Communication Law Review
Volume 17, Issue 1

Censoring Hate Speech in U.S. Social Media Content: Understanding the User's Perspective

Caitlin Carlson, Ph.D., Seattle University[*]

Tweets and Facebook posts containing racist or misogynistic slurs are a common part of the social media landscape in the United States. According to a Pew Research Internet Project survey, almost half of black social media users and one-third of female users in the United States said they frequently saw offensive images or humor on social networking sites (Pew Research Center, 2014). The pervasive nature of hate speech in social media content has received the attention of scholars, regulators, and journalists. Even social media organizations such as Twitter have recognized their own shortcomings in policing content. In July 2016, Twitter CEO Jack Dorsey responded publically to the ongoing harassment of actress and comedian Leslie Jones by deleting the account of Milo Yiannopoulos, who organized an online harassment campaign against Jones. In addition to the barrage of racist comments and memes directed at Jones, her account was also hacked and private photos were posted (Silman, 2016). Dorsey said in a statement that abusive behavior like this was not permitted under Twitter's hateful conduct policy (Silman, 2016). "We rely on people to report this type of behavior to us but we are continuing to invest heavily in improving our tools and enforcement systems to prevent this kind of abuse. We realize we still have a lot of work in front of us before Twitter is where it should be on how we handle these issues," said Dorsey (Silman, 2016). Unfortunately, absent any substantial changes on the part of the company, Dorsey's statement is nothing more than lip service. Also troubling is the extent to which he places the responsibility to identify this abuse onto users, rather than insisting the organization take the initiative to limit hate speech on its own platform.

Although each social media organization has a different approach to dealing with offensive content, all seemingly recognize the complex nature of the task at hand.

[*] *Dr. Caitlin Carlson is an Assistant Professor in the Department of Communication at Seattle University. Dr. Carlson may be reached at carlso42@seattleu.edu.*

The terms of service agreements users must electronically sign before they may access a particular platform absolve social media companies of most legal responsibility to protect hate speech, except that which can be characterized as a direct threat or harassment, both of which are prohibited by U.S. Federal Statutes (18 U.S.C. § 875(c), 47 U.S. Code § 223).

Despite the lack of substantial legal prohibitions against hate speech in the United States, many social media organizations, such as Facebook, work to minimize hate speech on their platforms by establishing policies prohibiting this content (Facebook, 2017). Posts that violate the established rules against hate speech on these sites can be immediately blocked by an algorithm designed to catch rule violations or flagged by users, evaluated by employees, and removed if deemed inappropriate (Rosen, 2013). Other organizations, such as Twitter, allow extreme language or images to be included in tweets unless they directly and specifically threaten another user (Twitter, 2017).

While it is likely that social media organizations have conducted their own market research about the business implications of hate speech regulation, little is known about users' expectations of social media companies to protect or police expression. This focus group study seeks to better understand online hate speech by asking users about their expectations of social media organizations to protect freedom of expression. This article will begin with a review of the legal prohibitions and protections for hate speech in the United States. A brief overview of the hate speech policies of major social media organizations, including Facebook, Twitter, Instagram and YouTube will then be presented and the theoretical arguments in favor of and against censoring hate speech will be explored. Next, the research question will be proposed and the focus group methodology explained. Findings from the interviews will be presented and discussed. The article will conclude with a call for social media organizations to increase their efforts to remove hate speech, but to do so in a more transparent way.

Literature Review

In the United States, most hate speech is protected by the First Amendment. For example, in *Snyder v. Phelps* (2011), the U.S. Supreme Court held that picketing fallen soldier's funerals with signs that said "god hates fags" did not meet the threshold for intentional infliction of emotional distress. Chief Justice John Roberts wrote for the majority that, "because this Nation has chosen to protect even hurtful speech on public issues to ensure that public debate is not stifled, Westboro must be shielded from tort liability for its picketing in this case," (*Snyder v. Phelps*, 2011). Unless expression falls into the categories of fighting words, incitement to illegal advocacy, true threats, or the rarely invoked notion of group libel, it is considered protected. Here, hate speech will be defined as expression that seeks to promote, spread or justify misogyny, racism, anti-Semitism, religious bigotry, homophobia, bigotry against the disabled (Foxman & Wolf, 2013).

