DANGEROUS SPEECH PROJECT

Why They Do It:

Counterspeech Theories of Change

Cathy Buerger, Ph.D.

September 26, 2022

DangerousSpeech.org

Table of Contents

Introduction

Around the world, tens of thousands of people regularly feed the trolls: they respond directly to online expressions of hate in an effort to improve discourse. We call them "counterspeakers." In our research at the Dangerous Speech Project (DSP), we have found many who have persisted at it daily or weekly, for years. Some do it alone, while others work in large groups to respond to online speech that they consider hateful (such as xenophobia, racism, and sexism) collectively, following codes of conduct and buoying each other's resolve to keep at it, though their efforts are generally unpaid, laborious, and emotionally taxing.

Their principal goal is nearly unanimous - and it is also different from what many researchers assume. The vast majority of the dozens of subjects I interviewed said they are not trying to reach (that is, change the minds or behavior of) the people to whose posts they respond. Instead, they want to reach the spectators – the people who read what they regard as hateful posts and the counterspeakers' responses. The spectators, they note, usually far outnumber people who post hateful content. This is a unique feature of online communication; the audience for any publicly-visible comment is often large, in part because people can continue to read it long after it was originally posted.

Their reasons for addressing spectators with counterspeech are grounded in four primary theories of change:

1) Some hope to change the spectators' views.
2) Some attempt to reach those who agree with them but don't yet dare to express those views online, since recruiting new counterspeakers would increase the amount of counterspeech.
3) Others choose to amplify negative content in order to strengthen norms against it among audience members.
4) Still others combine one or more of these three theories of change with an explicit effort to support those targeted by the hateful speech. In doing so, they seek to mitigate the negative impacts of the speech.

For some, their understanding of how best to counterspeak has transformed over time – their theories of change emerge and evolve in response to their experiences. This paper reports counterspeakers' theories of change as they describe them, and discusses the implications of this for counterspeech researchers.

Literature Review

There is a growing body of literature examining online responses to hatred, but there are only a handful of studies on the counterspeakers themselves.[1] This work is situated within the larger and better-developed body of work on online behavior change.[2] A tiny but promising handful of papers have explored the effectiveness of specific counterspeaking efforts.[3] This is the first paper to examine the goals of counterspeakers.

Previous scholarship on online responses to hatred has generally focused on the issue of effectiveness, most frequently defining success as the capacity of counterspeech to change the beliefs or behavior of the person to whom it responds, i.e. persuading them to apologize or stop posting harmful messages. Researchers have concluded that this is very difficult to achieve. For example, Miškolci et al. observed that responding directly did not stop the behavior (posting hateful content) of the original speaker.[4] Others have found that direct responses to hateful or harmful speech can occasionally change someone's online behavior. The effectiveness of a response was strongly dependent on factors such as the proportion of counterspeakers to hateful speakers,[5] whether they

---

[1] Buerger, Catherine. 2021. "# iamhere: Collective Counterspeech and the Quest to Improve Online Discourse." *Social Media+ Society* 7, no. 4 (2021): 1-17 https://doi.org/10.1177/20563051211063843. Ziegele, Marc, Teresa K. Naab, and Pablo Jost. "Lonely together? Identifying the determinants of collective corrective action against uncivil comments." *New Media & Society* 22, no. 5: (2020) 731-751. https://doi.org/10.1177/1461444819870130

[2] For examples, see Khan, M. Laeeq. "Social media engagement: What motivates user participation and consumption on YouTube?." *Computers in human behavior* 66 (2017): 236 - 247. https://doi.org/10.1016/j.chb.2016.09.024 Lewandowsky, Stephan, John Cook, Nicolas Fay, and Gilles E. Gignac. "Science by social media: Attitudes towards climate change are mediated by perceived social consensus." *Memory & cognition* 47, no. 8 (2019): 1445-1456 https://doi.org/10.3758/s13421-019-00948-y; Shahbaznezhad, Hamidreza, Rebecca Dolan, and Mona Rashidirad. "The role of social media content format and platform in users' engagement behavior." *Journal of Interactive Marketing* 53 (2021): 47-65. https://doi.org/10.1016/j.intmar.2020.05.001

[3] 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* 38, no. 5 (2021): 624-646. https://doi.org/10.1080/10584609.2020.1830322 Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. "Countering hate on social media: Large scale classification of hate and counter speech." *Association for Computational Linguistics*. (2020): 109, http://dx.doi.org/10.18653/v1/2020.alw-1.13.

[4] Miškolci, Jozef, Lucia Kováčová, and Edita Rigová. "Countering hate speech on Facebook: The case of the Roma minority in Slovakia." *Social Science Computer Review* 38, no. 2 (2020): 128-146. https://doi.org/10.1177/0894439318791786

[5] Schieb, Carla, and Mike Preuss. "Governing hate speech by means of counterspeech on Facebook." In 66th ICA annual conference, at Fukuoka, Japan, (2016): 1-23, https://www.researchgate.net/publication/303497937_Governing_hate_speech_by_means_of_counter speech_on_Facebook [https://perma.cc/KB42-QX3V]

were counterspeaking as part of a group,[6] the tone used by a counterspeaker,[7] or even specific characteristics of the people doing the counterspeaking – such as their race or perceived popularity.[8] These studies are useful, but they capture only a small sliver of the picture, by focusing on just one of the possible goals of counterspeakers.

