Counterspeech on Twitter: A Field Study

A report for Public Safety Canada under the Kanishka Project by

Susan Benesch[1,4], Derek Ruths[2], Kelly P Dillon[3], Haji Mohammad Saleem[2], and Lucas Wright[4]

[1]Berkman Klein Center for Internet & Society, Harvard University, Massachusetts
[2]School of Computer Science, McGill University, Montreal
[3]School of Communication, The Ohio State University, Ohio
[4]Dangerous Speech Project, Washington D.C.

Contact authors:
Susan Benesch - sbenesch@cyber.law.harvard.edu
Derek Ruths - derek.ruths@mcgill.ca

Table of Contents

Introduction

As hateful and extremist content proliferates online, 'counterspeech' is gaining currency as a means of diminishing it.[1] No wonder: counterspeech doesn't impinge on freedom of expression and can be practiced by almost anyone, requiring neither law nor institutions. The idea that 'more speech' is a remedy for harmful speech has been familiar in liberal democratic thought at least since U.S. Supreme Court Justice Louis Brandeis declared it in 1927.[2] We are still without evidence, however, that counterspeech actually diminishes harmful speech or its effects. This would be very hard to measure offline but is a bit easier online, where speech and responses to it are recorded. In this paper we make a modest start. Specifically we ask: in what forms and circumstances does counterspeech - which we define as a direct response to hateful or dangerous speech - favorably influence discourse and perhaps even behavior?

To our knowledge, this is the first study of Internet users (not a government or organization) counterspeaking spontaneously on a public platform like Twitter. Our findings are qualitative and anecdotal, since reliable quantitative detection of hateful speech or counterspeech is a problem yet to be fully solved due to the wide variations in language employed, although we made progress, as reported in an earlier paper that was part of this project (Saleem, Dillon, Benesch, & Ruths, 2016).

We have identified four categories or "vectors" in each of which counterspeech functions quite differently, as hateful speech also does: one-to-one exchanges, many-to-one, one-to-many, and many-to-many. We also present a set of counterspeech strategies extrapolated from our data, with examples of tweets that illustrate those strategies at work, and suggestions for which ones may be successful.

Review of Existing Research

Here we review previous studies of online counterspeech and also of the speech it is intended to counter. Both fields are still young, so literature is limited. Hate speech and extremist speech online have both been studied by multiple authors.[3] As we and others have noted elsewhere, hate speech is a widely used term but there is no consensus on its definition, in law or the social sciences (Benesch, 2014; Mendel, 2012). In general it means denigrating a person or people based on his/her/their membership in a group, but studies of hate speech, like laws, have employed a variety of definitions. In our own study, we avoided the term

---

[1]For example Facebook's COO Sheryl Sandberg said in a January 2016 speech at the World Economic Forum in Davos, Switzerland, "The best antidote to bad speech is good speech. The best antidote to hate is tolerance. Amplifying … counter-speech to the speech that's perpetrating hate is, we think, by far the best answer." http://www.recode.net/2016/1/21/11588986/want-to-combat-hate-speech-on-facebook-try-a-like-attack-says-coo

[2] Justice Brandeis asserted in his concurring opinion in *Whitney v California* that to expose "falsehood and fallacies" and to "avert the evil," "the remedy is more speech, not enforced silence" (*Whitney v California*, 1927, U.S. Supreme Court, p. 377)

[3] Cyberbullying also has an extensive literature, but is outside the scope of this project.

because of its ambiguity and the difficulty in coding consistently for it, instead using 'hateful speech' and 'dangerous speech,' as explained below.

Hate Speech Research

There is substantial literature describing and analyzing hate speech, online and offline. Within it, we have drawn on recent analytical studies such as *The Content and Context of Hate Speech* (Herz & Molnar, 2012) and Danielle Citron's *Hate Crimes in Cyberspace* (2014), e.g. for ideas on how to draw definitional boundaries. Other works usefully describe the breadth and variety of hate speech online, including *Viral Hate* (Foxman & Wolf, 2013), *Click Here to End Hate* (2014) by the human rights organization Muslim Advocates, and "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age" (Citron & Norton, 2011) .

Another line of research on hate speech consists of efforts to detect it automatically, online. This is much more difficult than it might at first seem, not only because of the definitional problem, but because of the wide variety of ways in which people express hatred and contempt: they do not always use slurs, and do not always use slurs hatefully. One study attempted to code for racism using a Naive Bayes classifier (Kwok & Wang, 2013). This work confirmed the definitional challenge of hate speech by showing annotators could agree only 33% of the time on whether texts contained hate speech. Another considered the problem of detecting anti-Semitic comments in Yahoo news groups using support vector machines (Warner & Hirschberg, 2012). Notably, the training data for this classifier were hand-coded. Manually annotating training data is the most familiar way to produce a dataset for training, but it admits the potential for hard-to-trace bias in the speech ultimately detected. A third study used a linguistic rule-based approach on tweets that had been collected using offensive keywords (Xiang et al., 2012). Like manually annotated data, keyword-based data has significant biasing effects as well.

Dinakar, Reichart, and Lieberman (2011) explored the detection of harassing (not necessarily hate speech) comments on YouTube. They identified sexuality, race & culture, and intelligence as the three major themes in user harassment. From the performance of classification systems that include naive Bayes and SVM, they concluded that label-specific classifiers are more effective than multiclass classifiers at detecting harassment. They further observed that blatant harassment is easier to model than expressions involving sarcasm and euphemism - which often appear in hate speech as well. The authors had to perform manual coding to generate small training and testing datasets.

In another effort at automated coding of hate speech, the Floating Sheep scholars' collective, which specializes in mapping geolocated data, published a map in 2013, identifying the ostensibly most hateful places in the United States with geolocated tweets that used offensive words (Stevens, 2013b). The blog was criticized for extrapolating broadly from a small and arguably skewed sample (tweets are not all geolocated, not all hatred is expressed with slurs, and some forms of hate, such as misogyny, were omitted). Some readers were offended, also,

that the blog published the hateful terms for which they searched. The researchers responded by clarifying their goals and methods (Stevens, 2013a). In another online effort a Canadian NGO, the Sentinel Project, launched a site in 2013 called HateBase, which invites Internet users to add to a list of slurs and insulting words in many languages. HateBase does not, as far as we know, attempt coding or detection.

Although we did not focus on extremist or terrorist speech in this project, we drew on relevant literature such as "Who Matters Online: Measuring Influence, Evaluating Content and Countering Violent Extremism in Online Social Networks" (Berger & Strathearn, 2013), "The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter" (Berger & Morgan, 2015), and "The Islamic State" (Barrett, 2014).

Counterspeech Research

There is a small body of existing research on counterspeech which is summarized in some detail here as it should be a useful contribution; no other review seems to be available. Nearly all of the work so far focuses on countering either hateful speech or extremist speech. Many authors observe, as we do, that counterspeech varies greatly, in tone and in communicative strategies, and several papers offer categories of counterspeech, providing useful frameworks for observation and further study. It is important to note, though, that some authors use the term 'counterspeech' broadly, to refer to any content that counters or contradicts hateful or extremist content generally - not necessarily in response to any particular statement or speech act. We use counterspeech to refer only to a response. For expression that counters another narrative or view in general - a much broader category that could include forms of education, propaganda, and public information - we use the term 'counter narrative.'

The literature thus far contains very limited evidence that counterspeech (or any particular variety of it) is effective in changing thought, online speech, or offline behavior. Nonetheless, some authors offer suggestions on which forms of counterspeech may be most useful for countering hatred and/or extremism.

Among features or techniques of counterspeech that other researchers have identified, the U.S.-based Anti-Defamation League identifies three in its "Best Practices for Responding to Cyberhate" (2014). They are: responding to the original speaker (the user who produced the hateful or dangerous speech), using comedy or satire, and correcting falsehoods. In a study of counterspeech on Facebook, the British think tank Demos distinguishes among constructive counterspeech and non-constructive counter hate (responding hatefully to hatred), fact checking, constructive discussion, satirical oppositional pages, and serious oppositional pages (Bartlett & Krasodomski-Jones, 2015). The Demos report also offers three types of metrics for measuring the success of counterspeech: quantitative metrics, such as user engagement and the spread of counterspeech; content metrics, such as sentiment analysis; and real world metrics that measure the long-term impacts of online counterspeech on offline behavior.

The Demos report also usefully sought to understand whether counterspeech reaches people who do not already agree with it, since if it does not, it cannot be effective at changing minds or discourse norms. Demos found that populist right wing Facebook pages got more likes and comments from people who did not already like those pages, than pages that countered hateful populist right wing views. (Bartlett & Krasodomski-Jones, 2015). This suggests populist right wing content reached a broader audience, and highlights the importance of disseminating counterspeech content to those who do not already agree with it.

Research into the usefulness of counterspeech against hateful speech has also found that the identity of the messenger and the content of the message are crucial. The conflict prevention activist Rachel Brown, in her *Defusing Hate: a Strategic Communication Guide to Counteract Dangerous Speech*, highlights the importance of the medium or platform for determining how and when a counterspeech message will reach the audience. Brown also notes that a particular speaker can make information seem more reliable, and the content of a message can influence the behavior of the audience (Brown, 2016). She recommends a centrally-planned strategy with a clear goal in which medium, speaker, and content are carefully tailored to the target audience.

Research on using speech and narrative to counter violent extremism (CVE) makes many similar points. A report on government anti-extremist programs published by the Institute for Strategic Dialogue (ISD), a London-based think tank that studies violent extremism and methods to diminish it, identifies five possible methods for counterspeakers: to erode the original speaker's intellectual framework, to mock or ridicule, to highlight the negative impact of extremist speech, to demonstrate inconsistencies in the extremist arguments, and to question the effectiveness of extremists at achieving their goals (Briggs & Feve, 2013). The report makes the important point that the goal for all of these possibilities is not to win the argument but to delegitimize extremist narratives. In a white paper on preventing extremism, the Quilliam Foundation, a British counter-extremism think tank, distinguishes counterspeech that tries to discredit extremist content, from counterspeech that offers alternative, more positive narratives, and counterspeech intended to inform and provide transparency around a topic (Saltman & Russell, 2014).