Legal Protections and Prohibitions for Hate Speech in the United States

Fighting words were defined by the Supreme Court in *Chaplinsky v. New Hampshire* (1942) as "those personally abusive epithets which, when addressed to the ordinary citizen, are, as a matter of common knowledge, inherently likely to provoke a violent reaction." An example of hate speech that might be considered fighting words would be a racial or gendered insult uttered about someone's mother. However, since the *Chaplinsky* decision, the Supreme Court has been reluctant to find much expression that falls within the narrow definition of this standard. For example, in *R.A.V. v. St. Paul*, a 1992 case dealing with the constitutionality of a cross burning statute, the Supreme Court said that what makes fighting words unprotected are their non-speech elements. Like regulating a sound truck, it is the noise or verbal cacophony caused by fighting words that the Supreme Court said warrants regulation (*R.A.V. v. St. Paul*, 1992). However, the government may not regulate the use of fighting words based on hostility or favoritism toward the underlying message because the First Amendment imposes a viewpoint discrimination limitation (*R.A.V. v. St. Paul*, 1992). Because the Minnesota ordinance applied only to fighting words that insult on the basis of race, color, creed, religion or gender and not to all fighting words, the Supreme Court

found it to be unconstitutional. In addition, this case established the doctrine that said content-based regulations must meet the threshold of strict scrutiny, the highest level of judicial review, which requires the government to demonstrate a compelling interest and for the regulation to be narrowly drawn. Thus, it is possible that content-neutral bans on fighting words may be used to limit hate speech on social media because of the "noise" it creates in that online environment.

The next category that hate speech may fall into is incitement to illegal advocacy. In the landmark case, *Brandenburg v. Ohio* (1969), which reversed the conviction of a Ku Klux Klan leader who had been convicted under Ohio's criminal syndicalism statute for advocating violent political and industrial reform, the U.S. Supreme Court said that the state may not limit advocacy unless it is "directed to inciting or producing imminent lawless actions and is likely to produces such action" (*Brandenberg v. Ohio*, 1969). The Supreme Court underscored the importance of the imminence requirement in the *Brandenburg* test when it said in *Hess v. Indiana* (1973) that the state could not punish a protestor who said, "we'll take to the streets" because it was mere advocacy of future unlawfulness and was not directed at anyone and was not likely to create immediate danger. The imminence standard in the *Brandenberg* test is what makes it difficult to apply this precedent to cases involving incitement online.

In addition to fighting words and incitement, some hate speech may be illegal if it rises to the level of true threats. Unlike incitement to illegal advocacy, the notion of true threats does not include an imminence requirement and therefore may be more suited to guiding regulations designed for an online environment. True threats are defined by the Supreme Court as threats that "encompass those statements where the speaker means to communicate a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals" (*Virginia v. Black*, 2003). In addition to the common law formulation, in the late 1990s Congress also adopted a law that articulates the elements of a true threat (18 U.S.C. § 875(c)).

In order to support a conviction under the true threats statute, the government must prove that there has been a transmission in interstate or foreign commerce of a communication containing a threat to injure or kidnap the person of another (18

U.S.C. § 875(c)). Cases regarding a violation of this statute often fail to meet the threshold of a direct threat to injure or kidnap an individual. In their application of this law, courts have considered the extent to which the threat is believable—an essential determinant of whether or not an utterance or expression should be considered a true threat. For example, the Sixth Circuit's decision in *United States v. Alkhabaz* (1997) held that emails exchanged between two parties presumably containing sexually threatening content about a mutual acquaintance did not rise to the level of true threats. The content of the emails did not meet the threshold required for true threats because "no reasonable person would perceive the e-mails as intending to effect change or achieve a goal through intimidation" (*United States v. Alkhabaz*, 1997).

Recently, a man serving 44 months in jail for being convicted of communicating a true threat to his ex-wife via Facebook appealed the decision of the Third Circuit, which held that if a statement causes a reasonable person to fear for her safety, that is a true threat (*Elonis v. U.S.*, 2015). Anthony Elonis' Facebook posts, the Third Circuit said, met the legal definition of a threat. Among the many other comments, Elonis posted the following to his ex-wife's Facebook page:

> "There's one way to love you but a thousand ways to kill you. I'm not going to rest until your body is a mess, soaked in blood and dying from all the little cuts. Hurry up and die, bitch, so I can bust this nut all over your corpse from atop your shallow grave. I used to be a nice guy but then you became a slut. Guess it's not your fault you liked your daddy raped you. So hurry up and die, bitch, so I can forgive you" (*U.S. v. Elonis*, 2013).

The Supreme Court reversed and remanded the Third Circuit's decision, stating that their instruction in the case to require only negligence with respect to the communication of a threat, is not sufficient to support a conviction under 18 U.S.C. § 875(c). Here, the Supreme Court had the opportunity to clarify how the true threats doctrine applies online and failed to do so. Legal scholar Joseph Russomano (2017) argues this failure should be construed as neglect of duty, "the Court chose not to address and clarify the true threats doctrine, an area of law marked by murky definitions and haphazard application. Moreover, the Court declined to specify what

required level of intent a threat is required, and from whose perspective the speech must be perceived to be a threat, for criminal conviction" (p.3). Instead, the Court ceded control to those who see the internet as "a virtual Wild West" free from regulation (Russomano, 2017, p. 3).