This paper argues that there is another question related to the effectiveness of counterspeech that is potentially more important for researchers: what role do spectators inspired by counterspeakers play in changing online discourse? Several studies serve as a useful jumping off point, though they do not ask this question explicitly. Some, for example, have examined how particular types of speech (e.g., pro- or anti-social speech) spread online by means of behavior modeling, imitation, and descriptive norm adoption. Han and Brazeal found that people exposed to civil comments (which they define as being readily perceived as "reasonable and courteous" even by those who disagree with them),[9] 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.

Masullo et al. studied the spectator reactions to comments intervening in disagreements between users on the Facebook pages of newspapers.[10] They found that spectators rated intervening comments that used a "high-person-centered response" (one that

---

[6] Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. "Impact and dynamics of hate and counter speech online." *EPJ data science* 11, no. 1 (2022): 3. https://arxiv.org/abs/2009.08392 [https://perma.cc/4EAV-HFV6] ; 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* 38, no. 5 (2021): 624-646. https://doi.org/10.1080/10584609.2020.1830322

[7] Bartlett, Jamie, and Alex Krasodomski-Jones. "Counter-speech examining content that challenges extremism online." *DEMOS, October* (2015). https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf [https://perma.cc/2KPE-R6TJ]; Frenett, Ross, and Moli Dow. "One to one online interventions: A pilot CVE methodology." Institute for Strategic Dialogue (2015). https://www.isdglobal.org/isd-publications/one-to-one-online-interventions-a-pilot-cve-methodology/ [https://perma.cc/5CQX-927X]; Ziegel, M. Jost, P., Bormann, M., and Heinback, D. 2018. Ziegele, Marc, Pablo Jost, Marike Bormann, and Dominique Heinbach. "Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments." *SCM Studies in Communication and Media* 7, no. 4 (2018): 525-554. DOI:10.5771/2192-4007-2018-4-525

[8] Munger, Kevin. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39, no. 3 (2017): 629-649. https://doi.org/10.1007/s11109-016-9373-5 ; Seering, Joseph, Robert Kraut, and Laura Dabbish. "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, (2017):111-125. https://dl.acm.org/doi/10.1145/2998181.2998277

[9] Han, Soo-Hye, and LeAnn M. Brazeal. "Playing nice: Modeling civility in online political discussions." *Communication Research Reports* 32, no. 1 (2015): 23. https://doi.org/10.1080/08824096.2014.989971

[10] Masullo, Gina M., Marc Ziegele, Martin J. Riedl, Pablo Jost, and Teresa K. Naab. "Effects of A High-Person-Centered Response to Commenters Who Disagree on Readers' Positive Attitudes toward A News Outlet's Facebook Page." *Digital Journalism* 10, no. 3 (2022): 493-515. https://doi.org/10.1080/21670811.2021.2021376

acknowledges people's emotions) more favorably than those that used "low-person-centered" speech. In a second study on spectator perception of newspaper comments, Ziegel and Jost found that "factual responses to uncivil comments increased observers' perceptions of a deliberative discussion atmosphere," which, in turn, increased their willingness to add their own comments to the thread.[11]

Two studies (Molina & Jennings and Han, Brazeal & Pennington) found that metacommunication comments (those that address the tone of a comment rather than its content, such as when a user scolds someone else for incivility rather than commenting on the opinions being expressed) do not increase civility but do engender additional metacommunication comments. Similarly, Miškolci et al. found that counterspeech comments seem to trigger additional counterspeech from the audience. The findings from my previous ethnographic work on the #iamhere network, a large counterspeaking network with over 150,000 members in at least 17 countries, coincided with this finding and provided some initial clues as to why this may be the case. Seeing someone document their dissent to a statement lowers the counterspeaker's epistemic load, that is, the amount of confidence one must feel in one's own opinions being right (or someone else's being wrong) to become willing to enter a discussion about the topic. Also counterspeakers perceive less risk of online retribution from the people to whom they respond, when they are not alone in speaking out.

Data

In this paper I report findings from several years of empirical study of responses to hatred online. At the DSP we have documented many individuals and groups from around the world responding to online content that they believe is hateful,[12] and I have interviewed over 50 counterspeakers. We located them in many ways – through online observation, mentions in the press, and word of mouth. The semi-structured interviews were done over Zoom, using a base list of questions that was adapted and expanded to capture the particularities of each effort.