Another common contention in CVE research is that extremism can be more effectively undermined by adding favorable, peaceful messages (such as counterspeech) than by deleting extremist content (Saltman & Russell, 2014; Hussain & Saltman, 2014; Bipartisan, 2012). For example, such research asserts that attempts to remove, monitor, or ban extremist content are stymied by the challenges of defining such content. Also, content is likely to reappear elsewhere on the Internet after it has been deleted (Saltman & Russell, 2014). Counterspeech and counternarratives seem to be more effective since they can better reach audiences that are vulnerable to extremism. Online counterspeech also allows for tracking viewership, and coordinating with offline counter narrative campaigns (Hussain & Saltman, 2014). Finally,

counterspeech can be conducted by civil society or platforms, which are usually more credible, as speakers, than governments.

This leads to another common feature of CVE research that is also found in the anti-hate speech literature -- the importance of the messenger, message, and messaging (tone) in constructing effective counterspeech (Saltman & Russell, 2014). The ISD report on government anti-extremist counternarratives asserts that the identity of the messenger is critical when countering online extremist speech (Briggs & Feve, 2013). According to the authors, to be an effective messenger one must have both credibility in the eyes of the audience and the expertise necessary to market a strong narrative (Briggs & Feve, 2013). Berger and Strathearn found in a study on online extremist narratives that white nationalist extremists were more likely to engage with new people online who seem to be conservative, than with liberals. The authors therefore recommend that interventions to reach white nationalist extremists be directed by conservatives (Berger & Strathearn, 2013). Similarly, a study of one-to-one interventions with individuals considered susceptible to violent extremist speech found that the tone and content of the intervention messages sent to such individuals greatly influenced the response rate (Frenett & Dow, 2015). Response rates to interventions online were highest when the tone of the message was casual rather than antagonistic and the content included offers of assistance and personal stories rather than personal questions or attempts to highlight the possible consequences of extremism.

These previous studies on countering hatred or extremism online focus on two quite different kinds of efforts: those intended to influence individuals, and those that seek to influence (or shift the norms of) groups. In the first category, there are several documented projects to persuade individuals to change their views, especially in cases where the target is considered vulnerable to extremist speech and likely to engage in violence. As Frenett and Dow (2015) demonstrate, such counterspeech is usually delivered by one person (even when acting as part of a larger effort), directly addressing another individual. Counternarrative campaigns are also organized and planned, but they attempt to change the norms or beliefs of a larger audience through changes in narratives. Such campaigns can be orchestrated by governments, civil society, private corporations, or individuals. They usually take place at the one-to-many or many-to-many level.

No previous published study seems to have done what we attempt in the present work: to collect and examine organic, spontaneous counterspeech, which appears when Internet users choose, for a variety of reasons and in a variety of ways, to reply to other users who express hatred or extremist views. Spontaneous efforts are varied but much more numerous, and therefore can have a larger, more sustained impact than programmed interventions and counternarrative campaigns.

Findings

This work was inspired by several examples of spontaneous counterspeech on Twitter that seem to have been successful, in the sense that they were followed by apologies or other

signs of favorable impact on the original account (the account to which the counterspeech responded). We were surprised to come across this, and resolved to discover and study more examples, hoping to learn something about how and why they worked. In each of the original cases (observed among Twitter users in Kenya, France, and the United States)[4], there was a three-step process which we eventually nicknamed a 'golden conversation,' since we came to see such a Twitter exchange as a valuable prize.

The three steps are these: one account produced a tweet that was hateful, racist, or that included a threat to commit violence. One or more accounts - apparently unknown to the original user - replied by rebuking that user, who then recanted or apologized. Such a response was surprising, and belied the familiar maxim 'don't feed the trolls.' If feeding can sometimes be successful, what would be the best rhetorical food?

Counterspeech surges on Twitter, we found, when people who don't know one another and who disagree deeply meet and argue online, often because they are interested in the same offline event. Usually people of very different convictions do not exchange them since, like most Internet users, they spend most of their time in like-minded silos, reading content they agree with and remaining with their own 'tribe' (Conover et al, 2011, Lewandowsky et al., 2012, p. 111, Zuckerman, 2013, Anderson & Raine, 2010, p. 18).

We found both extensive hateful speech and counterspeech in response to certain events that provoked widespread emotional response and debate, such as the Baltimore protests of April 2015, the June 2015 U.S. Supreme Court decision on same-sex marriage, or the November 2015 Paris nightclub massacre. In each of these, new hashtags were formed and people with opposing views 'met' on those hashtags. In some cases, different hashtags expressed opposing views, such as #lovewins and #1m1w or #stopislam and #loveislam. Some people tweeted on the hashtag of an opposing opinion, apparently in order to reach - and engage with - those with opposing opinions, and in some cases, people tweeted using two opposing hashtags.

Some conversations between people of opposing views are far from golden, unfortunately. Not only do they fail at improving discourse online, but counterspeech is sometimes as vitriolic and hostile as the speech to which it responds. Still worse, some respond to hateful or simply offensive speech with threats, dogpiling (many people sending hostile responses to an individual in a short period of time), and harassment. These are not to be confused with counterspeech.

[4] We first observed successful counterspeech on Twitter in Kenya 2013, during a project to study hateful and dangerous speech online during the months leading to a presidential election. *See* iHub Research, Umati Final Report, Sept. 2012 - May 2013 available at http://research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf. Subsequently, we worked with Twitter staff to find other examples of successful counterspeech, including in response to the selection of Nina Davuluri as Miss America 2014, and in response to homophobia on Twitter in France.

Methods

All data in the study are tweets, when possible captured in "conversations," or tweets in which users tagged other users, responding to their tweets. Virtually all of our data are in English, since we realized that coding and analysis must be done by native speakers who are familiar with connotations and cultural context. There are exceptions among our data, such as tweets whose meaning is conveyed almost entirely with images.

Figure 1. Data collection workflow

Twitter is a dynamic platform where users often react to offline events. Twitter allows researchers access to a portion of public tweets through the search API and the streaming API, and we developed user-friendly command-line-tools for those searches.

Both tools require a hashtag or keyword to initiate a session and the search tool returns relevant tweets from the past week, while the streaming tool starts the real-time collection of relevant tweets from that point forward. The output of both tools is a collection of tweets as JSON objects in a text file.

We also have developed additional tools to aid the analysis of the collected data. They include a parsing tool that converts the list of JSONs in the text file to an Excel friendly tabular format and a hashtag tool that lists the most frequent hashtags in the collection of tweets. Together, this canopy of tools allowed us to effectively collect tweets around events as they were happening and present the data in a format that is conducive to our analysis.

We often focused on offline events that led to conversation with racist, sexist, or bigoted overtones, such as terrorist attacks in Europe (e.g. Paris, Brussels), protests against U.S. police (e.g. Ferguson, Baltimore, McKinney, Waco), South Carolina's display of the Confederate flag, and same-sex marriage. Relevant hashtags like #baltimore or #mizzou appeared on Twitter. Multiple hashtags can appear in a single tweet, so each time we began collecting tweets with a hashtag, we also searched for the top 10 hashtags in that set of tweets, often leading us to the relevant sub-conversations.

To illustrate how sub-hashtags gave shape to conversations on Twitter on events, a word cloud generated from top hashtags in #stopislam dataset appears in Figure 2. Tweets containing #stopislam were collected for three days since it appeared as a top hashtag in tweets containing #Brussels, after the terrorist attacks there in March 2016. Users were tagging their tweets with hateful hashtags like #banislam or #muzzlemuzzies, and also with counterspeech hashtags like #loveislam or #muslimsforpeace. Generally, tweets containing competing hashtags were either hateful, "*#StopIslam dg #LoveIslam MAKE PEOPLE GET*

RID OF THE MUSLIM FAITH ETERNAL HELL," or they were counterspeech "*Let's make #loveislam trending then #stopislam. terrorism has no religion!*"

Figure 2. Word cloud illustrating top hashtags in #stopislam dataset, March 2016

Datasets were categorized based on topic, in an effort to include a variety of hateful, dangerous, and counterspeech examples (see Table 1). Within the datasets, we used sort and filter functions to find counterspeech. Searching Twitter's website by usernames, phrases, or hashtags also yielded conversations.

Our search tools capture data embedded in tweets such as the number of followers a user has, whether a tweet is a retweet, and how many times a tweet has been seen at the time of collection. Sorting by retweet status (yes/no) did not yield useful information as it only indicated whether a tweet had been retweeted. Retweets do not necessarily mean agreement with the original sentiment, as the RT function is used to inform, comment on, validate, agree or disagree with, or refer others to, the original tweet (boyd, Golder, & Lotan, 2010). Since retweeters' intent was often hard to determine, this variable was not particularly useful.

If a tweet appeared to be counterspeech, the original conversation was searched for on the user's Twitter feed. Sorting by tweet date was helpful to identify trends on hashtags, or multiple retweets, or even reclaiming of hashtags, as in the case of #KillAllMuslims in the wake of the January 2015 Charlie Hebdo shooting. The largest number of tweets on that hashtag in our data was on the day of the massacre, January 7. After that the number of tweets diminished, and then rose again, this time dominated by counterspeech.

Table 1. Data summary

Topic	**No. of datasets**	**No. of tweets**
Counter phrases (e.g. "I'm sorry")	6	68,036
LGBTQ (e.g. Obergefell v Hodges, U.S. Supreme Court ruling on same sex marriage)	18	55,234
Islamophobia (e.g. Brussels and Paris attacks)	23	91,847
Race (e.g. Baltimore protests)	37	305,273
Refugee/immigration (e.g. Syrian refugees, anti-PEGIDA protests)	8	22,577

Transgender (e.g. NC, Target bathrooms)	6	13,341
Misogyny (e.g. #getbackinthekitchen)	9	4345
Trump (i.e. U.S./Mexico wall, Pope, David Duke)	23	498,932

Coding

A preliminary codebook was developed to better identify hateful, dangerous, and counterspeech. Sixteen variables were included in the preliminary codebook: relevance, trend, direction, tone, hate speech, dangerous speech, counter speech, defensive speech, conciliatory speech, endorsement or reference of stereotypes, endorsement or reference of generalizations, counter speech education, counter speech ramifications, hypocrisy, media, media type, and deleted post.

