


# Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces

Social Media + Society  
January-March 2019: 1–15  
© The Author(s) 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2056305119832588  
journals.sagepub.com/home/sms  


**Anna Gibson**

## Abstract

How do moderation policies affect online discussion? This article analyzes nearly a quarter of a million anonymous comments over a 14-month period from two online Reddit forums matched in topic and size, but with differing moderation policies of “safe space” and “free speech.” I found that in the safe space, moderators removed significantly more comments, and authors deleted their own comments significantly more often as well, suggesting higher rates of self-censorship. Looking only at relatively low frequency posters, I found that language in the safe space is more positive and discussions are more about leisure activities, whereas language in the free speech space is relatively negative and angry, and material personal concerns of work, money, and death are more frequently discussed. Importantly, I found that many of these linguistic differences persisted even in comments by users who were concurrently posting in both subreddits. Altogether, these results suggest that differences in moderation policies may affect self-censorship and language use in online space, implicating moderation policies as key sites of inquiry for scholars of democratic discussion.

## Keywords

reddit, moderation, online moderation, communication accommodation theory, linguistic style matching, civil discussion, self-censorship

## Introduction

Habermas (1991) articulated the concept of the public sphere as a space for private citizens to come together as a public to deliberate. In the age of digital connections, the Internet has been theorized as a new space for these kinds of discussions of civic importance (Papacharissi, 2004; Price, Nir, & Cappella, 2006). The accessibility and visibility of networked digital media have been considered democratizing forces that can allow even more citizens to participate in the kind of public discussion needed for a democracy (Dylko et al., 2012; Papacharissi, 2002). Online discourse has had some striking positive effects on politics across the world through the last decade, from the Arab Spring uprisings of 2011 (Lotan, Graeff, Ananny, Gaffney, & Pearce, 2011; Wilson & Dunn, 2011) to #MeToo (Hawbaker, 2018; Pazzanese & Walsh, 2017). In a comparison of online comments to letters to the editor published in newspapers, McCluskey and Hmielowski (2012) found that online comments on a news story were more diverse in viewpoints and more likely to challenge institutions than conventional letters to the editor.

However, Internet discussion spaces still face the same issue that all other spaces for public discussion in the past have faced: how can—and should—these spaces be designed to meet the needs of a public sphere? Ever since the creation of online spaces for discussion, there has been controversy about the extent to which speech should be regulated and controlled (Pfaffenberger, 1996). According to a 2018 Pew poll, most US teens have experienced cyberbullying, including ridicule, threats, or false rumors (Anderson, 2018). Recent years have also seen horrific mass violence, such as the killing of Muslim minorities in Sri Lanka, as a result of misinformation and hate speech spread on online discussion spaces (Taub & Fisher, 2018).

In the face of all the potentials and dangers, what kind of speech should be allowed in online spaces, and to what end?

---

Stanford University, USA

### Corresponding Author:

Anna Gibson, Department of Communication, Stanford University, Building 120, Room 110, 450 Serra Mall, Stanford, CA 94305-2050, USA.  
Email: agibson2@stanford.edu



On its face, this is a normative question for political philosophers. However, it can also be approached empirically. This article will investigate one specific comparative case of how the values-driven moderation strategies of “safe spaces” and “free speech” shape discussion in two Reddit discussion forums.

### *Free Speech and Safe Space*

A policy of freedom of speech has been justified through its potential to unearth truth and create the social conditions necessary for democracy to flourish, as people can make informed decisions and check those in power (Ananny, 2018; Ash, 2016; Habermas, 1991). Speech is also a defining characteristic of humans, and it has been argued that only through total freedom of speech can a person fully explore and understand their own humanity (Ash, 2016, p. 73).

Culturally, freedom of speech from government censorship is dearly held American value; according to a 2015 Pew poll, Americans have the strongest support for free expression in the world (Wike, 2016). The same poll also reports that Americans are more tolerant of offensive speech than any other country.

A libertarian commitment to total freedom, including freedom of speech, has been a central part of Internet culture since at least the establishment of Usenet as an alternative to ARPAnet in 1979 (Pfaffenberger, 1996; Reagle, 2013). Many of the early prominent figures of the Internet advocated a “hacker” ethos, which celebrated the liberation of systems and information from any form of centralized control (Reagle, 2013; Turner, 2006). Early online spaces, such as Usenet, reflected that demand for freedom through cultural norms of users, as well as the technical affordances of the systems themselves (Pfaffenberger, 1996). Users abhorred the threat of censorship (Pfaffenberger, 1996) and uplifted the idea of the “rational, autonomous individual” (Turner, 2006). Pfaffenberger (1996) summarizes this attitude as “[it’s] up to the individual user, not some committee or administrator, to decide what’s worth reading” (p. 369). This libertarian view advocates equal access to discursive spaces for all without threat of intervention (Fiss, 2009). As the president of the American Civil Liberties Union has suggested, in a land of free speech, the solution to hate speech is more speech (as paraphrased in Marwick, 2017).

Fiss (2009) argues, however, that there is an inherent irony to free speech; even in the most open space, some voices will tend to have more access to this “free” discursive space than others because of social and economic hierarchies. Similarly, Fraser (1990), in her critique of the Habermasian public sphere, asserts that the bracketing process that Habermas proposes—to leave behind all individual markers such as class to engage in rational political deliberation—is simply impossible and minority groups must form their own subaltern counterpublics.

Indeed, there is substantial evidence that many “free” online spaces exhibit the sexist and racist tendencies of broader culture. Herring (1996) found that even in theoretically free and neutral spaces, women were discouraged from participating in favor of the domination of a small and vocal sect of men. Attempts by women to assert their presence were met by silence or attempts to delegitimize their statements (Herring, 1996, p. 486). Gray (2012) documented the persistent racist and sexist abuse experienced by black female gamers in Xbox Live gaming spaces. Reagle (2013) examined why women are disproportionately underrepresented on Wikipedia as compared with other online spaces, despite the site’s explicit embrace of a free culture. He suggests that free culture in online environments is laden with historical and cultural connotations. Thus, online spaces labeled as “free” may in fact signpost a space for argumentative, male-dominated discussion. In a more extreme example, Marwick and Lewis (2017) note that “commitments to ‘free speech’ in certain communities can serve as an on-ramp for far-right radicalization” (p. 46).

As Freeman examined in her 1972 paper “The Tyranny of Structurelessness,” even when formal structures of power are eliminated, informal structures persist and these structures tend to benefit some members and punish others. “Contrary to what we would like to believe, there is no such thing as a ‘structureless’ group” (Freeman, 1972, p. 152).

### *Safe Spaces*

The meaning of safe space has been widely contested in academic literature as well as popular culture (Barrett, 2010; Harris, 2015; Stengel & Weems, 2010). Several scholars have noted that the metaphor of a safe space is both overused and undertheorized (Boostrom, 1998). For the purposes of this empirical study, however, I will follow the Roestone Collective (2014) in their calls to understand safe spaces through their relational work, as sites for “negotiating difference and challenging oppression.”

In her book “Mapping Gay L.A.,” Kenney (2001) traces the development of the safe space discourse around the creation of radical feminist spaces in the 1970s: “the notion of safe space implies a certain license to speak and act freely, form collective strength, and generate strategies for resistance” (p. 24). She notes that in this context safe spaces acted as “a means rather than an end” for women to find strength and community.

The Roestone Collective writes that “the categories of safe and unsafe are socially produced and context dependent” (p. 1350). Thus, safe spaces arise in social contexts of specific threats to specific groups. In her 2018 ethnographic case study of one online safe space, the Facebook group Girl Army, Clark-Parsons identifies the need for such spaces in the context of rhetoric that seeks to silence women’s participation. Within these spaces, “marginalized users [can] speak

freely, seek support, and organize action against injustices faced outside the group's boundaries" (p. 2127).