After dozens of hours of observation and interviews, I conducted and published the first ethnographic study of a group of counterspeakers.[13] My findings are also informed by data collected during an International Counterspeakers' Workshop that the DSP organized and hosted in Berlin in 2018. Participants from Brazil, Canada, France, Germany, India, Israel, Slovakia, Sweden, and the United States gathered to share their

[11] Ziegele, Marc, and Pablo B. Jost. "Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments." *Communication research* 47, no. 6 (2020): 891-920. https://doi.org/10.1177/0093650216671854

[12] We have not tried to evaluate these judgements.

[13] Buerger, Catherine. "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse." *Social Media+ Society* 7, no. 4 (2021): 20563051211063843. https://journals.sagepub.com/doi/pdf/10.1177/20563051211063843

experiences fighting hatred online and to discuss what they found to be best practices for counterspeech.

The efforts included in this study are not limited to those we see as "effective," as one generally cannot know whether a response was effective without a separate study, and as noted above, effectiveness can be defined in many ways. The counterspeech I studied embodies a range of communicative and rhetorical techniques, including empathy, humor, and providing support to those targeted by hateful speech. Some efforts were planned in advance, while others emerged spontaneously, usually in response to a surge of hateful content online. Interviewees described their motivations and strategies, their challenges, and their definitions of "success." I sought to understand the goals of the counterspeakers, asking all of them whom they were trying to reach and why. In the sections below, I describe their theories of change in detail, providing examples for each category.

Making impact through the "Silent Crowd"

Most of the counterspeakers I interviewed said that their primary audience was not the original speaker, but others who encounter their response online - a group one counterspeaker called the "silent crowd." Members of this group are less visible online (and to researchers) since they are primarily reading content instead of adding it.

Counterspeakers described multiple reasons for trying to reach this audience instead of the original speaker. Many said that they simply did not think it was a good use of time to engage with those posting hateful comments online, as the chances of changing their behavior were low (an opinion supported by the literature). As one member of #iamhere explained, "the trolls will not be affected. They get energy from being debated with. It's other people that you try to stop from joining in on the hateful speech."[14]

Counterspeakers address the audience not only for the negative reason that they do not think they can convert trolls, but also for the positive reason that the silent crowd is a useful audience. Counterspeakers believe they can reduce the overall negative impact of hatred online in several ways: by educating and even changing the minds of people in the audience whose views are neither extreme nor entrenched, by encouraging others who may share their beliefs to become counterspeakers, or by publicizing the speech to bring attention to it and strengthen norms against it. These three categories are related, and in many cases, a group or person may see more than one of them as a pathway to improve online discourse. There are also several counterspeech efforts that combine one

[14] Interview with author, September 18, 2019.

of these with an explicit goal of providing support for spectators from the groups targeted by the hateful speech. In the following sections, I explore each of these theories of change.

1. The "movable middle"

By far the most common reason why counterspeakers said they thought "silent" readers are their most important audience is that some of these readers, group members posit, have not yet made up their mind about the topic being discussed, and therefore can be swayed toward one side or the other. Counterspeakers hope to sway the opinions of the "movable middle" against what they consider hateful content and toward their own beliefs. There is support for this strategy in the literature. For example, researchers have found that even a small group of counterspeakers can influence the discourse within an online space if the audience they are addressing holds relatively moderate views.[15]

The Lithuanian Elves (named such, according to their founder, "because elves fight trolls")[16] were founded in 2014 to counter Russian government disinformation campaigns that criticize democratic institutions and attempt to stoke intergroup conflict. The group began in Lithuania with 20 volunteers, and it has since expanded to 13 Central and Eastern European countries, by March 2022 reportedly growing to over an estimated 22,000 members.[17][18] In one example, a rumor (widely believed to be Russian in origin) spread that a German NATO soldier had raped a Lithuanian teenager.[19] The unquestionably false rumor first emerged in 2017, reemerging again during the 2022 Russia-Ukraine war.[20] The Elves work collectively to counter disinformation such as this to push back against Russian propaganda that claims Lithuania is a failed state and promotes a return to Russian control of the country.

---

[15] Schieb, Carla, and Mike Preuss. "Governing hate speech by means of counterspeech on Facebook." In 66th ICA annual conference, at Fukuoka, Japan, (2016): 1-23.
[16] Abend, Lisa. "Meet the Lithuanian 'Elves' fighting Russian disinformation." *Time,* March 6, 2022. https://time.com/6155060/lithuania-russia-fighting-disinformation-ukraine/ [https://perma.cc/5E5A-6CM7]
[17] *Id.*
[18] Because of the decentralized structure of the group and the fact that members participate anonymously, it is difficult to determine the exact number of members.
[19] Deutsche Welle. "Russia's information warfare targets German soldiers in Lithuania." *NATO Source*. February 24, 2017. https://www.atlanticcouncil.org/blogs/natosource/russia-s-information-warfare-targets-german-soldiers-in-lithuania/ [https://perma.cc/E3DZ-SVSQ]
[20] Nordstrom, Louise. "'We're at war': The 'Lithuanian Elves' who take on Russian trolls online." *France 24*. January 23, 2022. https://www.france24.com/en/europe/20220123-we-re-at-war-the-lithuanian-elves-who-take-on-russian-trolls-online [https://perma.cc/ANX3-KUSF]

As one member of the Elves said in an interview, the group's goal is to reach the "common people." "We hope they are understanding what they are reading. We are trying to make them not be poisoned by the fake news and the propaganda," he said.[21] Members of the Elves don't attempt to change the behavior of coordinated trolls backed by the Russian government. Instead, they attempt to put out enough counterspeech that those reading the comments identify the disinformation and recognize it as false.