Group libel or group defamation is the fourth and final legal category that certain hate speech may fall into. As Jeremy Waldron noted in *The Harm in Hate Speech* (2012), many countries, including Germany, Norway, and Denmark, use the term "group libel" instead of "hate speech" because of the difficulty associated with the task of determining what is dislike and what is hate. In the United States, the Supreme Court once chose not to overturn a fine imposed on the president of the White Circle league of America on First Amendment grounds, labeling his distribution of a leaflet on Chicago street corners urging people to "protect the white race from being mongrelized" as criminal libel (*Beauharnais v. Illinois*, 1952). However, the notion of group libel has largely been abandoned in the United States, where many states require the plaintiff to prove identification in defamation cases.

Finally, it is possible that hate speech may serve as evidence in a hate crime case. Hate crime laws work by increasing the criminal sentence if the prosecution can show that the perpetrator selected a victim based on their perceived race, nationality, religion, gender, or sexual orientation (Foxman & Wolf, 2013). In the United States, 45 states and the District of Columbia all currently have some form of hate crime law on the books (Foxman & Wolf, 2013). However, hate speech alone with no criminal conduct may not be subject to hate crime laws. Unless hate speech falls into one of the categories reviewed here, including fighting words, incitement to illegal advocacy, true threats, or group libel, it will most likely be protected by the First Amendment.

This approach differs greatly from Canada, South Africa, and European Union member countries, all of which prohibit the use of hate speech in-person and online. For example, the Canadian Criminal Code contains a cause of action against the public incitement of others to hatred. It says:

> "Every one who, by communicating statements in any public place, incites hatred against any identifiable group where such incitement is likely to lead to a breach of the peace is guilty of (a) an indictable offence and is liable to imprisonment for a term not exceeding two years; or *(b)* an offence punishable on summary conviction." (Canada Criminal Code, R.S.C. 1985, c. C-46, § 319(1)).

Recently Germany has begun to take a more assertive approach to enforcing the country's rules regarding hate speech online. The country recently passed a law that gives social media companies 24 hours to remove posts that obviously violate German law and have been reported by other users or be subject to hefty fines of up to 50 million euros (Knigge, 2017). As international corporations, social media organizations must navigate laws in a variety of countries. Perhaps if users are largely in favor of removal there would be an economic incentive for these organizations to adopt a more aggressive approach to removing hate speech across their platforms worldwide.

Social Media Hate Speech Policies

Currently, there are as many approaches to dealing with offensive content as there are social media platforms. Although the First Amendment protects most hate speech in in the United States (*Snyder v. Phelps*, 2011) social media organizations based in the United States are not required to permit hate speech on their sites. The terms of service users sign give social media organizations the legal authority to regulate hate speech on their sites however they like. Commercial ISPs and Social Media Organizations may voluntarily agree to prohibit users from sending racist or bigoted messages over their services (Foxman & Wolf, 2013). Such prohibitions "do not implicate First Amendment rights because they are entered into through private contracts and do not involve government action in some way" (Foxman & Wolf, 2013, p. 187). Therefore, these companies are able to decide how, when, and why they will remove content and can simply update the terms of service accordingly. This suggests that it may be possible to incentivize these companies to do more to regulate hate speech on their platforms.

Today, many social media organizations, including Facebook, have community guidelines that strictly prohibit hate speech. However, reporting the use of offensive terms like "Wetback" does not yield removal so confusion still exists regarding what counts as hate speech and whether it is removed. According to Facebook's current community standards, "organizations and people dedicated to promoting hatred against these protected groups are not allowed a presence on Facebook" (Facebook, 2017). Twitter, on the other hand, will only remove specific threats and abuse (Twitter, 2017). Offensive content is readily permitted on the site. Twitter's terms of service specifically state that, "Users are allowed to post content, including potentially inflammatory content, provided they do not violate the established rules" of the site (2017). As long as it is not a direct threat to an individual, which is prohibited under U.S. Federal Statute (18 U.S.C. § 875(c)), Twitter does not take issue with offensive tweets, images, or video (Twitter, 2017).

Instagram, which is owned by Facebook says in the long version of its community guidelines that users should share only photos they have taken themselves, post photos that are appropriate for a diverse audience, foster meaningful interactions, follow the law, respect other members of the community, maintain a supportive environment particularly when it comes to self harm or injury, and finally, to be thoughtful when posting newsworthy events (Instagram, 2017). Despite these rules, a cursory search of Instagram shows that there are still several visible posts that contain the "N-word" and other derogatory terms. YouTube, which is owned by Google, warns users in its community guidelines "not to cross the line," when it comes to hateful content. The policy says that the company "does not support content that, promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line" (YouTube, 2017). Finally, text based sites such as Reddit and 4chan tend to take a much more hands-off approach to regulating hate speech on their sites (Knigge, 2017).