Some variables were for reference or organizational purposes, like relevance, trend, direction, media, media type (if any), and whether the post was deleted. The directionality of a tweet (retweet, original text, and responsive) helped determine where conversations might be hiding. Coding tweets as hateful, dangerous, defensive, or conciliatory helped organize the data quickly for further probing. The remaining variables in the codebook assisted in shaping the context of the tweets and helped frame the discussion of the strategies observed.

Methodological Challenges

Using tweets as the unit of analysis, it was difficult to determine which part of a conversation a tweet formed, if any. Twitter exchanges are asynchronous, and users can come upon a tweet hours, days, or months after it has been posted. Parts of a conversation may have been deleted, or accounts suspended, or replies could appear out of order as 'side' conversations developed. Counterspeech tweets may not appear sequentially like a scripted or natural conversation, which can make it difficult to reassemble conversations. Also, since we had access to a stream of only a limited sampling of tweets (usually 10 percent), we were often unable to find all parts of a conversation. However, some Twitter conversations on hateful trending hashtags were the subject of news reports online (such as #FuckPhyllis and #KillAllMuslims), which made finding conversations on these hashtags much easier.

Lack of context often made it challenging to identify both hateful and counterspeech tweets by their content alone. For example, the tweet "*@realDonaldTrump we need to start catapulting illegals back to Mexico, They'll stop coming"* might have been posted by a Donald Trump supporter or an opponent sarcastically imitating a supporter. In this case, we checked the account and realized that the tweet was not sarcastic.

It was also difficult to find hateful and dangerous speech transmitted as web links or images, since our tools, which are text based, could not search that content. For example, one tweet collected in response to the Brussels attacks (March 2016) contained only a mention of

another user and a link. The link led to another tweet, reproducing a third one. The image is at Figure 3.

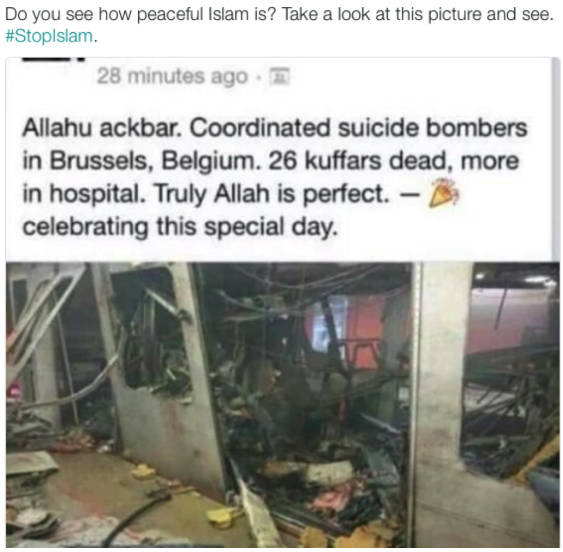

Figure 3. Image embedded in a tweet, illustrating difficulty of text-only analysis

Interrater reliability (IRR), or the extent to which coders agree when classifying the same content, is notoriously difficult to achieve. In social sciences IRR is generally acceptable if the alpha is above .80, or the coders agree at least 80% of the time. To improve our IRR, we coded sets of 10 tweets together. Then, to test a preliminary codebook, 100 random tweets from the #KillAllMuslims dataset were coded by one of the authors and a research assistant, resulting in Cohen's alphas between .42 and .85 (each variable has its own reliability score). Another 33 tweets coded using the same codebook produced a Cohen's Kappa of .37 and .82. It was especially difficult to reach agreement on whether tweets involved generalizations or stereotypes, and whether the overall tone was toxic, hateful, or dangerous. The data could be analyzed at various levels: the conversations (affording context), the individual tweets (with given data), or the components of the tweet (if including visuals, retweets, mentions, or combinations). Since the codebook was preliminary, broad coding and research assistant training did not go further. Future projects should determine (a) level of analysis and (b) a solid codebook or qualitative memoing system (Glaser & Strauss, 1967).

Definitions

Counterspeech is defined in this study as a response that takes issue with hateful, harmful, or extremist content. Counterspeech is considered successful when it is followed by a favorable response from the Internet user or users to whom the counterspeech was directed.

Three clarifying remarks are in order here. First, it is impossible to be sure, in an observational study like this one, that online counterspeech was in fact the inspiration for a favorable response such as an apology. Second, a 'favorable' response may have an unfavorable effect, ultimately, where users feel overwhelmed by angry or abundant

counterspeech and apologize or delete their accounts, simply to get it to stop. Third, counterspeech may have a positive impact on Internet users other than the original speaker, and where that 'audience' is large as it often is on Twitter, the impact of counterspeech on the audience can be much greater and more influential, than on one original account. This suggests a second definition of success: counterspeech that favorably shifts the expression and/or beliefs of Internet users. Past research has focused on counternarrative campaigns as a necessary vehicle for influencing a larger audience, but we believe that norms can be shifted when an audience witnesses public, direct, and organic counterspeech that they would not have been exposed to if it had been published as part of a counternarrative campaign or privately in a personal intervention. Since we did not attempt to measure such an effect, we define success in terms of a favorable response from the original account. That may or may not indicate a durable change in speech or behavior; this is discussed further in the Types of Responses section below.

As noted above, we avoided the term 'hate speech.' Its definitions specify different types of group markers, including (or not), religion, ethnicity, nationality, sexual orientation, disability, body type, age, or gender, among others. Where to draw the line between hate speech and speech that is merely offensive depends so much on prevailing social norms, context, and individual and collective interpretation that it is very difficult to code consistently. A recent study demonstrated a mere 33% agreement between coders from different races, when asked to identify racist tweets (Kwok & Wang, 2013). An additional ambiguity in the term 'hate speech' is in the word 'hate' itself which might refer to the speaker/author's hatred, or his/her desire to make the targets of the speech feel hated, or desire to make others hate the target(s), or the apparent capacity of the speech to increase hatred. Most often, it implies that the speaker feels or intends hatred.

Since it is difficult or impossible to know a speaker's intent, especially from a tweet, we propose the term "hateful speech" to focus on the *expression* of hate. This is a nuanced but useful distinction since expression is easier to detect than intent, and more likely to be linked to language's capacity to cause harm. **Hateful speech,** then, is speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity. We searched for counterspeech in response to this kind of content, and also in response to tweets that constitute '**dangerous speech**,' which we have defined in previous work as speech that can inspire or catalyze intergroup violence (Benesch, 2014). Intensely hateful and even violent speech is not hard to find online, including on Twitter, but most of our successful counterspeech responded, not surprisingly, to tweets that were not among the most vitriolic.

Vectors

We observed important distinctions in counterspeech conversations, according to the number of participants in each stage or side of an exchange. A conversation between two people is qualitatively different, for example, from a case in which many people respond to one tweet.

Not surprisingly, the effectiveness of counterspeech strategies varies with these distinctions as well.

We identified four types of counterspeech exchanges, or vectors, apparently for the first time in the literature: one-to-one, one-to-many, many-to-one, and many-to-many. Literature and policy discussions on hate speech usually conflate these categories also, although the distinctions are equally important in that context: harassment of an individual by a group of people, for example, is very different in nature and likely consequences, from hatred directed by one person against an entire racial or ethnic group. Below, we discuss each of the four vectors in the context of counterspeech.

One-to-One

Some of the most striking examples of "golden conversations," or cases in which counterspeech seems to convince a person to stop speaking hatefully, are conversations between only two people. We found extended exchanges between an original speaker and a counterspeaker who engages repeatedly, even in the face of apparently implacable resistance and a stream of hateful tweets.

Where someone seems firmly committed not only to hateful ideology but to declaring it publicly, we would not expect counterspeech to sway that person. Yet in some cases, it apparently has – and has even helped to bring about lasting change in beliefs, not only in speech. In these cases, we observe counterspeech strategies including: an empathic and/or kind tone, the use of images, and the use of humor. This successful counterspeech usually does not label the original speaker as hateful or racist, but does identify the speech as such.

A successful counterspeech conversation, in which nearly all of these strategies were used, took place on January 19, 2015, Martin Luther King Day in the United States. It began with this tweet:

"*In honor of MLK day today, I'm taking a vow to use the word "nigger" as many times as possible and in the most inappropriate times*"

A writer and activist[5] discovered the tweet and responded with an anti-hatred quote from King:

"'*Let no man pull you so low as to hate him*.' **—** *Martin Luther King Jr."*

The original account responded "*oh so you're one of those nigger lovers too?*" to which the counterspeaker replied, again citing King, "*hate destroys the hater.*"

The original speaker responded with egregious racist attacks on the counterspeaker ("*the only fucking fool is your inferior nigger ass*") who replied with quote after anti-hatred quote from

[5] We've erred on the side of not revealing the identities of people in the cases we describe in order to protect them and to preserve their privacy. We've made exceptions, however, for public figures like Donald Trump and/or those who have already chosen to discuss the case publicly.

Martin Luther King, until she wrote, "*I wish you peace and love and freedom from the hatred that hurts your heart*." Her interlocutor replied, "*who's that a quote from*."

"[t]*hat's me,*" the counterspeaker tweeted. "Sending love and hope to you." From there, the counterspeaker and her interlocutor exchanged more than a dozen tweets, in which she described lynchings and other violence against African-Americans, with photographs asking, "*you mock this?*" The original speaker replied that he was 14 years old. The counterspeaker asked "*does your mom know you spend your time on the Internet trying to hurt people*" – pointing out the harmful consequences that his Tweets could inflict on members of the group he was denigrating. In trying to make him feel accountable to his mother, the counterspeaker seems to have unexpectedly made an emotional connection with the stranger who had been continually attacking her. He replied, "*I doubt it. She's been dead for a year and a half so…*"

"*I'm sorry for your loss,*" the counterspeaker wrote. "*And I hope you find a better way to honor her.*" After several more exchanges, he wrote to the woman he had been attacking viciously, "*you're so nice and I'm so sorry*." This case was publicized online and the counterspeaker was praised for her forbearance (Payne, 2015).

Another striking example of one-to-one counterspeech is the case of Megan Phelps-Roper, who was fully convinced of the extreme homophobic tenets of the Westboro Baptist Church, which her grandfather Fred Phelps founded and in which she was raised - until she started a Twitter account to spread the views of the church. On Twitter she encountered people who challenged her views and engaged her in other ways, including humor and suggestions for music she might enjoy. Extended conversations with two of them (David Abitbol, a rabbi who founded the blog Jewlicious, and a lawyer with whom Phelps-Roper gradually fell in love) completely changed Phelps-Roper's views, by her own account. She ended up leaving the church. This case is recounted by Adrian Chen (2015) in a long article that Phelps-Roper has praised for capturing her transformation very well.