An even larger multinational coordinated effort to reach online spectators is #iamhere. The #iamhere network is, to our knowledge, the largest and most well-organized collective counterspeech effort in the world. The network, which began in Sweden in 2016, has around 150,000 members responding to hatred through 17 country-level Facebook groups (Australia, Bulgaria, Canada, Czech Republic, Estonia, France, Finland, Germany, India, Italy, Norway, Poland, Slovakia, Spain, Sweden, the United Kingdom, and the United States).

Members of the groups seek out hateful speech in comment threads of news articles posted on Facebook and then respond together, following a strict set of rules which require a respectful and non-condescending tone and prohibit spreading prejudice or rumors. Members also "like" each other's comments, pushing them to the top of comment threads, since Facebook ranks comments on public pages based on interactions ("likes" and replies).[22] This is a vital feature of #iamhere's model: they make use of Facebook's system to amplify their own civil, fact-based comments and bury hateful or xenophobic comments at the bottom of comment threads, making it less likely that others will see them.

Some #iamhere counterspeakers said they simply do not believe they would be able to successfully change the minds of those posting hatred, but that they will be able to have some impact on those who hold more moderate views. They observe that dispelling myths and making accurate information easily visible allows the readers to make up their own minds.

One said he doesn't counterspeak to make people see that they are wrong, but to show that there are different, non-hateful views. "These comment fields can make the impression that most people are hateful; they're not," he stated.[23] Another member shared a similar viewpoint: "Even if you write an answer for that side [those posting hatred], everyone else can read it too. If you go into a place where a lot of bad things are written, then people say, 'oh, God! That is what everyone thinks!' But this is not what

[21] Interview with author, October 12, 2021.
[22] Facebook Help Center. 2022. https://www.facebook.com/help/539680519386145 [https://perma.cc/QAG4-6374]
[23] Interview with author, August 27, 2019.

everyone thinks. A lot of people think differently; and that's important."[24] By providing factual information written in a civil tone, #iamhere participants hope to persuade readers not to be swayed by hateful content or misinformation.

In a similar vein, some scholars have started using social media – especially Twitter – to challenge misinformation in their area of expertise. One example that has become well known in the public conversation of race in America is Kevin Kruse, a history professor at Princeton University. Kruse writes lengthy threads correcting popular misrepresentations of American history on topics such as slavery, the drivers of the Civil War, and lynching.[25]

For several years, Kruse has had an ongoing exchange with conservative political commentator Dinesh D'Souza, who frequently pushes far-right conspiracy theories and uses historical references to promote the Republican Party. For example, D'Souza has argued that the current Democratic Party is racist because Democrats supported slavery in the 1860s – ignoring the party realignment that happened during the civil rights movement. Kruse, in response, writes extensive Twitter threads, fact-checking D'Souza (and many others), citing historical evidence and providing links to sources. Kruse has in the past explained that he doesn't respond to D'Souza with the hope of changing his mind, but rather he "does it for the people who encounter these counternarratives and think that they're wrong, but don't know enough to back it up; people who need a counterpoint to their uncle at Thanksgiving, or to their co-worker in the breakroom."[26]

As other counterspeakers have mentioned, documenting dissent can be a very powerful form of counterspeech, especially when hateful speech and misinformation have gone unquestioned in the past. When the dissenting voices are authorities in their fields, the counterspeech can carry even more weight.

2. Encouraging More Counterspeech

In addition to persuading members of the "movable middle," #iamhere members described another way that their method may be able to fight against hatred online: increasing the amount of counterspeech and, importantly, the number of counterspeakers in a particular space online.

---

[24] Interview with author, September 12, 2019.
[25] https://twitter.com/kevinmkruse [https://conifer.rhizome.org/dangerousspeech/theories-of-change/20220927181550/https://twitter.com/kevinmkruse]
[26] Pettit, Emma. "How Kevin Kruse Became History's Attack Dog." *The Chronical of Higher Education*. December 16, 2018. https://www.chronicle.com/article/how-kevin-kruse-became-historys-attack-dog/ [https://perma.cc/P9WK-TQBW]

During interviews, many members of the group told stories of their own experiences joining #iamhere – how they had felt alone and hesitant to speak against hatred they were seeing online. Many said that they did not counterspeak before joining the group. They were disgusted by the comments that they were reading, but felt too afraid to say anything. The emotional labor of counterspeaking was described by many interviewees, who said they found it especially hard when they did it alone. A solitary dissenting voice can draw attention and potentially garner attacks. But with the #iamhere model, members counterspeak as a group, leaving each individual minimally exposed within a comment thread. Members said this left them feeling safer or more protected.