Such spaces are never understood as perfectly safe, but perpetually doing the work of "negotiating and foregrounding difference" (Roestone Collective, 2014). This differs from the caricature of such spaces as infantile playgrounds (McKee, 2015; @RichardDawkins, 2015). Following the work of Foucault (1978), speakers are understood not just as neutral voices, but emanating from bodies situated within interconnected webs of power: "[power] is not something that is acquired, seized, or shared . . . power is exercised from innumerable points, in the interplay of nonegalitarian and mobile relations." Thus, the way in which power structures discourse and the relative positionality of speakers is salient in deciding what kind of language to allow in these spaces.

Policies of safe spaces are thus concerned with preventing the marginalization of voices already hurt by dominant power relations. This may be implemented through strict no-tolerance policies of "hate speech" or other discussion that would undermine the political project assumed in the space of the community. In practice, this often means that people can be censored or ejected from a space for not properly observing the standards of speech, tone, or style (Clark-Parsons, 2018). This includes not only hateful statements but also ignorantly prejudiced or unintentionally traumatizing topics without giving notice to readers in the form of trigger or content warnings (Manne, 2015).

### *Spiral of Silence*

In a highly influential 1974 paper, Noelle-Neumann coined the term "spiral of silence" to describe why members of the public with minority perspectives on value-laden topics stayed silent. According to her theory, if individuals perceive their opinions to be in the minority, they will refrain from expression out of fear of social isolation. This tendency is self-reinforcing, as a lack of public support for a minority position will discourage others with that opinion from expression. Thus, public expression is understood not through a Habermasian lens of rational debate, but performative social interaction (Scheufle & Moy, 2000). The theory has been met with limited, but positive, empirical support in the pre-digital age (see Scheufle & Moy, 2000, and Glynn, Hayes, & Shanahan, 1997 for reviews). Online support for the spiral of silence has been found across a variety of digital spaces (De Koster & Houtman, 2008; Gearhart & Zhang, 2018; Hampton et al., 2014; Liu, Rui, & Cui, 2017; Meyer & Speakman, 2016; Schulz & Roessler, 2012; Stoycheff, 2016). McDevitt, Kioussis, and Wahl-Jorgensen (2003) compared discussions about the topic of abortions between individuals who were face to face or connected through computer messaging and found that individuals in the computer condition were perceived as more moderate.

However, several recent studies suggest that the effects of a theorized spiral of silence are only exhibited by specific

people with regard to specific topics. Hayes, Glynn, and Shanahan (2005) developed a scale to measure a construct at the individual level they called the "willingness to self-censor." In an experiment designed to put participants in a spiral-of-silence situation, Hayes, Uldall, and Glynn (2010) found that the willingness to self-censor was driven almost entirely by those who had scored high on the scale; for individuals who were not dispositional self-censors, opinion climate made no difference in opinion expression. A study by Matthes, Morrison, and Schemer (2010) examined the role of "attitude certainty" with regard to the spiral of silence and concluded that individual differences in the strength of attitude affected susceptibility to a spiral of silence.

Topic, too, appears to make a difference. Gearhart and Zhang (2018) found that the degree to which a topic has been an enduring, emerging, or transitory subject of public debate affected the willingness of individuals to voice minority opinions. Porten-Che   and Eilders (2015) failed to find support for the spiral of silence, but conceded that their chosen topic of public discussion, climate change, actually had little to no moral conflict in its German context.

Neubauer and Kr  mer (2018) propose that the social environment in which the opinion is expressed may also affect the tendency of individuals to be susceptible to the spiral of silence. Different social environments may yield different expected sanctions for transgressions, changing the situational fear of isolation and correspondingly changing the stakes in deciding to share an opinion perceived to be in the minority.

### *Moderation Policies*

The history of online discussion is replete with stories of users deliberately deceiving (Van Gelder, 1985) and hurting (Dibbell, 1993) fellow forum users. Many online discussion spaces have some sort of moderation policies and/or structure in place to address the needs of users (Gillespie, 2018; Grimmelman, 2015).

Like other spaces for debate, online forums also use moderators—usually a computer program and/or a designated person—to determine and enforce the baseline rules of discussion. Moderators play an important role in preventing disruptive users like trolls or spam from taking over forums (Brunton, 2013). Moderator powers often include the ability to screen, modify, and delete comments, or ban users (Matias, 2016). Consequently, moderators have more power to affect the discussion in online forums than other forum participants. Therefore, the forum policies established by moderators, and the effects they have on discussion or spirals of silence, are important to investigate and understand.

There have been many studies of what kinds of online community structures facilitate community growth and longevity (see, for example, Aumayr & Hayes, 2017; Hinds & Lee, 2008; Lin et al., 2007; Wagner, Liu, Schneider, Prasarnphanich, & Chen, 2009), but this is the first study to

use a matching methodology to compare effects of differences in moderation policy.

## Method

### *Language Accommodation Theories*

Speakers adjust their communication within different groups and contexts with respect to self and group identity (Dragojevic, Gasiorek, & Giles, 2015). Such accommodation, often understood through the framework of communication accommodation theory (CAT), happens both at a conscious level and at an unconscious level. Language is often used to negotiate in-group social identity and is perceived in accordance with local sociocultural norms around language. A person who wants to be seen as a member of the “in-group,” for example, may consciously or unconsciously adopt the group’s linguistic features, such as intonation, accent, or word usage. Such changes are examples of accommodation. On the other hand, a person who is trying to distance themselves from the group may emphasize linguistic features that differ from the norms of the group; this is linguistic divergence. In addition, linguistic accommodation may serve an instrumental end, such as increasing communication efficiency (Dragojevic et al., 2015).

Linguistic style matching (LSM), as defined by Niederhoffer and Pennebaker (2002), builds on CAT. One prominent part of LSM is that within groups, word use covaries to reflect how “in-sync” speakers are. The mechanism in this situation is priming: the words one speaker uses prime the other’s response.

LIWC, or Linguistic Inquiry and Word Count, sorts and displays the words present in text into a variety of categories by frequency (Tausczik & Pennebaker, 2010). The word types included linguistic categories (e.g. word count, articles, and prepositions), social/affect categories (e.g. first-person singular pronoun usage, positive, and negative emotion), and cognitive categories (e.g. tentative, certainty). This tool has been shown as an effective way to study how groups converge under LSM using transcripts of conversations (Ireland et al., 2011; Niederhoffer & Pennebaker, 2002; Scissors, Gill, Geraghty, & Gergle, 2009; Scissors, Gill, & Gergle, 2008) as well as text from computer mediated writing and discussion (Gonzales & Hancock, 2008; Gonzales, Hancock, & Pennebaker, 2010; Ireland & Pennebaker, 2010).

Because LIWC analyzes categories of invisible structural “style” words, such as pronouns, that are theoretically rich and clinically tested but may otherwise go unnoticed, LIWC provides a method to study group language convergence at mass scale in an unbiased and meaningful way (Pennebaker, 2011).

### *Object of Study*

According to Alexa.com, as of 18 October 2017, Reddit is the eighth most popular website globally, ranking above

Amazon and Twitter. In the United States, it is the fourth most popular, below only Google, YouTube, and Facebook.

The site is organized into communities, called subreddits, which users can subscribe to or visit (“About,” n.d.). Users with accounts can post text, links, or images into these subreddits and also comment in response to the posts. Anyone can create an account without an email address and begin posting and commenting immediately. Users tend to remain anonymous or use pseudonyms and connect with strangers rather than friends or family (Bergstrom, 2011; Lamont, 2014; Shelton, Lo, & Nardi, 2015). Redditors can vote posts and comments “up” or “down” which will affect the post or comment’s public score and subsequently how easily other redditors and the general public will see that post or comment (Grimmelmann, 2015).