Admittedly, managing the process of removing reported posts, images, and videos is a complex task, particularly given the difficulty in determining what content is satirical or humorous and which contains actual harassment or threats. The shifting, subjective nature of what hate speech is and how it is defined in places across the globe makes its regulation and removal extremely challenging. This is reflected in the ever-changing nature of social media companies' terms of service, which often reference the specific community guidelines presented here. The terms of service users must agree to before accessing social media platforms come in two forms, "browsewrap" and "clickwrap" (Fradette, 2014). Browsewrap contracts are passive forms displaying and asserting the terms of service via a link. That page will indicate that the user agrees to the terms simply by using the site. Clickwrap agreements, on the other hand, "embody more classical aspects of contract formation" at least in form, but not in substance (Fradette, 2014, p. 958). Here, users must actively click "agree" to indicate their willingness to adhere to the stated rules. Although it should be noted that very few users actually read the terms or follow the terms link on the website (Fradette, 2014). Still, courts have indicated that failure to read the terms does not absolve the user from being bound by those terms. In *Fteja v. Facebook, Inc.* (2012) the Court noted that social media users are savvy enough to know what a hyperlink labeled "Terms of Service" will contain and that failure to click on the hyperlink and read the terms simply means that the user has failed to inform herself of the contract obligations.

Clearly, social media organizations have a substantial amount of discretion regarding what they consider hate speech and how aggressively they pursue the process of identifying and removing hate speech from their platforms. Perhaps if these companies can be convinced that users are in favor of removing hateful content, they may be more likely to do so. However, before considering what users think about this issue, it is necessary to understand the arguments legal communication scholars most regularly cite as the reason for protecting or prohibiting hateful expression.

Why Remove Hate Speech from Social Media?

The primary reason for banning hate speech in social media content is the psychological harm this content inflicts on its victims. Critical race theorist Mari Matsuda described this tendency for members of the defamed group to internalize the message and come to believe in their own inferiority as the most damaging impact of hate speech (1993). In addition, it is likely that the continued, unchecked presence of hate speech in social media may actually have a silencing effect on minorities, women, and other groups targeted by this speech. Rather than encouraging these individuals to participate in public discourse, hate speech may limit the amount of content generated by minorities and women by creating an environment that dissuades these individuals from speaking out against the various forms of discrimination they face (Delgado & Stefancic, 1997). For example, a woman may be less likely to continue to participate in a debate on Facebook about campus sexual assault after being called a "slut" by another commenter. According to legal scholars Danielle Citron and Helen Norton (2011), hate speech works to limit civic engagement online, thus curtailing the process of what they call digital citizenship (2011). In other words, the virtual environment created by hate speech dissuades the political speech of people of color, women, and members of the lesbian, gay, bisexual and transgender (LGBT) community.

Perhaps most importantly, the widespread adoption of racial and ethnic slurs has historically been associated with acts of violence, ranging from hate crimes targeted at individuals to mass genocide. Alexander Tsesis (2002) has conceptualized how the casual use of racial slurs can create a climate that will tolerate crimes against humanity such as slavery or the Holocaust. According to Tsesis, language plays a critical role in developing the psychological mechanisms necessary for engaging in racist conduct (Tsesis, 2002). Hate speech then, provides the "vocabulary and grammar depicting a common enemy . . . [and establishes] a mutual interest in trying to rid society of the designated pest" (Tsesis, 2002, p. 198). Throughout history, hate speech has been used to dehumanize various religious, ethnic, or racial groups in

order to make military action or physical violence against them more palatable (Tsesis, 2002).

Why Allow Hate Speech on Social Media?

A desire to encourage the free and open marketplace of ideas is perhaps the most frequently cited justification for allowing harmful or degrading speech (Weinstein, 1999). Truth, the theory holds, is made stronger and clearer when it has the opportunity to collide with all ideas, even erroneous ones (Weinstein, 1999).

Alexander Meiklejohn (1948) argued that all expression on political matters must be permitted in order to successfully execute the social contract of democratic self-governance. According to Meiklejohn, citizens of a democracy should be free to praise, criticize, or discuss all political and social issues. Legal theorist Thomas Emerson (1966) asserted that any attempts to curtail expression infringe upon our personal liberty. Those in favor of protecting hate speech say that efforts to censor this content will only send it further into the shadows, where it would be even more dangerous (Delgado & Stefancic, 1997). Social media then, provides racists, bigots, and others with a virtual "safety valve" where they can vent their anger through expression, as opposed to violent action. James Weinstein (1999) suggests that removing the outlet for hate speech could also lead to an increase in violence against women and minorities because online aggressions would be replaced with offline ones.

Social media users may also be opposed to efforts to prohibit hate speech because they believe the process of content removal gives too much control to social media organizations (Pew Research Center, 2015). Traditionally, public concern has focused on government censorship of information. Today, the unchecked power of social media organizations to regulate content seems a far greater threat to free expression than any form of government intervention.

While there are valid theoretical arguments on both sides of the debate about whether hate speech should be censored on social media platforms, little information exists about user perspectives on this issue. Public opinion data from the Pew Research Center indicates that 67 percent of Americans think people should be able to say things in public that are offensive to minorities (Pew Research Center, 2015).