When working within a comment thread, members of #iamhere attempt to elevate not only other group members' comments, but also civil, fact-based comments written by people not in the group. A moderator for #jagärhär, the Swedish group, recounted counterspeaking on her own before she joined the group in December 2016, only a few months after it began. One day, after she responded to a Facebook post denigrating asylum seekers, some of the commenters began attacking her. "I don't remember exactly what they said, but I remember it was aggressive, and that I didn't know exactly what I should do. I thought, should I keep responding? Should I just keep quiet?"[27] Before she had made up her mind, she noticed that others had joined her. People started "liking" her comment and others cited statistics about immigration, trying to refute claims that refugees were a danger to Sweden. The comments included a hashtag: #jagärhär. "I looked it up, and I decided to join. It was just in time. I had started losing some faith that responding was worth it."[28] "To see so much hate, that can eat you up at times," she said. After finding the group, she felt more hopeful. "I thought that I could make a difference with other people. We could do this together."[29]

Supporting the comments of non-group members has the extra benefit of aiding in recruiting and encouraging more participation in online discussion. As one member put it:

> In the end, it's about democracy, it's about debate, it's about freedom of speech that people will have the courage to say what they think. If you have lots of hate comments, maybe you are afraid, and you don't want to say what you think. But if we are 10-20 people arguing against the hate then I imagine that others will also want to do so, so that not only the people screaming the highest can say their opinion.[30]

[27] Interview with author, August 19, 2020.
[28] Interview with author, August 19, 2020.
[29] Interview with author, September 18, 2019.
[30] Interview with author, October 18, 2019.

There is some quantitative evidence to bolster the belief of #iamhere members that they are having some tangible impact. A team of Slovakian researchers collected and coded 60 comment threads (7,500+ comments) from Facebook related to the Roma. In half of the comment threads, they interjected counterspeech responding to anti-Roma comments, and in the other half, they did nothing. Where they intervened, there were significantly more pro-Roma comments following their counterspeech than in the control threads. As they noted, "If we enter Facebook discussions with pro-Roma comments, it motivates other followers of those particular Facebook profiles to join the discussion arguing in favor of the Roma as well."[31]

On the other hand, research on bystander intervention has shown that the presence of many willing interveners (either online or offline) can decrease feelings of individual responsibility to get involved, leading to people being less willing to intervene themselves.[32] It is likely that individual motivations for getting involved in counterspeech play a role in one's interpretation of the presence of other counterspeakers, but more research is needed to better understand this relationship.

3. Raising awareness about hateful content

Some counterspeakers choose to bring content that they believe to be hateful into more prominent public view so that more people become aware of it. At the DSP, we call this strategy "amplification." Those who use it often take content from a small online forum and post it to a larger one where more people will see it. At first glance, this seems counterproductive: why expose more people to content you're trying to counter? The argument for doing so is that it obliges members of a majority to recognize how their society attacks people different from themselves. By naming the content as harmful, counterspeakers attempt to uphold or even shift norms against it.

The Brazilian campaign Mirrors of Racism is an example of very literal amplification. In 2015, when the journalist Maria Julia Coutinho (known by her nickname Maju) became the first Black weather broadcaster for a prime time news show, *Jornal Nacional,* some Brazilians reacted with a torrent of racism against her and other Black Brazilians. In response, Criola, a Black women's civil rights organization in Brazil, partnered with the advertising firm W3haus to create an anti-racism campaign. They collected some vivid, crude racist comments, suggesting that Black people smell bad for example, and posted

[31] Miškolci, Jozef, Lucia Kováčová, and Edita Rigová. "Countering hate speech on Facebook: The case of the Roma minority in Slovakia." *Social Science Computer Review* 38, no. 2 (2020): 128-146. https://doi.org/10.1177/0894439318791786

[32] Obermaier, Magdalena, Nayla Fawzi, and Thomas Koch. "Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying." *New Media & Society* 18, no. 8 (2016): 1491-1507. https://doi.org/10.1177/1461444814563519

them in huge letters on billboards in five Brazilian cities, in the neighborhoods of the people who had posted the comments online. Each billboard also bore the phrase "Racismo virtual, consecuencias reales" ("Virtual racism, real consequences").

"The strategy of the campaign was to take internet racism out of the internet and expose it in the streets so that the population (of the region) could become aware of the damage caused by these virtual acts," said Criola's General Coordinator, Lúcia Xavier.[33]

To further amplify the content and the campaign itself, W3haus interviewed Brazilians about it, and posted the resulting videos. In one, passersby on the street react to a billboard.[34] One middle aged white man says, for example, that Brazilians forget that racism exists, but the billboard brings attention to it. In another video, the author of one of the racist posts stands in front of a billboard emblazoned with it – and his own blurred profile photo – and apologizes to a Black woman.[35] Posted online, these videos further amplified the racist content and the campaign beyond the communities where the billboards were located.