The structure of Reddit into subreddits makes it an interesting site for study because there is not a uniform policy of moderation across the site. The company takes a hands-off stance with regard to content in favor of letting individual subreddits make and enforce their own moderation policies, as detailed in a blog post from 2014 titled “Every Man Is Responsible for His Own Soul”:

We uphold the ideal of free speech on reddit as much as possible not because we are legally bound to, but because we believe that you—the user—has the right to choose between right and wrong, good and evil, and that it is your responsibility to do so. When you know something is right, you should choose to do it. But as much as possible, we will not force you to do it.

You choose what to post. You choose what to read. You choose what kind of subreddit to create and what kind of rules you will enforce. We will try not to interfere—not because we don’t care, but because we care that you make your choices between right and wrong.

The established norms of acceptable discourse in any given subreddit can vary wildly; moderators create and post rules and then are expected to enforce those policies accordingly. Reddit users can join or leave communities in reaction to those policies. Therefore, it can be posited that moderation policies act as an independent factor in the study of subreddit discourse. Studying the differences in discourse between subreddits with differing moderation policies may therefore provide evidence of the effects of those policies on commenter discourse.

Of course, there are many variables that may affect discussion within a subreddit: size, topic, and relative visibility are some of the most prominent factors. All of these may have much more profound effects on the observed measures of discourse than moderation style. Therefore, any investigation of subreddit discourse will need to account for these confounding variables. To compare the effects of moderation policy, I will use a matching strategy and analyze two subreddits that are alike in almost every way except for moderation policy.

## Subreddits in this Study

The two subreddits chosen for study in this article are r/lgbt ([www.reddit.com/r/lgbt](http://www.reddit.com/r/lgbt)) and r/ainbow ([www.reddit.com/r/ainbow](http://www.reddit.com/r/ainbow)).

According to RedditList.com (n.d.), as accessed 18 October 2017, r/lgbt has approximately 170,000 subscribers, and r/ainbow has approximately 49,000 subscribers. For context, of the over 4,000 subreddits tracked by RedditList.com, the top 100 subreddits have over 650,000 subscribers each. The median subreddit (r/grilling, ranked 2, 135th) only has approximately 24,000 subscribers. Even though these sizes are not a perfect match, they are acceptable relative to the spread of community sizes on Reddit.

Both subreddits have similar self-descriptions, identifying themselves as spaces for lesbian, gay, bisexual, and transgender (LGBT) and other to have discussions. The r/lgbt subreddit defines itself in the following way:

This subreddit is by and for people who are Gender and Sexual Minorities (GSM), including but by no means limited to LGBT (Lesbian, Gay, Bisexual and Transgender) people, and respect for our diversity and experiences is paramount. All are welcome to participate who agree to follow the rules outlined below and in: The r/lgbt FAQ

It also boasts a “Safe Space” badge icon on its sidebar. The r/ainbow subreddit posts this self-description:

A free area for the discussion of issues facing those who identify as gay, lesbian, bisexual, transgender, and all other sexual or nonsexual orientations and/or gender identities. Post links to articles, self-posts, photographs, experiences and whatever else is important to your experience of queer life. We encourage you to treat others with respect, start and/or engage in robust discussion and interact with the community. The more we know each other, the better we’ll get along.

These the two subreddits explicitly define the ideological bases for their policies and then link to each other as alternative, making them excellent sites for studying the effects of moderation policy. For example, r/lgbt explicitly bills itself as a safe space and links to r/ainbow as an alternative, less moderated space:

This is a safe space. Anyone can make a mistake and accidentally say something hurtful or triggering. If you find yourself corrected for making this error, please try to learn from it. This is not a place to tell people that they need to reclaim a pejorative so you can use it, that they should laugh at jokes about them, or that they otherwise just ‘shouldn’t be so sensitive.’ For lightly moderated LGBT-related discussion, I recommend r/ainbow. r/ainbow does not moderate discussion, but the community will expect that you treat them with respect. For more information, see r/ainbow’s FAQ.

Similarly, r/ainbow describes itself a free speech area and links to r/lgbt as an alternative, more moderated space:

This subreddit is lightly moderated. The community actively self-moderates offensive comments with downvotes, but comments are generally not removed except for violations of site-wide guidelines and as outlined below. If you prefer a more hands-on approach, try r/lgbt. r/lgbt requires trigger warnings, and removes comments and users for violations of their rules, which are detailed in their FAQ.

This subreddit is a free speech zone . . . It is also lightly moderated, which means that it’s up to you the community to downvote offensive posts and comments, and upvote constructive content. Please use your voting and posting powers to create the community you want to see.

In an exploratory study, there were very few other similarly paired subreddits. Often, it appeared that users who felt they were being censored started free speech subreddits in reaction to new policies in larger subreddits. Such new free speech subreddits, in addition to being smaller than the original subreddit by several orders of magnitude, generally centered reactionary politics and rejected the safe-space subreddits as being compromised by “SJWs” (frankenmine, 2015) or Social Justice Warriors, a pejorative term ascribed to individuals concerned with identity politics or feminism (Massanari & Chess, 2018). In contrast, the subreddits studied here, r/lgbt and r/ainbow, have relatively more similar political standpoints, activity, and subscribers.

## Research Questions

### Comment Deletion

Heavily moderated spaces entail more moderator intervention, including deleting offending comments. Along these lines, I wanted to explore whether moderators will delete more comments in a safe space than the free speech subreddit. This might also manifest as more pressure for users to delete their own comments, either due to intrinsic motivations like personal shame, or extrinsic factors, like other users downvoting that comment overwhelmingly. That is, I suspect more moderation to lead to not only greater amounts of moderator intervention, specifically to remove comments, but also to have that reflected in users’ own self-monitoring and self-censorship of their own comments.

*Research Question 1a.* Do moderators delete more comments in r/lgbt, the safe space, than r/ainbow, the free speech space?

*Research Question 1b.* Do more users delete their own comments in r/lgbt, the safe space, than r/ainbow, the free speech space?

### Participation

When a user creates an account Reddit, he or she is automatically “subscribed” to a number of popular subreddits, meaning

posts from those subreddits will comprise that user's Reddit homepage. A user can then change what subreddits' posts they want to see by either subscribing to or unsubscribing from as many subreddits as they want. Not all users who subscribe to a subreddit will post or even comment in that subreddit, remaining "lurkers" who watch without interacting through text. I was curious to see if there was a difference between the two subreddits in the proportion of users who commented in that subreddit.

As discussed in the literature review, safe spaces are concerned with minimizing the structural barriers that prevent individuals in different positions of power from participating in discussions. Thus, if safe space moderation policies are successful, a greater proportion of the subscribers for the safe space subreddit will actually participate in the subreddit by writing a comment. Conversely, it could be argued that the heightened standards for entry in a safe space dissuade individuals who are already hesitant about participation.

*Research Question 2.* How do the safe space, r/lgbt, and the free speech space, r/ainbow, differ in overall participation rates by subscribers?

### Language Use

Free speech and safe spaces have been characterized by the ways in which participants interact; discussion in free speech spaces has been described as overly rational and argumentative (Reagle, 2013), and discussion in safe spaces as overly emotional, deferent, and meek (Lukianoff & Haidt, 2015; Massanari & Chess, 2018). To explore whether these differences were real, I decided to analyze language use using LIWC2015 (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

The LIWC2015 program can analyze 90 different variables for each text, and thus I had many categories through which I could compare the two subreddits. Rather than simply going through all 90 variables and looking for differences, I chose to focus in on a few categories that pertain specifically to interpersonal attitude and boundary negotiation, as I believed these would most starkly draw out linguistic differences between the communities. I chose to examine five categories of language from LIWC: summary variables, linguistic dimensions, affective processes, personal concerns, and informal language.