However, when Millennials (those 30 and under) were asked the same question, 40 percent were in favor of preventing the public use of statements offensive to minorities (Pew Research Center, 2015). This focus group study seeks to understand the nuance in these perspectives by better understanding people's opinions about whether and how social media organizations should remove hate speech and why. Specifically, the research question this study seeks to answer is:

> RQ1*: Do users think social media organizations should remove hate speech from their platforms? If so, why? If not, why not?*

Method

Given how little information is known about why users do or do not support the removal of hate speech from various social media platforms, a focus group methodology was selected. This exploratory approach allows respondents to organically shape how researchers think about hate speech online, rather than having researchers predetermine the questions to be considered and analyzed (Morgan, 1997). It is my hope that this research will help guide future quantitative inquiries about the impact of hate speech on individuals and on public discourse.

Data Collection

To address the research question, six focus groups interviews (n=39) were conducted at a small Jesuit university in the Pacific Northwest in the fall of 2014. A total of 23 females and 16 males participated in the study. Each focus group had between 6 and 7 participants, all of whom were college students or their friends. Participants ranged in age from 18-30. Regrettably, participants were not asked to provide their ethnicity, which is a substantial limitation of this study. The fact that the college these students attend is in the Pacific Northwest may also be a limitation here given the regional trend toward liberal perspectives. Finally, the young age of study participants is also a potential limitation here as millennials tend to be in favor of greater speech restriction (Pew Research Center, 2015).

All of the participants reported that they were active social media users. Usage amounts ranged from infrequent use (once per week) to daily and even hourly use of

social media applications. A convenience sample was used to solicit the participation and a free meal was provided as an incentive. Validity for the interview guide was established by conducting two pilot focus group interviews. Each of the six focus group interviews lasted over one hour. The conversations were recorded and the audio was transcribed in full.

Data Transformation and Analysis

A transcript-based analysis was used to understand the responses and identify themes in the data. As recommended by Kruger and Casey (2000), the analysis was systemic, verifiable, sequential and continuous, as is the best practice. By gauging the frequency, specificity, emotion, and extensiveness of various comments, themes emerged (Krueger, 1998).

Findings

Most participants in this focus group study thought social media organizations should remove some or all hate speech from their platforms.[1] Participant's responses fell clearly into three groups: Those who did not think social media organizations should regulate content, those who favored the removal of certain hate speech that threatened or harassed others, and those who favored the removal of all hate speech from social media content. The justifications given for these positions will provide important insights about how people rationalize their positions. To begin, the respondents definition of and experiences with hate speech will be explored. Next, the reasons participants provided for why social media organizations should or should not remove hate speech from their platforms are presented and their implications discussed.

Defining Hate Speech

After discussing the frequency of their own social media use, participants were asked to define hate speech in their own words. Overwhelmingly, the responses indicated that most participants thought that hate speech meant slurs. Specifically, Ingrid (FG2) referred to hate speech as "*slurs against women or different ethnic groups.*"

[1] Pseudonyms are used for the names of focus group participants to protect their anonymity.

Thomas (FG5) said that hate speech included, "*things that like disparage someone who is unlike you and you want to make them feel bad because of that physical or like physiological difference.*"

In addition to racist and homophobic epithets, participants said hate speech also included threats or harassing posts. As Sean (FG3) noted, "*Hate speech is speech in a manner that insinuates, somebody means to harm. So threats fall under that category. Threats to your health, threats to your reputations and then um, I think that's under the law but I don't really know.*"

Not only did participants relate hate speech to threats, some participants also associated hate speech with historical oppression or mass genocide. Specifically, Ian (FG3) said, "*I think of 1930s Nazi Germany.*" Participants also mentioned that hate speech was "like cyberbullying" but different in that the latter involved actual harassment between two individuals or a small group. While most participants described hate speech as the casual or even satirical use of racial or homophobic slurs, cyberbullying was considered a far more personal phenomenon, one that was targeted at a specific individual.

Encountering Hate Speech on Social Media Platforms

When asked whether or not they had seen posts on social media that met the definition of hate speech established by the group, almost all participants reported that they had seen slurs or images that they would consider hate speech. The comments section of YouTube videos was mentioned by several participants as a virtual space that was overrun with hate speech. Often times, participants attributed this to the fact that users on that site were permitted to remain anonymous:

> Arianna (FG5): "*YouTube is like the scum of the earth.*"
> Haley (FG5): "*Yeah cause people don't have to put their picture or their name. It's kind of anonymous. Cause you can be gamergirl3400 and write crazy things.*"

Several participants also identified Twitter as a platform where hate speech was often present. In addition, participants noted that there are several Instagram accounts dedicated solely to posting racist or misogynistic content. Not only did users seem clear on the relationship between anonymity and hate speech, but the discussion among them suggested that they thought platforms that allowed users to maintain

their anonymity were more likely to contain offensive content. Only one participant, a female, reported being the victim of online harassment herself.

Participants were also keenly aware that the organizations and individuals they choose to follow had perhaps the greatest impact on the amount of hate speech they might be exposed to. In fact, many participants said that they avoided exposure to hate speech simply by unfriending or unfollowing the offending party. Finally, while YouTube and Twitter were mentioned most often by participants, Facebook, Instagram, Tumblr, Vine and YikYak were also cited as platforms where participants had encountered hate speech.