Another counterspeech campaign that used amplification is #MoreThanMean. In 2016, the hosts of the podcast "Just Not Sports," decided to try an innovative method to diminish the sort of harassment and contempt that women sports journalists frequently receive. ESPN's Sarah Spain and Julie DiCaro – two such journalists – agreed to help make a video to amplify the problem. With the video, Brad Burke, one of the co-creators of #MoreThanMean, hoped to shock the silent crowd of male sports fans into changing their behavior (by stopping harassing women or by speaking up when they see other men doing it). He described the set-up during an interview:

> I got a bunch of sports fans, and all I told them was we're going to do a mean tweets thing[36] where you read tweets to two Chicago sports reporters. I think they just thought it was going to be funny. I gave them an iPad so they could only see one tweet at a time. The first few were just lighthearted and then it takes a turn. It drops an F bomb and the C word and then it was like the air just got sucked out of the room. They became aware that this wasn't going to be fun.[37]

---

[33] Interview with author, October 31, 2017.
[34] https://vimeo.com/150728678 [https://conifer.rhizome.org/dangerousspeech/theories-of-change/20220927181243/https://vimeo.com/150728678]
[35] https://www.behance.net/gallery/63075249/Criola-Mirrors-of-Racism [https://conifer.rhizome.org/dangerousspeech/theories-of-change/20220927171037/https://www.behance.net/gallery/63075249/Criola-Mirrors-of-Racism]
[36] American talk show host Jimmy Kimmel often includes a segment on his show called "Mean Tweets" where celebrities read aloud tweets written by ordinary people insulting them.
[37] Interview with author, February 12, 2018.

The video shows the men – each sitting facing either Spain or DiCaro and looking them in the eye as Burke had instructed - struggling to read actual messages that the women had received from readers, such as "You need to be hit in the head with a hockey puck and killed," and "I hope you get raped again." The 3-minute video ends with the words "We wouldn't say it to their faces. So let's not type it."

Asked what they were hoping to accomplish with the video, Burke explained,

> We were never going to reach the hardest of hard-core trolls. The guy who threatens to rape Julie DiCaro isn't someone worth talking to. I wanted to reach the guy who scrolls past this stuff who doesn't think about it or the sports fan who reads it and piles on - like the last guy in the tackle. They don't realize it's a human interaction for the people taking part in the exchange. I heard from a lot of people who say 'I see this stuff, and I don't think to do anything.' If you're just online seeing Sarah Spain ribbing some guy then you might think she's in on it, but we need them to know that this is more than mean.[38]

Other counterspeakers also use amplification to reach the audience, including Logan Smith, founder of the Twitter account Yes, You're Racist. Smith founded the account in 2012, and at this writing it has over 320,000 followers. "It started out a light-hearted way to call out casual racism and calling attention to it," said Smith.[39] One day, he decided to search for Tweets containing the phrase "I'm not a racist, but…". "It was astonishing how many people were posting blatant casual racism often under their real names and thinking that it wasn't racist," Smith said.[40] So he decided to create a Twitter account devoted to retweeting their messages to call them out. He said that in the early years of Barack Obama's presidency, he heard many people claim that Obama's election proved that the United States was a post-racial nation. Smith said he wanted to make sure that "people knew racism still existed." As he said, "I'm a white man, I have no illusion that I'm the wokest person in the world or that I have any enlightened perspective on race. So where I've focused my effort is in helping white people like myself in recognizing casual racism."[41] Once they recognize it, Smith hopes they will be more likely to call it out when they see it.

Other Twitter accounts employ a similar strategy to Smith's, including Yes, You're Sexist (20,000 followers) which operates the same way with sexist Tweets, and Racism Watchdog (640,000 followers), where the account's founders retweet content they believe to be racist along with the words "bark!" or "woof!"

---

[38] Interview with author, February 12, 2018.
[39] Interview with author, August 10, 2018.
[40] Interview with author, August 10, 2018.
[41] Interview with author, August 10, 2018.

4. Supporting those targeted by hateful speech

Several counterspeech efforts combine one of the aforementioned theories of change with the goal of providing support to members of the audience who feel attacked by the speech. One example of this is White Nonsense Roundup (WNR). Started in 2015 and still active at this writing, WNR is a volunteer-run effort active on Facebook and Twitter that seeks to take some of the burden of counterspeaking – especially in response to people who claim they aren't racist – from people of color. As described by one of the project's founders:

> If someone is tired of explaining yet again why a certain statement is racist or why it's problematic to say "all lives matter" or to be colorblind and claim that you don't see color, what a great thing that we can do some of that work…It should be white people engaging in a lot of the labor and work of tackling our own skinfolks to figure this out.

Their model is therefore based around the goal of supporting the targets of racism by taking on some of the counterspeech labor that often rests on the shoulders of people of color. When people of color receive or see racist comments online, it is often left to them to respond. WNR seeks to remove at least some of that burden.