Word choice can be used to indicate the attentional focus, or the "gaze" of a speaker, and therefore identifying common themes in words can serve to reflect the speaker's mental state (Tausczik & Pennebaker, 2010). For example, as individuals identify more and more with a group, they will shift from using first-person singular pronouns ("I" and "me") to first-person plural pronouns ("we" and "us"). First-person plural pronouns, however, can also be used to insinuate distance, as in an exclusive we-and-not-you, Royal-we, or we-as-those-who-agree-with-me ("We need to make America great again"). Thus, the usage frequency of first-person

pronouns is not only a reflection of group identity but also status and arrogance, among other traits (Pennebaker, 2011; Pennebaker & Lay, 2002).

*Research Question 3.* How does language use differ between r/lgbt, the safe space subreddit, and r/ainbow, the free speech subreddit?

### Crossposters

To test whether these differences arise because of author traits or contextual cues in the individual subreddits, I decided to isolate a subset of comments created by authors who had concurrently posted in both subreddits. I can then examine whether these comments by "recent crossposters" exhibited the same kinds of differences in linguistic markers. If these differences persist in this subset of comments, it would provide evidence that language differences are influenced by contextual language usage rather than innate author characteristics.

*Research Question 4.* Do users who post in both the free speech and safe space subreddits show differences in the language they use in each subreddit?

### Data Analysis

A total of 2,76,574 comments from r/lgbt and r/ainbow made between 1 June 2016 and 31 July 2017 (inclusive) were downloaded from the BigQuery Reddit repository. These dates were chosen to include a number of recent American political events relevant to the LGBT community which might spark civic discussion, including the Orlando Pulse massacre (12 June 2016), the American presidential election (4 November 2016), and the presidential announcement through a tweet to exclude transgender individuals from the military (@realDonaldTrump, 2017). For each comment, the initial data set contained the comment author, subreddit, month and year posted, and full body text.

### Data Cleaning

A "bot" is a pre-programmed reddit user that posts a comment when triggered by a certain type of comment made by another user. Various bots do such diverse things as transcribe comic strips (u/ImageTranscribingBot), fix common spelling errors (u/CommonMisspellingBot), or post pictures of cats for sad people (u/ThisCatMightCheerYou). Because bots post indiscriminately, they were excluded from the data set for comment analysis.

There is no fixed list of bots, so I created my own list to exclude comments from those authors. First, all usernames and comments from the sample containing the word "bot" were examined to find posting patterns that corresponded to bot behavior, including multiple postings in a row of the

same or similarly formatted post. Suspected bots were further researched by looking up their username on Reddit and/or archived Reddit pages. This yielded a list of 62 bots and covered 1,170 comments in my sample.

Next, I compiled a list of alleged bots from Goodbot-Badbot’s voting site (Good Bot, Bad Bot). This list is automatically created through user votes and potentially contained false positive bot identifications. The list contained 1,116 bot names and covered 1,196 comments in my sample. Only 33 usernames appearing in the Goodbot-Badbot list were both in my data set and not on my previous list of bot names. These suspected bots were further investigated and yielded an additional 24 bots.

In total, my list of bots from my sample contained 86 usernames, and these bots created 1,168 comments. A total of 652 were in r/lgbt and 516 were in r/ainbow. After removing these comments, my data set contained 2,75,406 comments.

In my remaining data set of comments, many comment entries were incomplete. If a user deletes their comment or if that comment is removed by a moderator, Reddit indicates that this action has taken place by replacing the body of the comment as either “[deleted]” or “[removed],” respectively. If that action took place between the creation of the comment and the archiving of the comment into the BigQuery Reddit repository, the body of the comment is preserved as only “[deleted]” or “[removed].”

Many users go farther, by not just deleting their comments, but by deleting their user accounts as well. In this case, username attached to the comment will be replaced by “[deleted].” It is not uncommon to find vestigial comments in comment threads with a comment that says “[deleted]” posted by a “[deleted]” author.

My data set thus also contained many of these “null” comments. Because they are not useful for a linguistic analysis, I separated these comments into a separate data set.

**Table 1.** Authorship in Subreddits.

	r/lgbt (safe space)	r/ainbow (free speech)
Total authors	31,209	9,215
Authors who posted		
Only in one subreddit	27,787 (89.0%)	5,793 (62.9%)
Some point in both	3,422 (11.0%)	3,422 (37.1%)
Both in same month	2,154 (6.90%)	2,154 (23.4%)

**Table 2.** Relative Size of Comment Groups.

Subreddit	Total comments	Level 1		Level 2		Level 3	
		No. of comments	Total (%)	No. of comments	Total (%)	No. of comments	Total (%)
r/lgbt (safe space)	156,686	57,288	(36.6)	52,419	(33.5)	34,689	(22.1)
r/ainbow (free speech)	85,121	47,595	(55.9)	43,268	(50.8)	29,449	(34.6)
Total	241,807	104,983	(43.4)	95,687	(39.6)	64,138	(26.5)

### Creating the Data Set of Crossposters

In the remaining set of 2,45,146 comments, 3,339 had an author who deleted his or her account. Therefore, I cannot know whether or not the author of the comment was a crossposter. I therefore excluded those 3,339 authorless comments for the following analysis, leaving a total of 241,807 comments in the corpus.

To identify comments as being made by a crossposter, I filtered the corpus at both the author and comment level. I identified three levels at which a comment could be identified as being authored by a crossposter, here numbered from least strict to strictest:

Level 1. Filtered at the author level. The comment’s author posted in both r/ainbow and r/lgbt *at some point* within the 14-month sample window.

Level 2. Filtered at the author level. The comment’s author posted in both r/ainbow and r/lgbt *in the same calendar month* at some point within the 14-month sample window.

Level 3. Filtered at the comment level. The comment was written in the same calendar month as another comment by the same author in the other subreddit.

At level 1, 43.4% of comments qualified as being written by a crossposter, at level 2, 39.6%, and at level 3, my strictest level, only 26.5% of comments qualified. To maximize contextual posting validity, I chose to classify comments as having been authored by a crossposter if and only if they qualified under level 3. This subset thus consists of 22.1% of comments in r/lgbt and 34.6% of comments in r/ainbow. Please refer to Tables 1 and 2 for a detailed breakdown of the number of authors and comments in each subreddit and level.

## Results

### Research Questions 1a and 1b: Comment Removal and Deletion

There were a total of 13,680 comments with the body “[deleted].” Of those, only three comments had a username attached to them, and the rest had username “[deleted].” Of the 13,680 comments deleted by the author, 9,228 were posted in r/lgbt, the safe space, and 4,452 were posted in r/ainbow, the free speech space (see Table 3). A chi-square test

indicated that the difference in proportion of deleted comments is significant,  $\chi^2(1) = 5.52$  and  $p = .019$ .

A total of 16,580 comments had the body text body “[removed].” Of those, only three comments had a username attached to them, the rest had username “[deleted],” indicating a ban by moderators. Of the 16,580 comments removed by moderators, 14,910 were posted in r/lgbt, the safe space, and 1,670 were posted in r/ainbow, the free speech space (see Table 4). A chi-square test indicated that

**Table 3.** Comment Deletion Rate in Subreddits.

Subreddit	Comments deleted by author	Total comments	Comments deleted (%)
r/lgbt (safe space)	9,228	183,093	5.0
r/ainbow (free speech)	4,452	92,313	4.8

**Table 4.** Comment Removal Rate in Subreddits.

Subreddit	Comments removed by moderator	Total comments	Comments removed (%)
r/lgbt (safe Space)	14,910	1,83,093	8.1
r/ainbow (free Speech)	1,670	92,313	1.8

**Table 5.** Summary Statistics of Deletion and Removal Rate in Subreddits.

	r/lgbt (safe space)	r/ainbow (free speech)	$\chi^2$ test	$p$
Comments deleted by author	9,228 (5.0%)	4,452 (4.8%)	$\chi^2(1) = 5.52$	.019**
Comments removed by moderator	14,910 (8.1%)	1,670 (1.8%)	$\chi^2(1) = 3,937.6$	<.001***

the difference in proportion of removed comments is significant,  $\chi^2(1) = 3,937.6$  and  $p < .001$ .