Justifications for Censoring or Protecting Hate Speech

Responses to questions about whether social media companies should remove hate speech from their platforms revealed that most participants were in favor of some form of censorship. While a few participants advocated for a completely hands-off approach to content removal, most thought social media companies should remove some or all hate speech from their sites. This is in line with public opinion data, which indicates that young Americans are more inclined than their older counterparts to favor limiting free expression to protect individuals (Pew Research Center, 2015).

Do not Regulate or Remove Hate Speech

Participants who were opposed to having social media organizations remove hate speech from their platforms said that it was more dangerous to have these companies acting as arbiters of speech, than it was for users to be exposed to that content. In fact, many of these responses reflected the marketplace of ideas theory. Participants expressed concern about their ability to access accurate information.

> Jordan (FG1): "*I think as soon as these websites start engaging in like censorship we are going to start monitoring hate speech and preventing these things. It actually makes me not trust them as much, because now you don't know what they will block, because they are not going to keep a log and tell you what they are deleting.*"
> Lauren (FG6): "*As humans we have to have the discernment capabilities to filter out what's right and not right. We don't need someone else to censor stuff for us.*"

Participants also said that engaging in the process of content removal represented a "slippery slope" for social media organizations. Moreover, participants noted that the

volume of content, combined with its subjective nature, would make it difficult to regulate. As Elysse (FG5) said, "*It would be hard for social media companies to regulate all hate speech because there are always little things like people will get offended and they report, report, report.*" Thus, concerns about an unfettered marketplace of ideas and the difficulty presented by the process of content removal were the primary reasons participants said they were against social media organizations efforts to regulate or remove hate speech from their sites. The role the free exchange of ideas plays in ensuring an effective democracy was not mentioned by any of the participants.

Remove Only Hate Speech that Contains Threats or Harassments

Some users thought that social media companies should allow most hate speech but step in to remove those posts that could be characterized as direct threats or harassment.

> Kelly (FG1): "*I just don't think that unless it's directly affecting someone or threatening their safety or the safety of a group of people, unless it's directly targeting someone. If it's a general statement about race or gender, I don't think you can really regulate that or stop people from saying that.*"
> Chris (FG1): "*Yeah I agree with that, because if it was something minor and the company would lose traction and lose followers.*"
> Moderator: "*So you think the company should err more on the side of allowing this speech?*"
> Chris (FG1): "*Yeah allowing the speech but if it is a threat, such as 'I am going to kill you' then they should investigate.*"
> Alex (FG1): "*I think they should minimize harassments- like lets say someone keeps messaging you and you block them and they make another account to keep messaging you then I think Facebook can step in, but if that person has a swastika as their profile picture, you can control what you see and you don't have to go on that person's page. I think they should let that stay, because you are not, I mean you are kind of targeting people kind of with that, but you are not going out of your way to hurt people I guess if that makes any sense.*"

This exchange highlights the distinction participants made between direct threats targeted at individuals and unfavorable comments made more generally about a group of people, such as women or members of the LGBT community. These participants also acknowledged that social media organizations had their own reputations to consider in this decision-making process. It was therefore likely, participants said, that these organizations would make decisions based on their

business interests, rather than any perceived ethical responsibility to protect free expression.

Remove All Hate Speech from Social Media Content

Participants in favor of removing all hate speech cited the negative impact it has on those targeted as the primary justification for removal. Participants noted that the terms of service, which users sign and agree to before they are able to access a platform, make clear what kind of content is and is not permitted. Adherence to these rules was seen by many of the participants as the price of admission to be able to use these free applications.

> Elysse (FG5): "*Yes we have freedom of speech but social media companies, I'm pretty sure it's in the terms and conditions, the stuff that we never read, I'm sure something about hate speech is in there and they can ban people. That's not violating their freedom of speech that's violating their terms and conditions.*"

Finally, among participants in favor of removing all hate speech from social media platforms there was a fairly extensive debate about how social media organizations should manage this process. Some participants, such as Chloe (FG3) expressed a clear preference for user-driven reporting, "*They [social media companies] should give users the tools we need to report that stuff.*" Others, such as Thomas (FG5), favored a software-based automatic removal program, "*Well at the same time you can put it into programs that can detect those words every time they're used.*"

Discussion

The primary question this focus group study sought to answer was, "Do social media users think social media organizations should remove hate speech from their platforms?" Responses fell into three unique categories: Allow all hate speech on social media, remove only threatening or harassing hate speech targeting individuals, and remove all hate speech. Most participants felt that social media organizations had some obligation to police hate speech on their platform in an effort to mitigate the impact those comments have on members of the groups targeted by this content, such as people of color, women, and members of the LGBT Community.