A key distinction between these one--to-one cases and the other exchanges we studied is that these dialogues are long, going far beyond the three essential parts of what we affectionately call a golden conversation. It is no surprise that deep and/or lasting change in discourse and beliefs - difficult to achieve by any means, online or offline - can take many tweets. Another distinguishing feature of one-to-one conversations is that, even on Twitter, they are not always public. One can send a "direct message" on Twitter which, like an SMS or text message, is visible only to its designated recipient. In Megan Phelps-Roper's case, she and her new interlocutors also used one-to-one messaging apps others than Twitter, such as Words With Friends. In a less public online context, people may feel less guarded and therefore more open to dissenting views. On the other hand, if their conversations are invisible to the larger 'audience,' the audience can neither join in nor be favorably influenced by the conversation, except in rare cases when it is described elsewhere, as in Chen's article (2015).

One-to-Many

Some Twitter users have taken it upon themselves to try to change way in which others express themselves publicly on Twitter, by searching for the use of certain terms or phrases and rebuking those who use them. This sort of activist effort is one-to-many counterspeech.

In one example, Dawud Walid, an African-American Muslim, searched for variations of the word 'abeed' which means 'slave' in Arabic, where it was used in tweets to refer to black people. He sent an op-ed he had written, entitled "Fellow humans are not abeed" to many Twitter users who had tweeted the term. He received a variety of responses, from apologies and promises not to use the word again, to a tweet that repeated the word as many times as possible in 140 characters (Walid, 2013). Other similar efforts are @YesYoureRacist and @YesYoureSexist. In each of these cases, counterspeech is met with a range of responses, from apologies to angry argument. In another example of one-to-many counterspeech, some users deliberately tweet on a hashtag with which they disagree, such as #stopislam, to reach people who agree with it.

Many-to-One

In some cases, news of an objectionable tweet (or hashtag) goes viral, and many - sometimes thousands of - Twitter users join in counterspeech. This can be salutary where it catches enough of the attention of the original speaker to be successful but not harrassing, as in the case of a user who tweeted his outrage that Nina Davuluri (who he erroneously identified as an Arab) had been chosen as Miss America 2014. After receiving tweets that variously corrected his error and called him a racist, he first responded "*I didn't realize it would explode like that #unreal*" and then tweeted at Davuluri, apologizing. The furor died down quickly, and the user is still on Twitter, wishing happy birthdays to family members and tweeting photos of his fishing trips.

In other cases, however, huge numbers of angry Twitter users have overwhelmed others, rising to the level of harassment. Original speakers hastily delete tweets or even their accounts, but even that can be an insufficient refuge in the face of, for example, counterspeakers who contact their employers, demanding that they be fired for having posted perceived hateful or racist content. This was the case for Justine Sacco, who inadvertently created a number one worldwide trend on Twitter in December 2013 by tweeting (tongue-in-cheek, she insisted later), "*Going to Africa. Hope I don't get AIDS. Just Kidding. I'm white!"* as she boarded a flight to South Africa. Tens of thousands of furious tweets came within hours, and Sacco was soon fired from her job in public relations (Ronson, 2015).

The blog "Racists Getting Fired" made a practice of punishing people who made racist posts, by contacting their employers and, similarly, demanding that they be fired (McDonald, 2014). Such responses are no doubt successful at changing the online speech of their targets, but may only harden the hateful convictions of those targets, and constitute online mob justice.

Many-to-Many

Conversations among large numbers of people online are of interest, not least because of the impressive scale on which they often take place. In our experience, they are often catalyzed by offline events that are of strong interest to a large number and large variety of people. On Twitter, such conversations generally form around hashtags.

Hashtags can constitute hateful and dangerous speech, or counterspeech, and they often gather or inspire 'many-to-many' conversations. The use of "a hashtag can be seen as an explicit attempt to address an imagined community of users… as each user participating in a hashtag conversation acts potentially as a bridge between the hashtag community and members of their own follower network" (Bruns & Burgess, 2012, p. 804). Often, one hashtag represents one general view or normative group, such as #BlackLivesMatter, and others represent opposing or dissenting views, such as #BlueLivesMatter (which refers to the police for their blue uniforms), or #AllLivesMatter.

One of the most vitriolic hashtags we found, #KillAllMuslims, trended in the immediate aftermath of the Charlie Hebdo massacre of January 2015 - and then was quickly taken over by counterspeakers expressing their dismay that it existed. One counterspeech tweet that uses the hashtag was retweeted more 10,000 times: "*Not muslim but never thought about this b4 #CharlieHebdo #KillAllMuslims #Muslims pic.twitter.com/LL1pkPk6uk.*" The link was to an image of visual similarities among religious traditions, e.g. a Catholic nun in a habit and a Muslim woman in hijab.

Notably, hashtags can be more widely and quickly disseminated, when they trend, than any tweet. When #KillAllMuslims trended, for example, thousands of people on Twitter could not help but notice two things: the hashtag called for mass murder or genocide, and thousands of people had typed it and sent it, as part of their tweets. The fact that a hashtag is trending can also have a major impact on how Twitter users perceive norms on the platform. It is dismaying when hateful hashtags trend, and reassuring when counterspeech does. The hashtag #YouAintNoMuslimBruv, for example, trended after a bystander yelled the same phrase in December 2015, at a would-be attacker in London. A worthy topic for further study would be the norm-influencing capacity of hashtags around public events and controversies.

Taxonomy of counterspeech

We have observed numerous communicative strategies at work in counterspeech tweets, which we present and discuss here. Some are well described in persuasion and communication literature. We distinguish eight strategies: 1) presentation of facts to correct misstatements or misperceptions, 2) pointing out hypocrisy or contradictions, 3) warning of possible offline and online consequences of speech, 4) identification with original speaker or target group, 5) denouncing speech as hateful or dangerous, 6) use of visual media, 7) use of humor, and 8) use of a particular tone, e.g. an empathetic one. These are not exclusive;

counterspeakers often employ multiple strategies in a single tweet. Nor is our list comprehensive; it is intended as a start, based on data collected for the present project.

Presenting facts to correct misstatements or misperceptions

Like previous authors (Bartlett & Krasodomski-Jones, 2015; Anti-Defamation League, 2014; Briggs & Feve, 2013), we observed that counterspeakers often try to persuade by correcting misstatements. Unfortunately there is abundant research indicating that this rarely succeeds, especially in the first instance.[6] Even when presented with facts that correct an egregious mistake or falsehood, most people are unlikely to change their beliefs about the misperception as "false beliefs based on misinformation, are often held strongly and with (perhaps infectious) conviction" (Lewandowsky et al., 2012, p. 108). Paradoxically, individuals with minimal knowledge of facts and experience of the subject are least likely to be swayed by corrections (Kuklinski et al., 2000). Some researchers have even found a "backfire effect" where correction attempts result in firmer beliefs in the misperception or misinformation (Nyhan & Reifler, 2010; Nyhan & Reifler, 2015). Whereas the corrections may not be successful in changing the speech or attitudes of the original speaker, who may have a tightly held belief on the topic, they could persuade cyberbystanders i.e. members of the general audience who have not yet allied themselves firmly with either side.

Perhaps aware of how difficult it is to change minds even with facts, counterspeakers sometimes support their contentions with evidence or even invite others to look up the evidence for themselves. For example, one user who was against Target's new policy of gender-neutral bathrooms, said over 400,000 people had signed a petition to boycott Target. Another user tried to correct the number (<5,000), and invited the first user to check. The original speaker insultingly declined (see Figure 4).

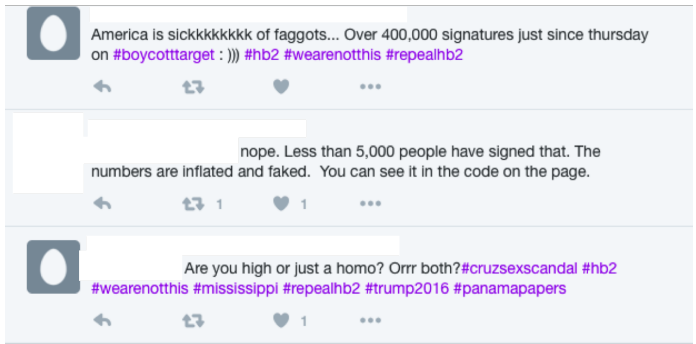

Figure 4. Counterspeaker using facts to correct misinformation and misperception

Counterspeakers sometimes go to striking lengths to persuade strangers that their facts or understanding are wrong. For example, in response to the suggestion that the image of the abolitionist Harriet Tubman be placed on an EBT (welfare) card instead of the US $20 bill, a

[6] Psychologists and other scholars use the term 'motivated reasoning' to describe people's extraordinary (and routine) efforts to reach the conclusions they seek, in spite of facts that do not support those conclusions. See, e.g. Dan Kahan, Neutral Principles, Motivated Congnition, and some Problems for Constitutional Law. Harvard Law Review, 2011.

counterspeaker engaged the original speaker in a debate about the proportions of white and black Americans on welfare (see Figure 5). The counterspeaker found data from a separate, more reputable source (the U.S. Department of Agriculture), and then posted a graphic to illustrate the data. The effort, like most of the others we observed, was unsuccessful.