In summary, I found significant differences in self-deletion and moderator removal of comments between the two subreddits, as summarized below in Table 5.

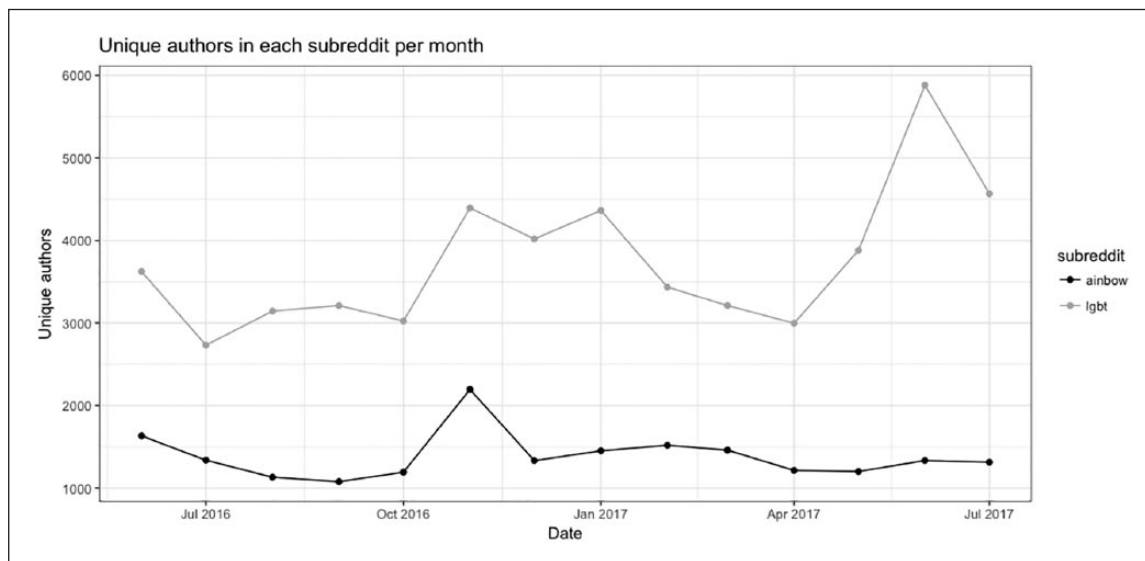
### Research Question 2: Participation

To examine participation, I first counted the number of unique authors per month in each subreddit from my comment corpus.

As illustrated in Figure 1, the number of unique authors spiked in both subreddits in November 2016, when the American presidential election took place. The number of unique authors also spiked in the free speech space r/lgbt in January 2017, when President Trump was sworn into office, and in June 2017, gay pride month in the United States.

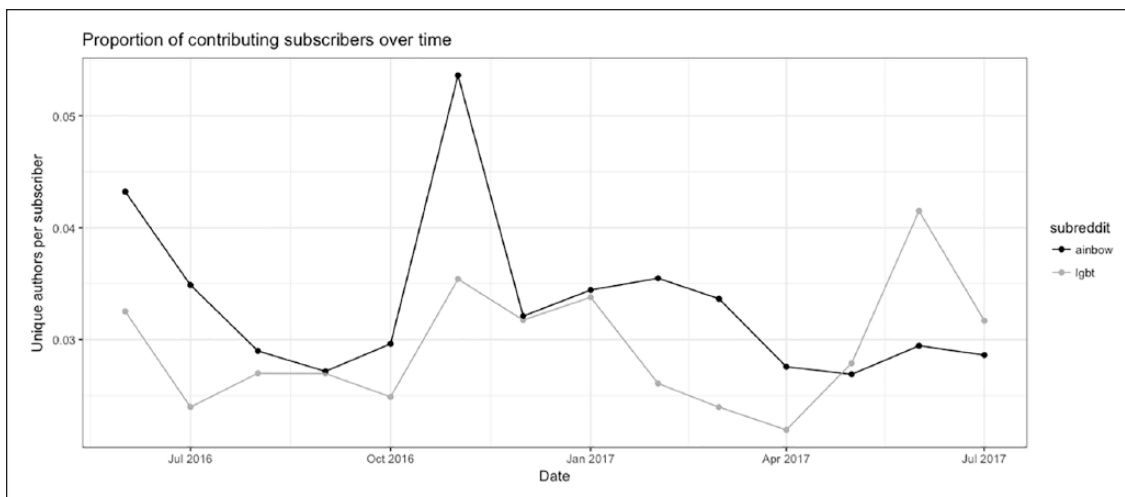
To track subscribers in the subreddits over time, I used the Internet Archive’s Wayback Machine. In my 14-month date range, I found 50 archived versions of the safe space r/lgbt and 28 archived versions of the free speech space r/ainbow. Each archived page displayed the number of subscribers to the subreddit on the day on which the page was archived.

From those data of dates and subscribers, I performed simple linear regressions to model the rate at which each subreddit gained subscribers, predicting the number of subscribers over time. Both communities grew at almost



**Figure 1.** Number of unique authors in each subreddit per month June 2016–July 2017.





**Figure 2.** Proportion of subscribers who authored comments June 2016–July 2017.

perfectly linear rates with significant regression results. In the safe space *r/lgbt*,  $F(1, 48)=2332$ , with an  $R^2=0.98$ ,  $p<.001$ . In the free speech space *r/ainbow*,  $F(1, 26)=10,330$  with an  $R^2=0.998$ ,  $p<.001$ .

I then estimated the number of subscribers for each subreddit using the first day of each month to represent the estimated subscribers for that month. (That is, the estimate for subscribers in June is based on the number of subscribers on 1 June.) I then divided the number of unique authors in that month by the number of subscribers to see what proportion of subscribers contributed a comment. The relative participation in each subreddit is graphed in Figure 2. Averaged by month over my 14-month window, 2.92% of subscribers commented in safe space *r/lgbt* and 3.33% of subscribers commented in free speech space *r/ainbow*. Overall, I found no significant difference in authorship rates between the two subreddits,  $t(23.9)=-1.65$ ,  $p=.11$ .

### Research Question 3: Language Use Across Subreddits

I used LIWC2015 to compare linguistic measures between comments in the two subreddits over my 14-month sample. LIWC2015 classifies the percentage of words in each text file that fit into predetermined and validated categories (Pennebaker et al., 2015). Each comment body was processed as a separate file. As the category percentages would be especially high for short comments, I excluded all comments with a word count less than 10 from the data set (following Ireland & Iserman, 2018).

Like many online spaces, very few users are responsible for a large percentage of comments (Sun, Rau, & Ma, 2014; Van Mierlo, 2014). In this sample, the top 1% most prolific comment authors created more than one-third of the total comments collected (34.5%), and the top 10% wrote two-thirds of the total comments (66.8%).

Thus, to create a normal distribution of comments and prevent the possibility that this small percentage of users was overly skewing the results, I removed comments by the top 9.44% most prolific posters from the data set, including comments by authors who deleted their accounts, leaving only named users who posted 10 or fewer comments over the 14-month sample ( $N=30,734$  authors).