Participants who were in favor of allowing hate speech in social media content

were adamant that they did not want social media companies making decisions about what information they would or would not have access to, particularly because these organizations are susceptible to financial pressures. Their responses highlight a desire for an unfettered marketplace of ideas. Instead of the government censoring speech, users were concerned with the virtually unchecked power social media companies currently have to promote or eliminate all kinds of posts, images, or videos from their sites with no outside checks and balances, let alone reporting mechanisms.

Another issue raised by participants was the extent to which the job of reporting offensive content is left to users. Social media organizations put users in the driver's seat when it comes to censorship, requiring them to flag offensive content before the company can take any action. Given the extent to which social media companies rely on users to report offensive content, it is no surprise that hate speech continues to be pervasive on many of these platforms. Users are not trained or compensated to play such a large role in managing the complex process of content removal. As a result, hate speech proliferates on social media. For example, in 2013 one-half of black social media users and one-third of female users said they *frequently* saw offensive images or humor on social networking sites (Pew Research Center, 2013).

Participants who said hate speech should be permitted on social media did not see this content as important political speech. Many First Amendment scholars, including Alexander Meiklejohn (1948) and C. Edwin Baker (2012), consider controversial speech about matters of political importance an essential part of the process of democratic self-governance. However, participants in this study did not use the word democracy even once. Instead of the U.S. Bill of Rights, participants referenced a social media organization's terms of service as the ultimate rulebook governing content removal decisions. Participants seemed unfamiliar with the historical connection between democracy and free expression in the United States.

Respondents who advocated for the removal of all hate speech from social media platforms cited the harm it causes its victims as the primary reason for censorship. Participants who called for this approach were concerned about the

psychological damage this content can have on members of the targeted groups. This position echoed the arguments of critical race theorist, Mari Matsuda (1993), who said that the most detrimental impact of the widespread use of hate speech is the negative impact it has on a victim's self-esteem. Members of the groups targeted come to believe in their own social, political, or physical inferiority. The preference by these respondents to minimize the psychological damage caused by hate speech tracks with existing public opinion data, which indicates that approximately 40 percent of millennials are okay with limiting speech offensive to minorities (Pew Research Center, 2015).

In discussing their own experiences with hate speech in social media, users recognized a link between anonymity and the proliferation of hateful or offensive content. Users said they were most likely to encounter hate speech on sites that permitted anonymity. Thus, one of the easiest paths Twitter and other platforms could take to reducing hate speech is to require real names and identities to be used. This would likely decrease the volume of harassing and hateful comments like those leveled at comedian Leslie Jones on Twitter.

While the majority of participants thought some or all hate speech should be banned from social media platforms, they recognized the difficulty in determining where the line between harmful and meaningful expression should be drawn. What was surprising was the extent to which participants saw the responsibility to draw that line, by drafting and enforcing content regulations against hate speech, as solely the job of social media organizations. Often referencing the terms of service, participants said that social media organizations should have nearly absolute power to regulate content as they saw fit. Participants rightly saw social media platforms as private virtual spaces owned by for-profit organizations that are well within their rights to create the rules that apply in those spaces. If respondents wished to avoid hate speech on a particular social media platform, they said it was their responsibility to unfriend, unfollow, or report any accounts generating the offensive content. This perspective underestimates the power users have to publically pressure these organizations to make changes in their community guidelines and to increase the amount of resources

dedicated to minimizing hate speech in social media content.

Conclusion

There is no easy answer to dealing with the problem of hate speech in social media content. Most participants in this study favored the prohibition and removal of some or all hate speech because of the psychological damage it has on members of the targeted groups, like people of color and women. Users were generally unfazed by the extent to which social media organizations relied on them to report and remove racist or homophobic slurs from their feeds. However, they continued to be frustrated by the amount of hate speech in social media content, particularly on those sites that allow users to remain anonymous. Those in favor of permitting hate speech in social media content said that they did not want publically traded, financially motivated companies to have unchecked power to regulate expression.

While there are justifiable concerns about unchecked censorship by social media organizations, the overwhelming preference reported by these participants was for social media companies to increase their efforts to remove hate speech from their sites and I second that call. One easy way for all social media platforms to do this would be to prohibit anonymity and strengthen the terms of service to more strictly prohibit racist and misogynistic slurs, along with threats of violence. In addition, social media organizations could track and publish descriptions of content removed under the hate speech sections of their community guidelines. This would provide much needed transparency in the content removal process. Finally, social media organizations can and should dedicate additional resources to actively identifying and removing reported content. For example, the role artificial intelligence technologies could play in this process should be explored extensively, along with other potential solutions that do not require the unpaid labor of users to be successful. Finally, future research should work to quantify the user perspectives identified in this exploratory study in order to incentivize social media organizations to do more to address this growing problem.