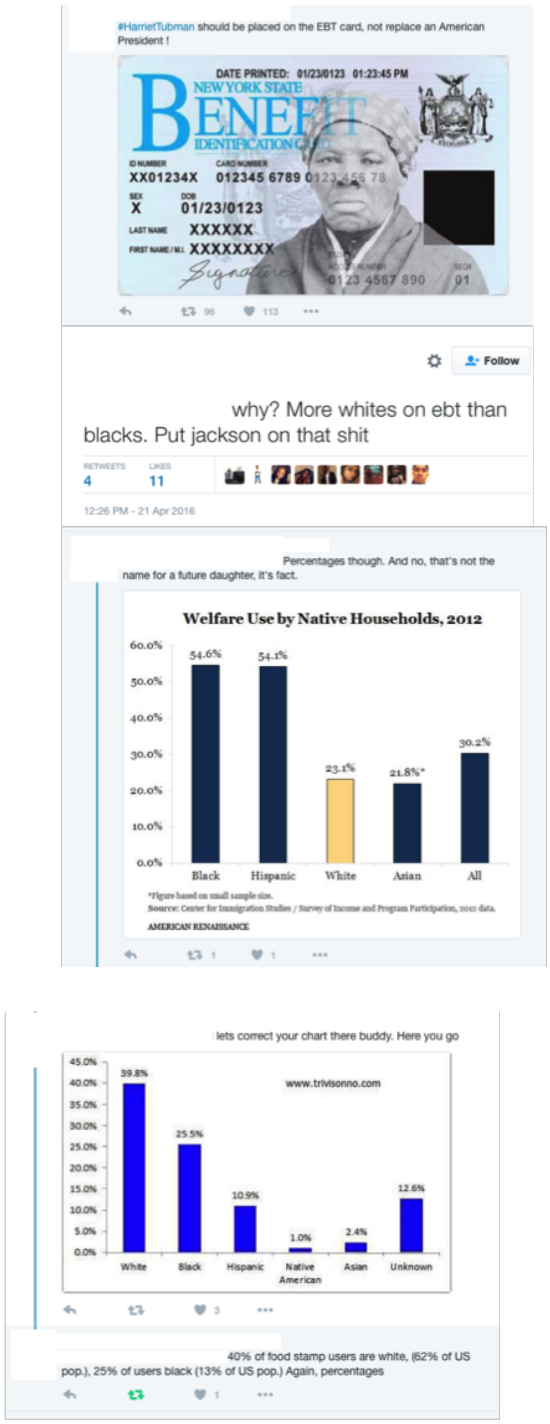

Figure 5. Counterspeaker marshalling statistics

In a discussion of Islamic prayer rooms in American public schools, one speaker tried to use historical events to correct the original speaker's assertions, "*Point of historical fact Spanish soldiers murdered Pueblo Indian religious leaders because they were competition."* One of

those tagged responded to this counterspeech, and a polite exchange continued but the fact did not apparently change any minds on the topic.

Pointing out hypocrisy or contradictions

In a strategy similar to correcting mistakes or misstatements, counterspeakers often pointed out hypocrisy or contradictions. This, too, may cause an original speaker to dig in his or her heels, since it is disturbing to be called a hypocrite. In order to discredit the accusation, the individual may explain away and rationalize their previous behavior, or if they are persuadable, resolve to avoid the dissonant behavior in the future (Beauvois, Joule, & Brunetti, 1993). But to members of the audience, whom we can call cyberbystanders, pointing out hypocrisy may be persuasive.

In one example, a user tweeted "'*#KillAllMuslims!!!!' 'Was hitler a bad person?' 'Yes how dare he kill someone just cos of their religion' Ummm... Ok*." Another example of calling out hypocrisy or contradictions is to paint the original speaker with the same brush he or she is using to paint a group. For example, one counter speaker tweeted "*If a muslim twtd #KillAllChristians he must b a "terrorist" bt a christian twtng #KillAllMuslims is simply exprsng his freedom" while another simply stated, "#KillAllMuslims " .. but yet they call us terrorists.*"

Warning of offline or online consequences

Warning users of the possible consequences of their hateful or dangerous speech is a common strategy among counterspeakers. Like the previous strategy, this may lead an original speaker to hold more tightly to opinions. However, in our data, this counterspeech strategy appeared to be successful, leading to recanting or deletion of offending tweets. This may be because people forget, until reminded, that their speech online can have serious consequences offline. Online, they can become disinhibited, acting as if they "live in an make-believe dimension, separate and apart from the demands and responsibilities of the real world" (Suler, 2004, p. 323). It may also be that warnings or threats are effective because they have been carried out in well-publicized cases, such as successful demands that people be fired from their jobs for content that they posted online.

In one such case, a Texas schoolteacher[7] posted a racist diatribe on Facebook in the wake of a scandal over police brutality in McKinney, Texas (Michael & Young, 2015). Screenshots of her diatribe made their way onto Twitter, and were circulated widely. In response, many users tagged her employer, Frenship Independent School District, on Twitter calling for her ouster, "*@FrenshipISD She has got to GO!! Get rid of her NOW. #[teacher's name]," "@FrenshipISD If #[teacher's name] has any black kids in her 4th grade class PLEASE get them out of there! #Mckinney," "@FrenshipISD is the racist #[teacher's name] fired yet or is*

[7] We avoid repeating the name of the teacher since she has not chosen to discuss the case publicly, nor is she a public figure.

this type of ideology encouraged in your district?" The teacher was eventually fired for failing to abide by the staff code of conduct on social media.

In many other cases, counterspeech features warnings of the risk of future unemployment. Responding to misogynist and racist tweets regarding University of Illinois at Urbana–Champaign Chancellor Phyllis Wise's decision not to cancel classes on a very cold day in January 2014, counterspeakers were quick to warn of real world consequences (Simeone, 2014). A recurring theme was the implications on future employment. Some tweeted, "*All of you students who tweeted racist comments, GOOD LUCK GETTING A JOB DUMMIES #fuckphyllis*" and "*Ppl defending #FuckPhyllis focusing on wrong thing. You should be worrying ab how to undo your tweets come your eventual employment search*." Another reminded students that their professors and future employers might see their tweets (see Figure 6).

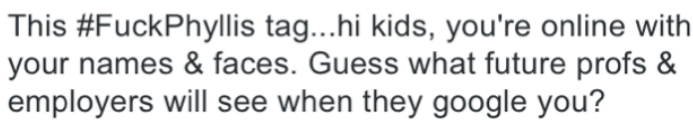

Figure 6. Image of tweet warning users of offline consequences

One user did report attempting to punish students by contacting their employers, tweeting "*11 students reported to Wise. 1 reported to employer. see if they'll be laughing abt #fuckphyllis Monday #diversity Madness*" and later "*lmfao...taking screencaps and names of these racist kids. Found the employer for one....still working*." Another user tweeted, *"As an employer who has recruited many fresh out of college students, I do have second thought to hire anyone out of UIUC seeing #fuckphyllis."* Some alumni even complained that the hateful tweeting would put a burden on them: "*Police yourselves, U of I students; burden of #fuckphyllis also falls on the shoulders of us alumni who try to get u hired when u graduate.*"

Online consequences of hateful speech are blocking and ostracism. Many counterspeakers documented their efforts to limit instances of hateful speech in their personal feeds as warnings to the general audience: *"The situation in #Baltimore brought out the #racists. My friend's list is down a dozen so far blocked and unfriend-ed," "Calling the #protesters #animals and defending #Baltimore for killing #FreddieGray you're gonna get blocked! #AmericaIsBetterThanThat," "Anyone using the hashtag #KillAllMuslims on my TL get blocked," "Who Ever Came up with this #KillAllMuslims he/She should be blocked from all social network."*

Counterspeakers sometimes also try to provoke empathy by pointing out to the original speaker that hateful or dangerous speech may have painful consequences for others online. For example, the writer who repeatedly replied to another Twitter user's racist tweets on Martin Luther King Day 2015, as described above, tweeted at him, "*remember that there is a good chance that someone coming across your hateful tweets is suffering from depression too.*" In other cases, counterspeakers warn of harm to the groups denigrated in hateful tweets,

"*If UofIllinois students can say #FuckPhyllis and be racist/sexist to their Chancellor - then I fear for other Asians and/or women on campus.*" However, warning original speakers of the offline consequences that they themselves might face appears to be the most effective type of this strategy.

Unfortunately, warnings can easily swell into threats and harassment, sometimes by large numbers of people. Counterspeakers can also target the wrong individuals, such as people who have written something offensive or insensitive, but not hateful or dangerous - or even people who simply happen to share the name of the intended recipient. No matter what content they are countering, would-be counterspeakers should be aware of the line between counterspeech, on the one hand, and trolling, harassment, dogpiling, or even threats of physical harm on the other. It is both wrong and ineffective to attack in those ways, and it must not be confused with counterspeech.

Affiliation

Affiliation is "...establishing, maintaining, or restoring a positive affective relationship with another person or group of persons" (Atkinson, Heyns, & Veroff, 1954, p. 407). As social creatures, we understand we are "a member of numerous social groups and the membership contributes, positively or negatively, to the image" one has of him or herself (Tajfel, 1974, p. 69). People may also more likely to credit the counterspeech of those with whom they affiliate, since they tend to "evaluate ingroup members as more trustworthy, honest, loyal, cooperative, and valuable to the group than outgroup members." (Kane, Argote, & Levine, 2005, p. 58).

If a counterspeaker personally affiliates with a hateful speaker, it can reduce the perceived distance between the two (Tanis & Postmes, 2003). Counterspeakers, using implied social identities either in the text of hateful or dangerous tweets or from the hateful speaker's profile information, can use the information gleaned to affiliate with the original speaker's assumed or perceived social identities. In some cases, counterspeakers use a shared identity to claim that certain speech is unacceptable for a member of that group. In the tweet below, the speaker uses his identity as a "right winger" to both associate with, and distance himself from, the original speakers (Figure 7). As noted above, Berger and Strathearn's research indeed found that conservatives are better suited to reaching out to white nationalists online than liberals, because of a perceived affiliation (2013).

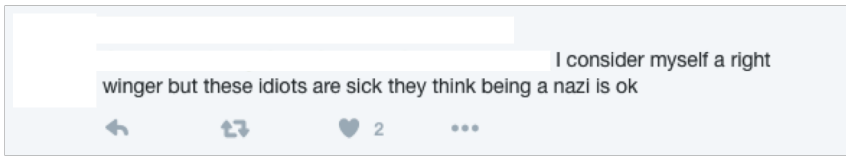

Figure 7. Illustration of counterspeaker affiliating with original speaker.

In the #FuckPhyllis data, alumni asserted their affliation with the students who started the trend to lend credibility to their counterspeech, and to declare the students' behavior

out-of-bounds for their shared identity. For example: "*As a UofI alumna, I am completely embarrassed by this #FuckPhyllis hashtag. Reality check kids: professors & employers can see everything.*" This affiliation tactic was also successful in recruiting other counterspeakers. It can facilitate helping or intervening behaviors, most likely due to group norms, such as expectations to help one another or behave like one another (Rutkowski, Gruder, & Romer, 1983).