To analyze the data, I constructed linear mixed models fit by REML for each linguistic dimension to minimize error arising from individual differences (Barr et al., 2013) using the *lme4* package for *R* (Bates, Maechler, Bolker, & Walker, 2015). In each model the estimate for the LIWC measure ( $Y_{si}$ ) was calculated using *subreddit* as a fixed effect ( $\beta_0$ ) and *comment author* as a random effect ( $S_{0s}$ ), thus allowing intercepts to vary by author. I constructed a model that would allow slopes to vary as well, but the model did not converge for all data. The simpler model, below, was therefore used for all estimates in Tables 6 and 7. The *t*-test results were calculated using Satterthwaite's method

$$Y_{si} = \beta_0 + S_{0s} + \beta_1 X_i + e_{si} \text{ (equation from Barr et al., 2013)}$$

### Hypothesis 4: Language Use Among Crossposters

The same statistical tests were then run using only the “cross-poster” subset of the data analyzed above.

## Discussion

First, I found significant differences in moderator deletion rates between the two subreddits, with the safe space subreddit, *r/lgbt*, demonstrating a higher rate of both post removal by moderators and users deleting their own comments. This result affirms that moderators effectively implemented differing moderation policies in my sample.

**Table 6.** Comments by Lower 90% of All Non-deleted Authors.

LIVC measure	r/lgbt <sup>a</sup> (safe space)	r/rainbow <sup>b</sup> (free speech)	t	p
Summary variables				
Word count	52.47	54.65	2.78	**
Words/sentence	15.09	14.73	3.79	***
Linguistic dimensions				
1st person singular	4.34	3.62	-13.67	***
1st person plural	0.59	0.69	5.46	***
2nd person	2.57	2.35	-5.30	***
3rd person singular	1.25	1.27	0.59	
3rd person plural	1.15	1.17	1.10	
Impersonal pronouns	6.98	6.67	-6.02	***
Articles	5.84	6.06	4.97	***
Prepositions	11.78	11.91	2.31	*
Auxiliary verbs	10.88	10.64	-4.21	***
Common adverbs	5.91	5.77	-1.34	
Conjunctions	6.27	6.03	-5.50	***
Negations	2.66	2.75	2.65	**
Affective processes				
Overall affect	6.96	6.74	-3.81	***
Positive emotion	4.25	3.64	-12.01	***
Negative emotion	2.54	2.92	10.20	***
Anxiety	0.35	0.36	0.54	
Anger	1.06	1.30	9.56	***
Sadness	0.35	0.34	-0.46	
Personal concerns				
Work	1.56	1.77	6.78	***
Leisure	0.82	0.78	-1.98	*
Home	0.23	0.23	-0.20	
Money	0.43	0.48	3.21	**
Religion	0.62	0.64	1.16	
Death	0.17	0.21	4.80	***
Informal language				
Swear words	0.52	0.63	6.05	***
Netspeak	0.97	0.90	-2.59	**
Assent	0.46	0.44	-1.32	
Nonfluencies	0.28	0.26	-1.84	
Fillers	0.05	0.04	-2.19	*

LIVC: Linguistic Inquiry and Word Count.

<sup>a</sup>N=49,393 comments.<sup>b</sup>N=14,356 comments.

\*\*\*p&lt;.001; \*\*p&lt;.01; \*p&lt;.05.

I operationalized spiral of silence effects slightly differently than they have been treated in the literature, a distinction captured by the difference between expressive and withdrawal behaviors as defined in Chen (2018). Generally, spiral of silence predicts a limit of expressive behaviors, that is, the decision to limit future posting. However, my data only allows me to observe withdrawal behaviors, that is, deleting comments that have already been posted.

The results described above suggest a greater spiral of silence in the safe space subreddit, as we see a significantly greater proportion of comment deletion. However, this must be tested experimentally, as the spiral of silence may also be

affected by individual differences (Gearhart & Zhang, 2018) and social norms (Neubaum & Krämer, 2018), and we have no way to know how individuals self-sorted into these subreddits. The difference in deletion rate may be an effect of the population of the safe space having a relatively higher willingness to self-censor, or it may be that the social sanctions in the safe space for minority opinions are relatively more severe than in the free speech space. In either case, the relationship between ideology and self-censorship should be further investigated.

I did not find any significant difference in participation rates between the two subreddits. This result is interesting,

**Table 7.** Comments by Crossposters in Lower 90% of All Non-deleted Authors.

LIWC measure	r/lgbt <sup>a</sup> (safe space)	r/ainbow <sup>b</sup> (free speech)	t	p
Summary variables				
Word count	51.19	52.49	0.89	
Words/sentence	15.51	15.64	0.66	
Linguistic dimensions				
1st person singular	3.93	3.54	-3.80	***
1st person plural	0.59	0.72	3.26	**
2nd person	2.32	2.19	-1.51	
3rd person singular	1.16	1.33	2.45	**
3rd person plural	1.20	1.19	-0.25	
Impersonal pronouns	6.89	6.63	-2.47	*
Articles	5.94	6.09	1.52	
Prepositions	11.80	11.97	1.41	
Auxiliary verbs	10.73	10.60	-1.07	
Common adverbs	5.91	5.77	-1.34	
Conjunctions	6.22	6.01	-2.14	*
Negations	2.65	2.79	2.01	*
Affective processes				
Overall affect	6.51	6.69	1.45	
Positive emotion	3.95	3.73	-2.21	*
Negative emotion	2.50	2.89	4.93	***
Anxiety	0.31	0.37	2.24	*
Anger	1.09	1.31	4.27	***
Sadness	0.32	0.31	-0.30	
Personal concerns				
Work	1.63	1.77	0.03	*
Leisure	0.87	0.78	-1.74	
Home	0.25	0.25	-0.31	
Money	0.41	0.46	1.70	
Religion	0.61	0.61	0.17	
Death	0.19	0.21	0.96	
Informal language				
Swear words	0.53	0.61	0.05	*
Netspeak	0.99	0.94	-0.87	
Assent	0.43	0.46	0.75	
Nonfluencies	0.27	0.26	-0.44	
Fillers	0.04	0.04	-0.34	

LIWC: Linguistic Inquiry and Word Count.

<sup>a</sup>N = 5,135 comments.<sup>b</sup>N = 4,176 comments.

\*\*\*p &lt; .001; \*\*p &lt; .01; \*p &lt; .05.

but ultimately inconclusive in this context, as I theorized that participation might be driven by inherent inclusiveness of safe spaces or salience of group identification. Either mechanism, or both, might be at play here. In the future, studies that carefully control for safe space moderation policy as well as salience of group identity may be able to find a more specific mechanism. Participation may also be more carefully studied over shorter periods of time for more granularity, which might reveal more meaningful results. The participation rates were not found to have a main effect, but rather spiked at various times. This could be further studied to understand what kind of comments

were being made in each community at the times of these spikes.

The statistical models revealed many significant differences in language use between r/lgbt and r/ainbow. First, there were many differences along linguistic dimensions. Users in the safe space used more first-person singular pronouns (I or me), second person pronouns (you), impersonal pronouns, auxiliary verbs, and conjunctions. In the free speech space, users used more words in every comment, more words per sentence, more first-person plural pronouns (we or us), more articles, more prepositions, and more negations. There were also differences in affect. In the safe space,

users used more overall affect, driven by a higher use of positive emotion words. In the free speech space, authors used more negative emotion words and more words indicating anger. The range of personal topics discussed differed between the two spaces. In the safe space, users discussed leisure relatively more frequently, whereas in the free speech space, users were more likely to discuss work, money, and death. Finally, there were also differences in informal language. Users in the safe space were relatively more likely to use netspeak and fillers, while in the free speech space, users used more swear words.

After identifying these differences, I isolated the comments of crossposters to identify which linguistic tendencies were driven by space rather than individual differences. Among the crossposters, most of the language difference persisted. Crossposters in the safe space were more likely to use first-person singular pronouns, impersonal pronouns, and conjunctions, while in the free speech space they were more likely to use first-person plural pronouns, third person singular pronouns, and negations. Crossposters were also more likely to use positive emotion words in the safe space, but in the free speech space use negative emotion words, anxiety-related words, and anger-related words. They were also more likely to discuss work and use swear words in the free speech space.