References

18 U.S.Code § 875 (c): Interstate Communications.
47 U.S. Code § 223: Telecommunications.
Baker, C. E. (2012). Hate speech. In M. Herz & P. Molnar (Eds.). *The content and context of hate speech* (57-62). Cambridge, UK: Cambridge University Press.
Beauharnais v. Illinois, 343 U.S. 250, 253–254 (1952).
Brandenberg v. Ohio, 395 U.S. 444, 447 (1969).
Canada Criminal Code, R.S.C. 1985, c. C-46, § 319(1).
Chaplinsky v. New Hampshire, 315 U.S. 568, 571-72 (1942).
Citron, D. & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review, 91*, 1435-1484.
Delgado, R. & Stefancic, J. (1997). *Must we defend Nazis? Hate speech, pornography and the new First Amendment.* New York, NY: New York University Press.
Elonis v. United States, 135 S.Ct. 2001 (2015).
Emerson, T. I. (1966). *Toward a general theory of the First Amendment*. New York, NY: Vintage.
Facebook. (2017). *Facebook community standards*. Retrieved from https://www.facebook.com/communitystandards
Fteja v. Facebook, Inc., 841 F. Supp. 2d 829 (S.D.N.Y. 2012).
Flores, A., & James, C. (2013). Morality and ethics behind the screen: Young people's perspectives on digital life. *New Media & Society*, *15*(6), 834-852.
Foxman, A.H. & Wolf, C. (2013). *Viral hate: Containing its spread on the internet.* New York, NY: Palgrave Macmillan.
Fradette, J.E. (2014). Online terms of service: A shield for First Amendment scrutiny of government action. *Notre Dame Law Review*, *89*(2), 947-984.
Harris, C., Rowbotham, J., & Stevenson, K. (2009). Truth, law and hate in the virtual marketplace of ideas: Perspectives on the regulation of Internet content. *Information & Communications Technology Law*, *18*(2), 155-184.
Hern, A. (2015, February 5). Twitter CEO: We suck at dealing with trolls and abuse, T*he Guardian*. Retrieved from http://www.theguardian.com/technology/2015/feb/05/twitter-ceo-we-suck-dealing-with-trolls-abuse
Hess v. Indiana, 414 U.S. 105 (1973).
Instagram. (2017). *Community guidelines.* Retrieved from https://help.instagram.com/477434105621119
Knigge, M. (2017, July 5). Hate speech curb should look beyond Facebook, Twitter, *Deutsche Welle*. Retrieved from http://www.dw.com/en/hate-speech-curb-should-look-beyond-facebook-twitter/a-39550114
Krueger, R.A., & Casey, M. A. (2000). *Focus groups*. Thousand Oaks, CA: Sage.
Krueger, R.A. (1998). *Analyzing and reporting focus group results*. Thousand Oaks, CA: Sage.
Matsuda, M.J. (1993). *Words that wound.* Boulder, CO: Westview Press.
Meiklejohn, A. (1948). *Political freedom.* New York, NY: Harper Brothers.
Morgan, D. (1997). *Focus groups as qualitative research* (2nd ed.). Thousand Oaks, CA: Sage Publications.
Pew Research Center. (2014, Oct. 22). *Online harassment survey.* Retrieved from http://assets.pewresearch.org/wpcontent/uploads/sites/14/2014/10/PI_OnlineHarassment_72815.pdf
Pew Research Center. (2015). *Spring global attitudes survey* [Data file]. Retrieved from http://www.pewglobal.org/datasets/2015/
R.A.V. v. City of St. Paul, 505 U.S. 377 (1992).
Rosen, J. (2013, April 29). The delete squad: Google, Twitter, Facebook and the new global battle over the future of free speech, *New Republic*. Retrieved from

http://www.newrepublic.com/article/ 113045/free-speech-internet-silicon-valley-making-rules#

Russomanno, J. (2016). Facebook threats: The missed opportunity of *Elonis v. United States*. *Communication Law & Policy*, *21*(1), 1-37.

Silman, A. (2016, Aug. 24). A timeline of Leslie Jones's horrific online abuse, *NY Mag*. Retrieved from http://nymag.com/thecut/2016/08/a-timeline-of-leslie-joness-horrific-online-abuse.html

Slagle, M. (2009). An ethical exploration of free expression and the problem of hate speech. *Journal of Mass Media Ethics*, 24, 238-250.

Snyder v. Phelps, 131 S. Ct. 1207, 1219-1220 (2011).

Tsesis, A. (2002). Prohibiting incitement on the Internet. *Virginia Journal of Law and Technology*, *5*, 1–40.

Twitter. (2017). *Twitter terms of service*. Retrieved from https://twitter.com/tos

United States v. Alkhabaz, 104 F.3d 1492 (6th Cir. 1997).

United States v. Elonis, 730 F.3d 32 (3rd Cir. 2013).

Waldron, J. (2012). *The harm in hate speech.* Cambridge, MA: Harvard University Press.

Weinstein, J. (1999). *Hate speech, pornography and the radical attack on free speech doctrine*. Boulder, CO: Westview Press.

YouTube. (2016). *Community guidelines*. Retrieved from https://www.youtube.com/yt/policyandsafety/communityguidelines.html