "troll-like" - and then write their own comment. They 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.

- **Molina, Rocío Galarza, and Freddie Jennings. (2018). "The Role of Civility and Metacommunication in Facebook Discussions."** ***Communication Studies*****,** ***69*****(1), 42–66.**
  This study uses 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 encouraging civility, and a control group with no comments. Results show that exposure to civility and metacommunication increases participants'

willingness to write a comment and that their comments were more likely to be modeled on the condition comments (i.e. civility begets civility, metacommunication begets metacommunication.)

- **Han, Soo-Hye, and LeAnn M. Brazeal (2015). "Playing Nice: Modeling Civility in Online Political Discussions."** ***Communication Research Reports*****,** ***32*****(1), 20–28.**
  This online experiment also found that exposure to civility increases willingness to participate and heightens civility in a participant's comment. The researchers did not find an effect of incivility on participant 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).**
  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." They did not find any relationship between exposure to an uncivil discussion containing metacommunication and comment civility, but exposure 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.**
  In an online experiment, researchers exposed treatment groups to conversations with varying proportions of uncivil comments. It was a between-subject design, so each group was only exposed to 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' hostile cognitions (measured through the 21-question State Hostility Scale).

- **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.**
  This observational study of Twitch live chatrooms used an interrupted time series model, where data is collected at several, equally-spaced points of time, to measure imitation effects for pro-social behavior, anti-social behavior, and questions. Results show that all three types of behavior result 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).

The Counterspeaker

Adding nuance, some studies found evidence that certain specific variables related to the counterspeaker, such as their race or level of influence, were important in determining whether or not the counterspeech was effective. Munger (2017) found that a speaker's race and number of followers had an impact on how the person's ability to persuade another person to change their behavior. His study found that 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 rebuked by what appeared to be a black counterspeaker or a white speaker with fewer followers. 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.

- **Berger, J.M., and Bill Strathearn. (2013). "Who matters online: Measuring influence, evaluating content and countering violent extremism in online social networks."** ***The International Centre for the Study of Radicalisation and Political Violence.***
  The authors developed a scoring system to study the influence of white supremacists on Twitter. Users were given an "influence" score (how frequently their tweets received a measurable reaction), and "exposure" score (the likelihood that a user will respond to another person's tweet), and a combined "interaction" score. The study found that influence was highly concentrated among a few users (50% of measurable influence came from 1.5% of users).

- **Briggs, Rachel and Sebastian Feve. (2013). "Review of programs to counter narratives of violent extremism."** ***Institute of Strategic Dialogue*****.**
  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.**
  The author used an experiment to test the impact of identity and social status on successful group norm promotion. The experiment used an intervention to rebuke accounts that had used anti-black slurs on Twitter. Munger used bots variously identified as black or white and as high- or low-status (many vs. few followers) and documented the difference in reaction. He found that 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 an apparent black counterspeaker or a white speaker 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.**
  This study used an interrupted time series model to study behavior imitation (which we refer to in this lit review as "contagion") on Twitch. The authors found that messages coming from

authoritative users (moderators and paid subscribers) were imitated more frequently than those coming from less authoritative users.

Descriptive Studies

This literature review is focused on studies that can help determine whether counterspeech is effective, but we also summarize a rich body of literature that is more descriptive in nature. These studies illuminate many types of interactions that fall under the term "counterspeech." They also reveal different ways of defining success in counterspeech interactions,. Some of these studies describe a limited number of detailed cases 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), as they provide examples of 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.**
  The article describes the responses to anti-Muslim ads placed in the public transportation systems of three cities: New York, Detroit, and Washington, D.C. It argues that although counterspeech is not always effective (for example, if may have limited utility in communities where most of the people support the hateful speech), it educates the public about the issue, allowing for anti-hatred coalitions to form within communities, and therefore should be viewed as a positive remedy for harmful speech.

- **Benesch, Susan; Derek Ruths, Kelly P. Dillon, Haji Mohammad Saleem, and Lucas Wright. (2016). "Counterspeech on Twitter: A Field Study." Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/**
  A qualitative analysis of counterspeech interactions found on Twitter. The authors categorize counterspeech conversations by the number of people involved, describing four different "vectors": one-to-one, one-to-many, many-to-one, and many-to-many. The authors also enumerate a variety of potentially effective counterspeech strategies that that they have gleaned from their analysis.

- **Briggs, Rachel, and Sebastian Feve. (2013). "Review of programs to counter narratives of violent extremism."** ***Institute of Strategic Dialogue*****.**
  Largely focused on what governments might do to counter extremist messaging, the report divides what the authors call "counter-messaging" into three categories: government strategic communication, alternative narratives, and counter-messaging. This distinction is useful for thinking about the different goals and audiences of counterspeech interactions. The article

concludes with an appendix of a 18 case studies illustrating each form of counter-messaging. The case studies include examples of both online and offline efforts, primarily from Europe and the United States.

- **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.**
  The article's findings are based on an annotated dataset of counterspeech comments from YouTube (n=13,924) that the authors assembled using multi-level annotation. The article contains a link to the dataset, which is available to all readers. They use this dataset to categorize counterspeech interactions and to support various insights about counterspeech related to its effectiveness and "linguistic structure."

- **Richards, Robert D., and Clay Calvert. (2000). "Counterspeech 2000: A New Look at the Old Remedy for Bad Speech."** ***BYU L. Rev.***
  This article uses five case studies of offline counterspeech to examine whether, and under what conditions, it might serve as an effective remedy to harmful speech. The case studies highlight counterspeech campaigns organized by groups or companies, and the authors define harmful speech broadly, ranging from speech supportive of the Klu Klux Klan to messages that damage the reputation of a business. The article argues that large-scale counterspeech campaigns are most effective when they are able to leverage media connections in order to increase their audience.

- **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.**
  The article uses case studies to outline a "spectrum of force" of feminist counterspeech, with responses ranging in level of force directed against the speaker from negative comments made directly to misogynist speakers to positive messages supporting targets of misogyny. 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.**
  A condensed version of the aforementioned Benesch et al. (2016), 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" varies, at least in part, due to the number of people involved in the conversation. The article goes through each of the four vectors in detail, explaining factors related to each that might impact the effectiveness.

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 against cases of 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 as well, so we have included a select number of articles related to the topic.

- **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.**
  This is a literature review of research on cyberbullying bystander behavior, covering a range of research to understand why some bystanders choose to get involved. The review also offers recommendations on how to increase bystander interventions.

- **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.**
  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 it is countered.

- **Markey, Patrick M. (2000). "Bystander intervention in computer-mediated communication."** ***Computers in Human Behavior*****,** ***16*****(2), 183–188.**
  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 a target of harassment received help. This "bystander effect" was eliminated when a bystander was addressed by name in a call for help.

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. As this review demonstrates, the field of counterspeech research is emerging and largely undefined. We would therefore 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.