Pronoun use has been linked to understanding of one's position in social hierarchies, with lower-status individuals using "I" much more frequently, and high-status individuals using "we" more frequently (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2014). Difference in pronoun usage between these subreddits may reflect normative positions about the proper relationship of the individual to the group at large. For example, in safe spaces, in accordance with processes of relational negotiation as discussed above, individuals are encouraged to not make assumptions about groups of other people; speaker gaze, therefore, should focus more on oneself rather than the group at large, leading to more use of the singular first-person pronoun. Correspondingly, there is no such norm in free speech spaces, and individuals are given free rein to express how they believe wide swaths of people and groups behave, leading to more use of the plural first-person pronoun.

Altogether, these results suggest that implemented moderation policies were able to effectively set norms around style, affect, and topic. In our sample of thousands of infrequent posters, we found prominent trends. Generally, language in the safe space is more positive and discussions are more about leisure activities. Language in the free speech space is relatively negative and angry, and material personal concerns of work, money, and death are more frequently discussed.

One limitation of this study is that it lacks access to the specific mechanics and practices of the moderators of these subreddits, and instead the moderation practices have been summed up in broad, ideological terms.

If online spaces are indeed the future of democratic discussion, then this research suggests that moderation policies, on both ideological and practical grounds, should be a featured issue of inquiry for their role in shaping discussion. This applies not only to volunteer moderation on sites like Reddit but also to otherwise opaque or invisible corporate moderation which takes place at a broad scale. If moderation does indeed shape and constrain public discourse, as these results suggest, then the ways in which moderation policies shape and constrain public discourse for the purposes of a democracy at a much larger scale are indeed worth investigating further.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### References

- About. (n.d.). Reddit. Available from <http://www.redditinc.com>
- Ananny, M. (2018). *Networked press freedom: Creating infrastructures for a public right to hear*. Cambridge, MA: MIT Press.
- Anderson, M. (2018, September). *A majority of teens have experienced some form of cyberbullying*. Washington, DC: Pew Research Center.
- Ash, T. G. (2016). *Free speech: Ten principles for a connected world*. New Haven, CT: Yale University Press.
- Aumayr, E., & Hayes, C. (2017). Separating the wheat from the chaff: Evaluating success determinants for online Q&A communities. In *ICWSM* (pp. 476–479). Retrieved from [https://www.insight-centre.org/sites/default/files/publications/eaumayr\\_icwsm2017.pdf](https://www.insight-centre.org/sites/default/files/publications/eaumayr_icwsm2017.pdf)
- Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.
- Barrett, B. J. (2010). Is "safety" dangerous? A critical examination of the classroom as safe space. *Canadian Journal for the Scholarship of Teaching and Learning*, *1*(1), 9.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01
- Bergstrom, K. (2011). "Don't feed the troll": Shutting down debate about community expectations on Reddit.com. *First Monday*, *16*(8). Retrieved from <https://firstmonday.org/article/view/3498/3029>
- Boostrom, R. (1998). The student as moral agent. *Journal of Moral Education*, *27*, 179–190.
- Brunton, F. (2013). *Spam: A shadow history of the Internet*. Cambridge, MA: MIT Press.
- Chen, H. (2018). Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors. *New Media & Society*, *20*, 3917–3936.

- Clark-Parsons, R. (2018). Building a digital girl army: The cultivation of feminist safe spaces online. *New Media & Society, 20*, 2125–2144.
- De Koster, W., & Houtman, D. (2008). “Stormfront is like a second home to me” On virtual community formation by right-wing extremists. *Information, Communication & Society, 11*, 1155–1176.
- Dibbell, J. (1993, December 23). A rape in cyberspace: How an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *The Village Voice*. Retrieved from <https://www.villagevoice.com/2005/10/18/a-rape-in-cyberspace/>
- Dragojevic, M., Gasiorek, J., & Giles, H. (2015). Communication accommodation theory. In *The international encyclopedia of interpersonal communication* (pp. 1–21). Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118540190.wbeic006>
- Dylko, I. B., Beam, M. A., Landreville, K. D., & Geidner, N. (2012). Filtering 2008 US presidential election news on YouTube by elites and nonelites: An examination of the democratizing potential of the internet. *New Media & Society, 14*, 832–849.
- Every man is responsible for his own soul. (2014, September 6). Redditblog. Retrieved from <https://redditblog.com/#!/every-man-is-responsible-for-his-own-soul/20140906>
- Fiss, O. (2009). *The irony of free speech*. Cambridge, MA: Harvard University Press.
- Foucault, M. (1978). *The history of sexuality*. New York, NY: Pantheon Books.
- frankenmine. (2015). A list of SJW-compromised subreddits and viable alternatives for them. *Reddit*.
- Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text, 25/26*, 56–80.
- Freeman, J. (1972). The tyranny of structurelessness. *Berkeley Journal of Sociology, 17*, 151–164.
- Gearhart, S., & Zhang, W. (2018). Same spiral, different day? Testing the spiral of silence across issue types. *Communication Research, 45*, 34–54.
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT: Yale University Press.
- Glynn, C. J., Hayes, A. F., & Shanahan, J. (1997). Perceived support for one’s opinions and willingness to speak out: A meta-analysis of survey studies on the “spiral of silence.” *Public Opinion Quarterly, 61*, 452–463.
- Gonzales, A. L., & Hancock, J. T. (2008). Identity shift in computer-mediated environments. *Media Psychology, 11*, 167–185.
- Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research, 37*, 3–19.
- Good Bot, Bad Bot. Retrieved from [https://web.archive.org/web/20170926011216/https://goodbot-badbot.herokuapp.com/all\\_filter](https://web.archive.org/web/20170926011216/https://goodbot-badbot.herokuapp.com/all_filter)
- Gray, K. L. (2012). Intersecting oppressions and online communities. *Information, Communication & Society, 15*, 411–428. doi:10.1080/1369118X.2011.642401
- Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology, 17*(1), Article 2. Retrieved from <http://digitalcommons.law.yale.edu/yjolt/vol17/iss1/2>
- Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Cambridge, MA: MIT Press.
- Hampton, K. N., Rainie, L., Lu, W., Dwyer, M., Shin, I., & Purcell, K. (2014). *Social media and the “spiral of silence.”* Washington, DC: Pew Research Center.
- Harris, M. (2015, November 11). What’s a ‘safe space’? A look at the phrase’s 50-year history. *Splinter News*. Retrieved from <https://splinternews.com/what-s-a-safe-space-a-look-at-the-phrases-50-year-hi-1793852786>
- Hawbaker, K. T. (2018, December 6). #MeToo: A timeline of events. *Chicago Tribune*. Retrieved from <https://www.chicagotribune.com/lifestyles/ct-me-too-timeline-20171208.html>
- Hayes, A. F., Glynn, C. J., & Shanahan, J. (2005). Willingness to self-censor: A construct and measurement tool for public opinion research. *International Journal of Public Opinion Research, 17*, 298–323.
- Hayes, A. F., Uldall, B. R., & Glynn, C. J. (2010). Validating the willingness to Self-Censor Scale II: Inhibition of opinion expression in a conversational setting. *Communication Methods and Measures, 4*, 256–272.
- Herring, S. C. (1996). Gender and democracy in computer-mediated communication. In R. Kling (Ed.), *Computerization and controversy: Value conflicts and social choices* (2nd ed., pp. 476–489). San Diego, CA: Academic Press.
- Hinds, D., & Lee, R. M. (2008, January 7–10). Social network structure as a critical success condition for virtual communities. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)* (p. 323). New York, NY: IEEE.
- Ireland, M. E., & Iserman, M. (2018). Within and between-person differences in language used across anxiety support and neutral Reddit communities. In K. Loveys, K. Niederhoffer, E. Prud’hommeaux, R. Resnik, & P. Resnik (Eds.), *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 182–193). New Orleans, LA: Association for Computational Linguistics.
- Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology, 99*, 549–571.
- Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science, 22*(1), 39–44.
- Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology, 33*, 125–143.
- Kenney, M. (2001). *Mapping gay LA: The intersection of place and politics*. Philadelphia, PA: Temple University Press.
- Lamont, T. (2014, February 7). Reddit: How to win the internet. *The Guardian*. Retrieved from <https://www.theguardian.com>
- Lin, H., Fan, W., Wallace, L., & Zhang, Z. (2007, January). An empirical study of web-based knowledge community success. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences* (pp. 1530–1605). New York, NY: IEEE.