When someone tags them in a post, a volunteer reads through the thread and then figures out the best way to respond. Although volunteers at WNR sometimes seek to change the behavior of an author of racist speech with counterspeech, more frequently they attempt to reach the larger audience. As one of the WNR founders noted, "If no one says anything, and it's just this racist comment left sitting there, what are we supposed to think? That all white people agree with that? So it's important to counter that. Also, if there is someone else silently reading the conversation – they may learn something."[42]

Alexandra Tweten, who founded and runs the Instagram account Bye Felipe, also seeks to educate the audience while providing support to those targeted by hostile speech, although she uses very different methods. The account is dedicated to "calling out dudes who turn hostile when rejected or ignored."[43] She does so by sharing screenshots submitted by women of conversations they have had with men (often on dating apps) where men lashed out, responding with degrading or violent comments, after being ignored or rejected. The account was started in 2014 and as of August 2022, it had 449k followers.

---

[42] Interview with author, January 30, 2018.

[43] https://www.instagram.com/byefelipe [https://conifer.rhizome.org/dangerousspeech/theories-of-change/20220927170105/https://www.instagram.com/byefelipe/]

Tweten says that she started Bye Felipe as an inside joke - a way to provide space for women to "make fun of dudes,"[44] and the account still plays this role – there are generally hundreds of comments on each post mocking the men in the screenshots. Even the name "Bye Felipe" captures this original intent, inspired by the slang phrase "Bye, Felicia," used as a dismissive send off. "These guys are just trying to take women down, and I just wanted to flip the script and at the same time make women feel better,"[45] said Tweten. But after the account received some press in *The Atlantic* (an article that dubbed Tweten a "Feminist Tinder-Creep-Busting Web Vigilante")[46] near the end of 2014, Tweten started to think of her goal in a new way: "It was kind of like oh, maybe I can use this to start a national conversation about online harassment against women. Maybe I could show other men what it's like to be a woman," she said.[47] Through her account, Tweten attempts to reach the audience while also providing a space for women who have received messages like those posted to feel a sense of solidarity and mock the men involved – decreasing the power of the misogynistic speech.

The counterspeakers from WNR and Bye Felipe respond to hatred by speaking to the larger audience, but they do so with the explicit understanding that some members of that audience experience the speech differently. In the example of Bye Felipe, women may want a space to vent about their own similar experiences and find solidarity with others posting in the comments. For those who haven't experienced an interaction like the ones Tweten posts, the screenshots show them that this type of behavior occurs and clearly communicates that men should not treat women that way.

Other Pathways for Challenging Hatred through Counterspeech

Although most of the counterspeakers I interviewed said that their goal was to reach those in the "silent" or "reading" audience, there were some who had other objectives. A few did say that they try to reach the people who have posted hatred online. Some want to shame, punish, or simply annoy the person who posted hateful content. Others take an educational approach, seeking to change the behavior or even the beliefs of those to whom they respond.

---

[44] Interview with author, January 12, 2018.
[45] Interview with author, January 12, 2018.
[46] Khazan, Olga. "Rise of the Feminist Tinder-Creep-Busting Web Vigilante." *The Atlantic*. October 27, 2014. https://www.theatlantic.com/health/archive/2014/10/rise-of-the-feminist-creep-busting-web-vigilante/381809/ [https://perma.cc/L63H-SZVP]
[47] Interview with author, January 12, 2018.

People being shamed or "canceled" as a result of their speech or behavior has become commonplace. Sometimes referred to as digital vigilantism or "digilantism," these efforts – often organized around a hashtag - try to irritate, shame, or punish the original speaker.

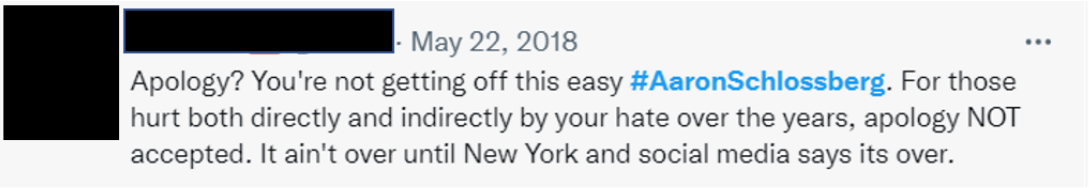
In 2018, Manhattan attorney Aaron Schlossberg was captured on video berating employees at a restaurant for speaking Spanish. In part of the tirade, Schlossberg said, "My guess is they're not documented. So my next call is to ICE to have each one of them kicked out of my country." "I pay for their welfare," he continued.[48] The video went viral on social media, and Schlossberg was quickly identified.

Writer and civil rights activist Shaun King shared the video on Instagram along with the following text:

> "Who this this [sic] bigot in Midtown Manhattan? What's his name? Please share this. Here he is harassing & insulting two women for speaking Spanish...TO EACH OTHER in the middle of Manhattan. Trump has empowered ugly white people like this to say whatever they feel like saying. My job is to make them known for it. Act like this and we will expose you to the world.[49]

The post was viewed over 250,000 times. Those who take part in digilantism do so for many different reasons. As this message shows, some do so in order to punish with the hopes of deterring others from similar behavior. As King wrote, "act like this, and we will expose you to the world."