In another use of affiliation, users from a group vilified in tweets sometimes disclose their affiliation in efforts to counter those tweets. For example, in response to #KillAllMuslims, one tweeted "*As a Muslim I'm disgusted and offended that this is even a hashtag. #KillAllMuslims.*" Identifying as Muslim in the face of online Islamophobia was found in other events. One user, in response to the Garland, TX shooting tweeted, "*As a Muslim living in Texas the tweets under the #garlandshooting hashtag are both alarming and disturbing. Praying for peace & tolerance.*" Others humanized a dangerous proposal in the hashtag #KillAllMuslims, "*#KillAllMuslims i'm a muslim and i don't think i should be killed bc someone did something*," "*Can't believe that this is trending, I'm a Muslim myself so you want me dead? #KillAllMuslims*." The effects of racist speech on a target group were spelled out by self-disclosing target group members, "*As an Asian American woman, trends like #FuckPhyllis make it clear that no amount of success can erase racist and sexist attacks."*

Some take it upon themselves to apologize on behalf of identities they presumed they had in common with hateful speakers. For example, one user tweeted, "*As a christian, I'm sorry for the hashtag - #KillAllMuslims. It's a shame that humanity is gone from the modern world.*" Others, having identities in common with both a target group and the original speakers of the hateful speech, speak out against the hateful speech generally (see Figure 8).

Figure 8. Image of tweets affiliating with target and hateful speaking group.

Denouncing hateful or dangerous speech

Counterspeakers often identify speech as hateful, especially racist or misogynist. As one user tweeted, "*I know you can't always change a racist's mind and tbh i dont want to bother wasting my time on them but ill be fucked if i dont at least humiliate and shame them."*

Denouncing the content and not the speaker can be useful, and even educational if the speaker is unaware they are engaging in hateful or dangerous speech (see Figure 9).

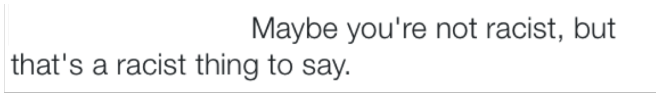

Figure 9. Counterspeaker distinguishing speech from speaker

Some counterspeakers simply call out tweets as racist (see Figure 10). Others elaborate on why speech is hateful or dangerous: "*YES. IT DOES. The entire tag is misogynistic and encourages rape. #fuckphyllis.*"

Figure 10. Counterspeaker's denunciation, quoting an original speaker's tweet.

Where denunciation is directed at a person, not at content, it can be indistinguishable from hostile name-calling, as in the exchange in Figure 11. One Twitter user makes a critical comment about atheists (and atheism) and another responds with an insult: "piss off, bigot." A second user instead explains why the first tweet might be understood as bigoted. After that, the original speaker says her tweet was "snotty" and clarifies her point.

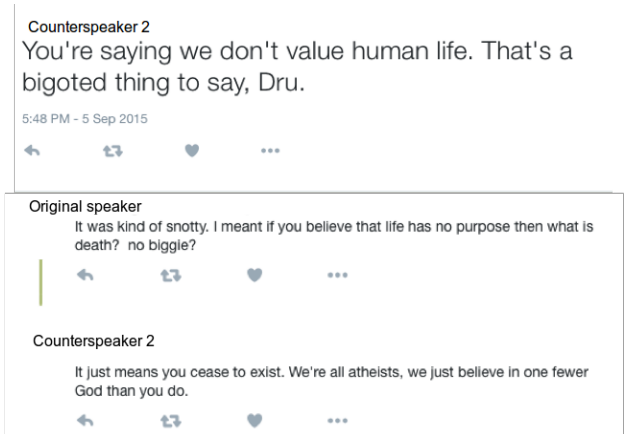

Figure 11. Two counterspeakers; different strategies

Counterspeakers often denounce hashtags. For example, one user tweeted asking rhetorically, "*why is #FeminismIsAwful a thing? It supports equality. Why are people against that*?" Many took to Twitter to denounce the hashtag #KillAllMuslims, until tweets condemning it outnumbered those supporting it. Examples are "*#KillAllMuslims is a vicious and cruel trend and it just needs to stop*" and "*#KillAllMuslims is literally the most disgraceful thing I've seen on Twitter*" and one of the most often retweeted posts in the wake of the Charlie Hebdo shootings in January 2015 was "*To those who trended #KillAllMuslims You're no different than Al Qaeda, Osama & ISIS You are humanity at its worst!*"

Visual communication

Twitter allows users to embed images (i.e. memes, graphics, photographs, animated gifs, videos) in tweets, and counterspeakers often use them, since images are more persuasive than text alone. They "send people along emotive pathways where textual/verbal material leaves them in a more rational, logical and linear pathway of thought" (Joffe, 2008, p. 84). Online specifically, visuals can "augment textual information via paratasis, that is, by being placed next to such information as a coordinate, supportive structure" (LaGrandeur, 2003, p. 124). In one example, a heavily retweeted counterspeech post contained the text "*the fact that #KillAllMuslims is/was trending makes me sick so here's a friendly reminder*" with an image illustrating the proportion of terrorists among all Muslims (see Figure 12).

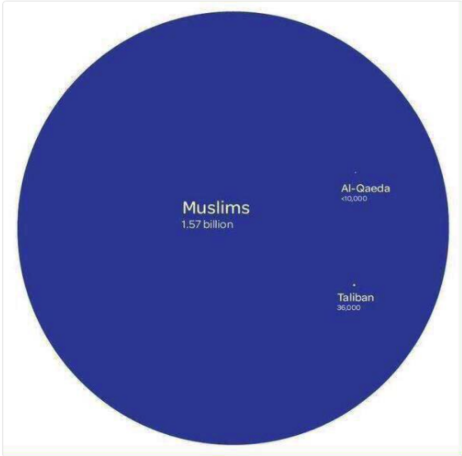

Figure 12. Visual communication example.

Others try to put current hateful speech into historical context (see Figure 13).

Figure 13. Visual comparison of Islamophobia, Nazis, and KKK.

Visual caricatures can make the original seem ludicrous, and may limit the effects of hateful or dangerous speech. For example, users have tried to undermine perceptions of the strength and power of ISIS by altering ISIS photos, which abound on Twitter, with rubber-duck heads and accoutrements (see Figure 14).

Figure 14. Spoofing ISIS.

Users also embed animated visuals like gifs and videos in counterspeech. For example, many tweeted ESPN's #MoreThanMean video of men reading hateful and dangerous tweets sent to female sports reporters with a comment "*We should be respectful of others in person or online*" (see Figure 15).

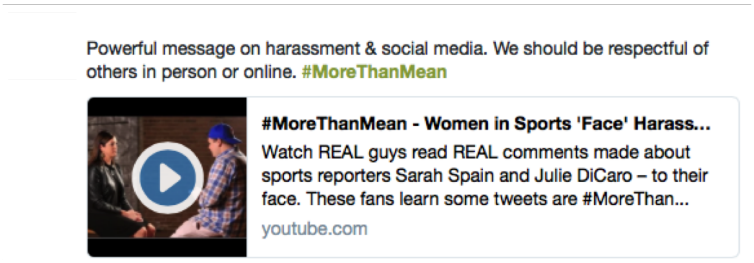

Figure 15. Video counterspeech.

Humor

Humor seems to have special power as a counterspeech strategy. It may shift the dynamics of communication, de-escalate conflict, and draw much more attention to a message than it would otherwise garner. Some perceive humor as a linguistic or communicative performance, which "puts the act of speaking on display – objectifies it, lifts it to a degree from its interactional setting and opens it to scrutiny by an audience" (Baumann & Briggs, 1990, p. 73). Researchers have found that humor in online settings eases hostility, offers support to other online speakers, and encourages social cohesion. (Marone, 2015, p. 61). On Twitter, humor often gets the attention of a very large and diverse audience, as illustrated below, and is a relatively common strategy. It comes in many forms, of course, including caricature and sarcasm, and can vary immensely in tone, from conciliatory to provocative or even aggressive.

Humor is often successfully conveyed in images, so it allows people who do not share a language to counterspeak together - often in large numbers and across cultural and national boundaries. One example of this came in July 2014, after Turkish Deputy Prime Minister Bülent Arinc said (giving his ideas for proper female conduct) that women should not laugh loudly in public. With hours of the speech, #direnkahkaha (Turkish for 'resist and laugh') was one of the top ten trending hashtags worldwide (Bacchi, 2014). Inspired, outraged, and amused, tens of thousands of women - and men - posted laughing selfies. In many cases, the images were also funny, as they showed dogs, cats, goats, and horses apparently laughing.

Another amusing image spread like wildfire on Twitter in April 2014, after Dani Alves, a Brazilian soccer player, was subjected to an all-too-familiar humiliating racist gesture when a spectator threw a banana at him on the field. Alves picked up the banana, peeled it and ate it. Another player, Neymar da Silva Santos Jr., quickly posted an image of himself eating a banana, on the hashtag #Somostodosmacacos, meaning "we are all monkeys." That hashtag, in English, spread quickly as well. In this case, Neymar had prepared the image in advance with the advice of an advertising firm, since both he and Alves had been pelted with bananas many times before - and thousands of people posted banana-eating selfies in spontaneous support. See Figure 16.

Figure 16. Tweet using a hashtag in two languages, to counter racism.

In some cases, humor is used to soften a message that would otherwise be blunt, and perhaps to make it more persuasive. A popular meme, the "It's Time To Stop Posting cat" was used by one user to protest Sinn Féin president Gerry Adams' use of the 'n-word'. (see Figure 17). In response to that critique and others, Adams apologized and deleted his tweet quickly, but said he had been innocently comparing the suffering of African-Americans with that of Irish nationalists. "*While there are parallels between people in struggle, the tweet was inappropriate*" he tweeted.