- Liu, Y., Rui, J. R., & Cui, X. (2017). Are people willing to share their political opinions on Facebook? Exploring roles of self-presentational concern in spiral of silence. *Computers in Human Behavior, 76*, 294–302.
- Lotan, G., Graeff, E., Ananny, M., Gaffney, D., & Pearce, I. (2011). The Arab Spring: The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication, 5*, 31.
- Lukianoff, G., & Haidt, J. (2015). The coddling of the American mind. *The Atlantic, 316*(2), 42–52.
- Manne, K. (2015, September 19). Why I use trigger warnings. *The New York Times, Opinion*. Retrieved from <https://www.nytimes.com/2015/09/20/opinion/sunday/why-i-use-trigger-warnings.html>
- Marwick, A. E. (2017). Are there limits to online free speech? *Data & Society: Points*. Retrieved from <https://datasociety.net/output/are-there-limits-to-online-free-speech/>
- Marwick, A. E., & Lewis, R. (2017). *Media manipulation and disinformation online*. New York, NY: Data & Society Research Institute.
- Massanari, A. L., & Chess, S. (2018). Attack of the 50-foot social justice warrior: The discursive construction of SJW memes as the monstrous feminine. *Feminist Media Studies, 18*, 525–542.
- Matias, J. N. (2016). *The civic labor of online moderators*. *Internet Politics and Policy Conference*, Oxford, UK, 22–23 September.
- Matthes, J., Morrison, K. R., & Schemer, C. (2010). A spiral of silence for some: Attitude certainty and the expression of political minority opinions. *Communication Research, 37*, 774–800.
- McCluskey, M., & Hmielowski, J. (2012). Opinion expression during social conflict: Comparing online reader comments and letters to the editor. *Journalism, 13*, 303–319.
- McDevitt, M., Kioussis, S., & Wahl-Jorgensen, K. (2003). Spiral of moderation: Opinion expression in computer-mediated communication. *International Journal of Public Opinion Research, 15*, 454–470.
- McKee, R. (2015, November 11). College safe spaces COLOR. *The Augusta Chronicle*. Retrieved from <http://caglecartoons.com/viewimage.asp?ID={85355D74-4534-4A5A-BE23-B87458AF34CC}>
- Meyer, H. K., & Speakman, B. (2016). Quieting the commenters: The spiral of silence's persistent effect on online news forums. *ISOJ, 6*(1), 51.
- Neubauer, G., & Krämer, N. C. (2018). What do we fear? Expected sanctions for expressing minority opinions in offline and online communication. *Communication Research, 45*, 139–164.
- Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*, 337–360.
- Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication, 24*(2), 43–51.
- Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society, 4*(1), 9–27.
- Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society, 6*, 259–283.
- Pazzanese, C., & Walsh, C. (2017, December 21). The women's revolt: Why now, and where to. *The Harvard Gazette*. Retrieved from <https://diversity.harvard.edu/news/women%E2%80%99s-revolt-why-now-and-where>
- Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin: University of Texas at Austin.
- Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*, 271–282.
- Pfaffenberger, B. (1996). "If i want it, it's ok": Usenet and the (outer) limits of free speech. *The Information Society, 12*, 365–386.
- Porten-Che , P., & Eilders, C. (2015). Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate. *Studies in Communication Sciences, 15*, 143–150.
- Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. *Communication Theory, 16*, 47–74.
- Reagle, J. (2013). "Free as in sexist?" Free culture and the gender gap. *First Monday, 18*(1). Retrieved from <https://firstmonday.org/article/view/4291/3381>
- @realDonaldTrump. (2017, July 26). After consultation with my Generals and military experts, please be advised that the United States Government will not accept or allow..... [Twitter Post]. Retrieved from <https://twitter.com/realdonaldtrump/status/890193981585444864>
- Reddit.com Traffic Statistics. (n.d.). Alexa. Retrieved from [www.alexa.com/siteinfo/reddit.com](http://www.alexa.com/siteinfo/reddit.com)
- RedditList. (n.d.). Retrieved from [www.redditlist.com](http://www.redditlist.com)
- @RichardDawkins. (2015, October 24). A university is not a "safe space." If you need a safe space, leave, go home, hug your teddy & suck your thumb until ready for university [Twitter Post]. Retrieved from <https://twitter.com/richarddawkins/status/658022567085801472>
- Roestone Collective. (2014). Safe space: Towards a reconceptualization. *Antipode, 46*, 1346–1365.
- Scheufle, D. A., & Moy, P. (2000). Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International Journal of Public Opinion Research, 12*(1), 3–28.
- Schulz, A., & Roessler, P. (2012). The spiral of silence and the Internet: Selection of online content and the perception of the public opinion climate in computer-mediated communication environments. *International Journal of Public Opinion Research, 24*, 346–367.
- Scissors, L. E., Gill, A. J., Geraghty, K., & Gergle, D. (2009, April 4-9). In *CMC I trust: The role of similarity*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA.
- Scissors, L. E., Gill, A. J., & Gergle, D. (2008, November 8-12). *Linguistic mimicry and trust in text-based CMC*. Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, San Diego, CA.
- Shelton, M., Lo, K., & Nardi, B. (2015, June 8-11). *Online media forums as separate social lives: A qualitative study of disclosure within and beyond Reddit*. Iconference proceedings, Newport Beach, CA.
- Stengel, B. S., & Weems, L. (2010). Questioning safe space: An introduction. *Studies in Philosophy and Education, 29*, 505–507.

- Stoycheff, E. (2016). Under surveillance: Examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism & Mass Communication Quarterly*, 93, 296–311.
- Sun, N., Rau, P., & Ma, L. (2014). Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, 38, 110–117.
- Taub, A., & Fisher, M. (2018, April 21). Where countries are tinderboxes and Facebook is a match. *New York Times*. Retrieved from <https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html>
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24–54.
- Turner, F. (2006). How digital technology found utopian ideology: Lessons from the first hackers' conference. In D. Silver, A. Massanari, & S. Jones (Eds.), *Critical cyberculture studies* (pp. 257–269). New York: New York University Press.
- Van Gelder, L. (1985, October). The strange case of the electronic lover. *Ms. Magazine*. Retrieved from <http://lindsayvangelder.com/clips/strange-case-electronic-lover>
- Van Mierlo, T. (2014). The 1% rule in four digital health social networks: An observational study. *Journal of Medical Internet Research*, 16(2). Retrieved from <https://www.jmir.org/2014/2/e33/>
- Wagner, C., Liu, L., Schneider, C., Prasarnphanich, P., & Chen, H. (2009, June). Creating a successful professional virtual community: A sustainable digital ecosystem for idea sharing. In *Digital Ecosystems and Technologies* (pp. 163–167). New York, NY: IEEE.
- Wike, R. (2016, October). *Americans more tolerant of offensive speech than others in the world*. Washington, DC: Pew Research Center.
- Wilson, C., & Dunn, A. (2011). The Arab Spring: Digital media in the Egyptian revolution: Descriptive analysis from the Tahrir data set. *International Journal of Communication*, 5, 25.

### Author Biography

Anna Gibson is a PhD candidate in the Communication Department at Stanford University. Her research interests include digital media and labor.