After Schlossberg released an apology several days later, Twitter users reacted again, one writing:



As a result of the public outcry, the company that managed the building housing Schlossberg's private law firm decided to terminate his lease.

In addition to shaming Schlossberg on social media, internet users also doxxed him – releasing identifying information including his home address. NYC resident Luis

[48] https://www.instagram.com/p/Bi1oSD4g4v7/ [https://conifer.rhizome.org/dangerousspeech/theories-of-change/20220927164714/https://www.instagram.com/p/Bi1oSD4g4v7/]
[49] *Id.*

Magaña used social media to organize a Latin themed party to be held outside of Schlossberg's apartment. Amplified on Twitter through the hashtag #LatinPartyNYC, the idea took off. Hundreds of people attended the peaceful protest party where crowds enjoyed a mariachi band paid for by a GoFundMe page that raised over $1,000 for the event.

Employing a very different strategy, some counterspeakers try to educate the person to whom they are responding about why their speech is objectionable. One such counterspeaker is activist Dawud Walid. Longstanding executive director of the Michigan chapter of the Council on American-Islamic Relations (CAIR Michigan), Walid has written several books on the intersection between race, Islam, and activism. In 2014, Walid used his expertise to counterspeak, publishing a piece called "Fellow humans are not '*abeed*,"[50] ("*abeed*" means "slaves" in Arabic and is sometimes used as a slur to refer to Black people). Walid thought the term was becoming more common, offline and on social media. In the essay, he described the historical meaning of the word and how it evolved to be derogatory. Walid wanted to convince those who were using the term to stop.

After writing the article, Walid took to Twitter. "When I first got on Twitter it was to highlight anti-Muslim hatred as part of my job with CARE. I later came to the conclusion that in order for Muslims in America to push back against Islamophobia, we had to deal with our own inter-Muslim tribalism and racism," Walid said.[51] He searched for use of the word "abeed" as a slur, in Tweets written in English and Arabic. When he found it, he would reply with a link to his article, writing "please read this." Walid said he sent the article to several hundred people. Many, unsurprisingly, did not reply. For those who did, Walid said there were three types of responses, from those who: 1) tried to defend their use of the term, telling him to stop being so touchy, 2) defiantly tweeted back the word "abeed" (in one case, repeated as many times as the character limit of a tweet would allow) 3) apologized, saying that they "didn't know *abeed* meant slaves"[52] and that they had grown up hearing their parents or grandparents use the term "without animus."[53] Walid said that he hoped his counterspeech helped people understand why the word is offensive. "I don't think people can be bullied out of being racist. People's moral consciousness has to be revived" he said.[54]

---

[50] Walid, Dawud. "Fellow humans are not 'abeed.'" *The Arab American News*. September 20, 2013. https://www.arabamericannews.com/2013/09/20/Fellow-humans-are-not-abeed/ [https://perma.cc/XKU3-LF7U]
[51] Interview with author, January 12, 2018.
[52] Walid, Dawud. "Responses to my calling out the term 'abeed.'" November 24, 2013. https://dawudwalid.wordpress.com/2013/11/24/responses-to-my-calling-out-the-term-abeed/ [https://perma.cc/P5SE-92XY]
[53] Interview with author. January 12, 2018.
[54] Interview with author. January 12, 2018.

Implications for Researchers

Researchers often set out to study whether counterspeech is "effective," but they seldom consult counterspeakers before defining what they mean by "effective." Going forward, researchers should be responsive to the various goals of counterspeakers when designing studies to test the effectiveness of counterspeech. Evaluating whether counterspeech has had an impact on the beliefs or behavior of the speaker, or whether it has empowered or changed the minds of members of the audience would require different methods. Being sensitive to different pathways through which counterspeech could enact discourse change also opens up new research questions for analysis. Future possible research directions include:

- Examining the link between specific counterspeech strategies and goals
- Studying whether the goals of counterspeakers change over time (and if so, how and why?)
- Determining whether discourse changes in comment threads following counterspeech can be attributed to audience members changing their minds or the activation of new counterspeakers.

As illustrated in this paper, most of the counterspeakers with whom I spoke attempt to reach those in the "silent" audience. Counterspeakers have honed in on this audience in hopes of persuading them on an issue or convincing them to become counterspeakers themselves. Sometimes, they attempt to educate the audience while also providing support for those targeted by hateful speech. Knowing this is useful to researchers who are attempting to evaluate their efforts. It may complicate things as well, as members in the "silent" or "reading" audience are not as visible as those actively taking part in comment threads.

The qualitative data on which this paper is based helps us to zoom in on the motivations behind counterspeech interventions. It provides a clearer understanding of whom counterspeakers are trying to reach and why. Acknowledging their different goals and theories of change in our studies will provide a much more nuanced and complete picture of how counterspeech works – or doesn't – in the wild.