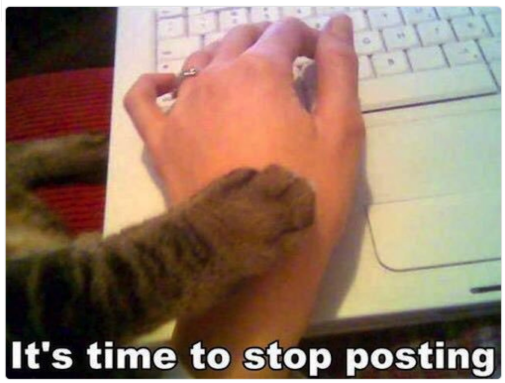

Figure 17. Cat meme used to soften a demand

Using caricatures, by contrast, is a way to counter hateful or dangerous speech by mocking it. For example, several counterspeakers took on a user who tweeted "*racist is a derogatory word used only to describe white people who oppose their genocide. Anti-racist is code word for anti-white. Diversity is a codeword for white genocide.*" One user replied with a caricature of the NFL team Washington Redskins' logo, using white stereotypes (e.g. striped ties, golf clubs, blond hair, lantern jaw) to mock the original user (see Figure 18).

Figure 18. Image of caricature of stereotypes by counterspeaker

Humorous counterspeech has been used even in response to violent extremist content on Twitter. For example, the ISIS rubber duck pictures in Figure 14 are meant to make intimidating images and messages less frightening or powerful. Similarly, when the human rights activist Iyad El-Baghdadi retweeted part of a call to action from ISIS in 2015, Muslim Twitter users responded with humorous reasons why they could not join (see Figure 19), referring to popular culture and to the Quran. Like the ISIS rubber duck images, the humor in these counterspeech tweets can weaken the ISIS content by trivializing it.

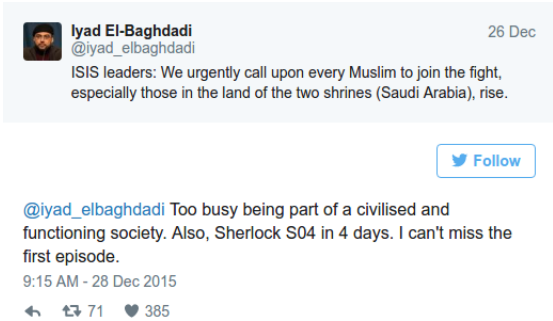

Figure 19. Humorous response to ISIS recruitment attempts

Often Twitter humor is also sardonic or sarcastic, like this tweet: *"Hey let's trend #KillAllMuslims but still call them terrorist because it makes so much sense!!!!"* Another counter speaker commented on Donald Trump's proposed ban on admitting Muslims to the United States, *"@realDonaldTrump congratulations! You're the new Hitler! #MuslimBan*."

Users, as well as researchers (Maynard & Greenwood, 2014), have difficulty identifying sarcasm reliably and sometimes give the benefit of the doubt, *"Can't tell if you're being sarcastic or not. Do you really not know that black face is offensive and a racist act?"* On their own use of sarcasm, some counterspeakers leave no room for doubt, tagging their tweets with a hashtag or some form of modified code like /sarcasm (Liebrecht, Kunneman, & van den Bosch, 2013). "*I've never had a snow day in my life, even with temps at -46 with windchill, so I'm really feeling sorry for you all #sarcasm #fuckphyllis,"* or by simply

announcing it "*yes thats right your hashtags aren't racist at all. Neither are the photos you have posted. Just to make sure, i'm being sarcastic.*"

Tone

Twitter counterspeech occurs along a very wide tonal and emotional spectrum, from tweets that are just as obscene, angry, and vicious as the tweets to which they respond, to kind and empathic messages. Tweets of different tone have, of course, different effects. Hostile tweets can persuade an original speaker to delete tweets or even a whole account, but are unlikely to either de-escalate the conversation or persuade the original speaker to recant or apologize. For example, one counterspeaker tweeted "*fuck not all muslims belong to ISIS*" in response to another who used the hashtag #KillAllMuslims. Elsewhere on the spectrum are many varieties of tone: empathic, kind, and polite, or civil, speech.

In general, expressions of empathy have been shown to lead to positive attitudes towards an adversary when trust between speakers is high, but negative effects if trust is low (Nadler & Liviatan, 2006). It has also been seen as a speech accommodation tactic, leading to convergence where speakers begin to adopt each other's speaking styles (Street & Giles, 1982). Empathy can help a speech partner save face in disclosure, helping close the interpersonal distance between two speakers to engender trust and credibility of arguments.

Counterspeakers used empathic tones in their tweets against hateful hashtags as well as towards individual users who were tweeting hateful content. The hashtag #KillAllMuslims yielded much of this, including suggestions for new hashtags to trend. For example, one user tweeted "*#KillAllMuslims #KillAllChristians What atrocious hashtags. #LetLoveLive*" and "*#KillAllMuslims How about #SaveHumanity ?? :")*" Another user was quite polite in denouncing the hashtag: "*#KillAllMuslims I strongly disagree with this tag. It is disgusting I believe in #Love & #peace it's a real shame #ParisShooting."* Even when faced with explicit death wishes, some responded with empathy (see Figure 20).

Figure 20. Empathic response to hateful speech.

In an extended one-on-one exchange with a user intent on tweeting racism on Martin Luther King Day 2015, the counterspeaker repeatedly countered with empathy, "*still wishing you love*," "*try as you might, I'm never going to wish you anything but love*," and "*I would never*

laugh at your murder. I would never taunt your grieving. I would never mock your fight for equality." Her empathetic tweets apparently succeeded; for a detailed description of this case, see page 14 above.

Types of Responses

We have identified five counterspeech response types, which may serve as an important starting point for understanding when counterspeech is successful, with regard to both the author of a hateful tweet and its audience. These responses are: an apology or recanting, the deletion of the tweet or account, the creation of more hateful speech, a sustained and civil conversation, and eliciting new counterspeech from the audience.

Apology or recanting

The response to counterspeech which most clearly suggests success is when the original speaker recants or apologizes. We found such responses to be rare, but not unheard of. We observed them when original speakers claimed to be unaware that their tweets were hateful, and when they were rebuked repeatedly, or by several different accounts. For example, in the Miss America incident, a user at first denied that he was racist, and apologized only after several counterspeakers contradicted him. This suggests that continued counterspeech as original speakers slowly retreat can eventually lead to an apology. Without interviewing those who recant or apologize, however, we cannot know why they have done so, or whether their speech remains less hateful, or whether their views have changed. Sometimes, original speakers may back down or even apologize simply to get counterspeakers to leave them alone, and may still speak hatefully in the future.

Deletion

Another common response to counterspeech is that the original speaker deletes the hateful tweet, or the entire account from which it was sent. We observed this as a response to many-to-one counterspeech and in response to warnings of offline consequences. As mentioned above, many counterspeakers responded to the #FuckPhyllis hashtag by pointing out to students that future employers would be able to read those tweets. In response, many students deleted their hateful tweets.

Measuring the success of this type of response requires a careful consideration of its long term and short term effects. While pushing the original speaker to remove hateful content is a short term success, it does not ensure that the speaker will communicate less hatefully in the future. Even if an account is deleted, the speaker may continue in the same vein on another platform or on a different account on the same platform. The outcome that most indicates long term success is a meaningful change in the original speaker's future speech, but this may also constitute silencing, a tactic that is often used to curb the speech of women, minority groups, and dissenters. This highlights one reason for our strict definitions of both hateful speech and counterspeech: we believe it is essential not to silence controversial views.

More hateful speech

Supporters of the slogan "Don't feed the trolls" should be unsurprised to find that one response to counterspeech that we observed is more hateful speech, not only from the original speaker, but sometimes also from others on Twitter who join in to refute the counterspeaker. We observed this response coming from both those who enjoy eliciting a reaction, those who seem to sincerely hold hateful beliefs, and some who are both, e.g. racist trolls. It is a reaction similar to that of those with strongly held beliefs who respond to fact-checking by becoming even more entrenched in their views. This sort of response obviously increases hateful speech online, but that may, in turn, inspire more counterspeakers to step in.

Civil conversation

We have also observed repeated exchanges between those who disagree strongly but remain civil. Success is uncertain, since the original speaker often refuses to acknowledge any wrongdoing, but diffusing a situation so that hateful speech is no longer being produced is at least a partial success. The contact hypothesis suggests that prolonged civil exchanges between individuals who share different views and normally would not interact can help reduce distance between the two factions or opinions (Allport, 1954; Dekker, Belabas, & Scholten, 2015). It is difficult to prove "in the wild" that the contact hypothesis is working, but the theory suggests sustained civil conversations are a successful form of counterspeech.

More counterspeech

The final response that we have observed is the creation of more counterspeech by other Twitter users, or 'cyberbystanders.' It is difficult to know whether subsequent counterspeakers were drawn into a conversation by the initial counterspeech, but we have indications that this is the case and can signal success in the form of favorable impact on the audience. During the #FuckPhyllis incident, one University of Illinois alumna credited another counterspeaker with inspiring her own counterspeech against the hashtag. She said, "*I knew [redacted] was an alum and she inspired me to use social media for good, but I guess my classmates didn't get the memo. #FuckPhyllis*." Similarly, dozens of Twitter users followed the example of comedian Michael Ian Black, after he discovered Megan Phelps-Roper's Twitter account and began counterspeaking to it. Cyberbystanders, like offline bystanders, typically choose not to act when they perceive that many others are present and watching (Dillon & Bushman, 2015; Dillon, 2015), but they are much more likely to do so once someone else acts (or counterspeaks) (Markey, 2000).

This response could be successful in the sense that it may have significant impact on the audience. We speculate that a surge in counterspeech might acclimate other users to it, so that some of them begin to practice it. If a group of new counterspeakers becomes vocal, even if still in the minority, that may be influential in shifting discourse norms (Rosenberg, 2011; Gladwell, 2000).

This type of response is not without its dangers, however. Like counterspeech that silences, counterspeech which creates more counterspeech can lead to dogpiling. Avoiding this, like silencing, requires good judgment and self-control on the part of counterspeakers.

Conclusions and Ideas for Future Research

We hope to have provided some useful ideas on counterspeech, which is worth further exploration although it is surely not, by itself, a solution to the acute problem of vitriol and extremism online.

We suggest that future research continue our field study approach, gathering larger datasets and working in other languages and national contexts, in order to compare datasets. Work on these future datasets should include a codebook or qualitative memoing system.

It may also be useful to conduct laboratory experiments on the communicative strategies for social media that we've witnessed online. It would be especially important to test the effects of various forms of counterspeech on an 'audience,' which would probably be more feasible in a laboratory experiment than in an observational study like ours.

Another line of research would focus on the use of hashtags as hateful speech or counterspeech. It would be useful to examine multiple cases in which hateful hashtags are reclaimed, for example, especially when they trend after events that ignite strong emotions in large populations of people, such as bombings and shootings.

Other important questions to be tackled include: Which sorts of counterspeech work best for what sorts of subjects (people)? Are certain types of counterspeech (using particular strategies) more effective in response to particular forms of hateful or dangerous speech or extremism (e.g. racism vs. misogyny, white supremacy vs. Islamic extremism)? Does counterspeech differ - in nature and effectiveness - when conversations are among strangers with no ties, people who know each other, or people with some offline shared identity or connection (e.g. students and alumni of the University of Illinois)?

Finally, it could be of great interest to interview individuals who changed their speech or attitudes after online exchanges, to learn more about how and why such change takes place, and to interview counterspeakers to learn about their reasoning and thought process as they debate people who are, most often, perfect strangers.

Works Cited

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Perseus Books.

Amichai‐Hamburger, Y., & McKenna, K. Y. (2006). The contact hypothesis reconsidered: Interacting via the Internet. *Journal of Computer Mediated Communication*, *11*(3), 825-843.

Anderson, J. Q., & Rainie, L. (2010). The future of social relations. *Pew Research Center's Internet & American Life Project*, *2*.

Anti-Defamation League. (2014). *Best practices for responding to cyberhate.*

Barrett, R. (2014). The Islamic State. *The Soufan Group.*

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-speech: Examining content that challenges extremism online. *Demos*.

Bauman, R., & Briggs, C. L. (1990). Poetics and performance as critical perspectives on language and social life. *Annual Review of Anthropology*, 19, 59–88.

Baym, N. K. (1995). The performance of humor in computer‐mediated communication. *Journal of Computer Mediated Communication*, *1*(2), 0-0.

Bacchi, U. (2014, July 30). #DirenKahkaha: Turkish women in social media 'laugh protest' against Erdogan's deputy. *The International Business Times*. Retrieved at http://www.ibtimes.co.uk/direnkahkaha-turkish-women-social-media-laugh-protest-against-erdogans-deputy-1459005.

Berger, J. M., & Morgan, J. (2015). The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Project on US Relations with the Islamic World*, *3*(20).

Berger, J.M., & Strathearn, B. (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.*

Beauvois, J. L., Joule, R. V., & Brunetti, F. (1993). Cognitive rationalization and act rationalization in an escalation of commitment. *Basic and Applied Social Psychology, 14*, 1-17.

Benesch, S. (2013). Proposed guidelines on dangerous speech. *The Dangerous Speech Project.* Retrived from http://dangerousspeech.org/guidelines

Benesch, S. (2014). Defining and diminishing hate speech. In *Freedom from hate: State of the world's minorities and indigenous peoples 2014* (pp. 18-25). London: Minority Group International.

Bipartisan Policy Center. (2012). *Countering online radicalization in America.* Washington, DC: Neumann.

boyd, d., Golder, S., & Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (pp. 1-10). IEEE.

Brewer, M. B., & Brown, R. J. (1998). Intergroup relations. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4 ed., Vol. 2, pp. 554–594). New York: McGraw-Hill

Briggs, R., & Feve, S. (2013). Review of programs to counter narratives of violent extremism. *Institute of Strategic Dialogue*.

Brown, R. (2016). Defusing hate: A strategic communication guide to counteract dangerous speech. *United States Holocaust Memorial Museum.*

Bruns, A., & Burgess, J. E. (2011, August). The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*.

Chen, A. (2015, November 23). "Unfollow: How a prized daughter of the Westboro Baptist Church came to question its beliefs." *New Yorker*. Retrieved from http://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper

Citron, D. K. (2014). *Hate crimes in cyberspace*. Cambridge, MA: Harvard University Press.

Citron, D. K., & Norton, H. L. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*, 1435.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *ICWSM*, *133*, 89-96.

Dekker, R., Belabas, W., & Scholten, P. (2015). Interethnic Contact Online: Contextualising the Implications of Social Media Use by Second-Generation Migrant Youth. *Journal of Intercultural Studies*, *36*(4), 450-467.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying. *The Social Mobile Web*, *11*, 02.

Dillon, K. P. (2015, November). "Help me help you: Victim communication in cyberbystander intervention in cyberbullying," Paper presented at the National Communication Association 101st Annual Conference, Human Communication & Technology Division, Las Vegas, NV.

Dillon, K. P., & Bushman, B. J. (2015). Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, *45*, 144-150.

Frenett, R., & Dow, M. (2015). One to one online interventions: A pilot CVE methodology. *Institute for Strategic Dialogue.*

Foxman, A. H., & Wolf, C. (2013). *Viral hate: Containing its spread on the Internet*. New York, NY: Macmillan.

Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics*, *14*(2), 219-236.

Gladwell, M. (2006). *The tipping point: How little things can make a big difference*. Boston: Little, Brown.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Herz, M., & Molnar, P. (2012). *The content and context of hate speech: Rethinking regulation and responses*. New York, NY: Cambridge University Press.

Hussain, G., & Saltman, E.M. (2014). Jihad trending: A comprehensive analysis of online extremism and how to counter it. *The Quilliam Foundation.*

Joffe, H. (2008). The power of visual material: Persuasion, emotion and identification. *Diogenes*, *55*(1), 84-93.

Kane, A. A., Argote, L., & Levine, J. M. (2005). Knowledge transfer between groups via personnel rotation: Effects of social identity and knowledge quality. *Organizational Behavior and Human Decision Processes*, *96*(1), 56-71.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schweider, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics, 62*(3), 790–816.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the 27th Association for the Advancement of Artificial Intelligence Conference, 1621-1622*.

LaGrandeur, K. (2003). "Digital images and classical persuasion" in Mary E. Hocks & Michelle R. Kendrick (Eds.) *Eloquent images: Word & image in the age of new media*. Cambridge: MIT Press.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13(3*), 106-131.

Liebrecht, C. C., Kunneman, F. A., Bosch, A. P. J. van den (2013). The perfect solution for detecting sarcasm in tweets #not. In Balahur, A.; Goot, E. van der; Montoyo, A.

(Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 29-37.

Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, *16*(2), 183-188.

Marone, V. (2015). Online humour as a community-building cushioning glue. *The European Journal of Humour Research*, *3*(1), 61-83.

McDonald, S.N. (2014, December 2). 'Racists Getting Fired' exposes weaknesses of Internet vigilantism, no matter how well-intentioned. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/morning-mix/wp/2014/12/02/racists-getting-fired-exposes-weaknesses-of-internet-vigilantism-no-matter-how-well-intentioned/

Mendel, T., Herz, M., & Molnar, P. (2012). Does international law provide for consistent rules on hate speech? *The content and context of hate speech: Rethinking regulation and responses*, 417–429.

Michael, K., & Young, A. D. (2015, June 11). "Frenship ISD teacher apologizes after McKinney-related segregation post," *Lubbock Avalanche-Journal*. Retrieved from http://lubbockonline.com/education/2015-06-10/frenship-isd-teacher-apologizes-after-mckinney-related-segregation-post#.V1iqzasaLZo

Muslim Advocates. (2014). *Click here to end hate: Anti-Muslim bigotry online & how to take action.* Retrieved from http://www.muslimadvocates.org/wp-content/uploads/Click-Here-to-End-Hate.pdf

Nadler, A., & Liviatan, I. (2006). Intergroup reconciliation: Effects of adversary's expressions of empathy, responsibility, and recipients' trust. *Personality and Social Psychology Bulletin*, *32*(4), 459-470.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303-330.

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*(3), 459-464.

Oring, E. (1992). *Jokes and their relations*. Lexington: University Press of Kentucky.

Payne, S. (2015, January 20). An amazing woman fields a troll on MLK Day and it was nothing short of inspirational. [Blog]. Retrieved from https://www.dailykos.com/story/2015/1/20/1359055/-An-amazing-woman-feeds-a-troll-on-MLK-Day-and-it-was-nothing-short-of-inspirational

Prior, M. (2003). Liberated viewers, polarized voters: The implications of increased media choice for democratic politics. *The Good Society, 11*, 10–16.

Ronson, J. (2015, February 12). How one stupid tweet blew up Justine Sacco's life. *The New York Times.* Retrieved from http://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html

Rosenberg, T. (2011). *Join the club: How peer pressure can transform the world*. New York, NY: WW Norton & Company.

Rutkowski, G. K., Gruder, C. L., & Romer, D. (1983). Group cohesiveness, social norms, and bystander intervention. *Journal of Personality and Social Psychology, 44(3)*, 545-552.

Saleem, H. M., Dillon, K. P., Benesch, S., Ruths, D. (2016). A web of hate: Tackling hateful speech in online social spaces. *European Language Resources Association.*

Saltman, E. M., & Russell, J. (2014). The role of Prevent in countering online extremism. *The Quilliam Foundation.*

Simeone, M. (2014, January 29). Twitter outrage, charted: The partial anatomy of the #FuckPhyllis trend, or why I don't trust BuzzFeed. [Blog]. Retrieved from https://suffenus.wordpress.com/2014/01/29/twitter-outrage-charted-the-partial-anatomy-of-the-fuckphyllis-trend-and-why-i-dont-trust-buzzfeed/

Stevens, M. (2013, May 13). FAQ: The geography of hate. [Blog]. Retrieved from http://www.floatingsheep.org/2013/05/faq-geography-of-hate.html

Stevens, M. (2013, May 13). The geography of hate. [Blog]. Retrieved from http://www.floatingsheep.org/2013/05/hatemap.html

Street, R. L., & Giles. H. (1982). Speech accommodation theory: A social cognitive approach to language and speech behavior. In M. Roloff & C. Berger (Eds.), *Social cognition and communication* (p. 205-255). Beverly Hills, CA: Sage.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321-326.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information, 13*, 65-93.

Tanis, M., & Postmes, T. (2003). Social cues and impression formation in CMC. *Journal of Communication*, *53*(4), 676-693.

Walid, D. (2013, November 24). "Responses to my calling out the term 'abeed'." Retrieved from https://dawudwalid.wordpress.com/2013/11/24/responses-to-my-calling-out-the-term-abeed/

Warner, W. & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, 19–26. Association for Computational Linguistics.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 1980–1984. ACM.

Zuckerman, E. (2013). *Digital cosmopolitans: Why we think the Internet connects us, why it doesn't, and how to rewire it*. WW Norton & Company.