Defining Hate Speech
By Andrew F. Sellars
Berkman Klein Center
December 2016
https://cyber.harvard.edu/publications/2016/DefiningHateSpeech

Sellars is overviewing the various attempts to define hate speech by academics, legal experts, and online platforms. In lieu of defining hate speech he identifies 8 emerging themes and common traits that are used in defining hate speech:

1. Targeting of a Group, or Individual as a Member of a Group
2. Content in the Message that Expresses Hatred
3. The Speech Causes a Harm
4. The Speaker Intends Harm or Bad Activity
5. The Speech Incites Bad Actions Beyond the Speech Itself
6. The Speech is Either Public or Directed at a Member of the Group
7. The Context Makes Violent Response Possible
8. The Speech Has No Redeeming Purpose

Sellars points out that there are divergent theoretical fields that deal with hate speech: free speech theory and critical race theory, but there is little overlap in these two realms.

Can we know it when we see it

Justice Stewart famously asserted that "I know it when I see it" when referring to identifying obscenity. It seems to be the consensus that this approach is not applicable to identifying hate speech due to the variety of forms of speech and contexts which one could identify as hate speech.

There are instances where specific epithets or insults are used and an outsider or scholar may see hate speech but the speaker/recipient do not. Henry Louis Gates Jr. for this reason asserted that we should not "spend more time worrying about speech codes than coded speech." Since a lot of hate speech can be coded or masked by symbols.

The discussion of Stewart's "I know it when I see it" and context points to a central difficulty in defining hate speech since it requires assessing the subjectivity and intention of both the perpetrator and the victim. However, only some definitions include the component of intention on the part of the perpetrator, and definitions also vary on how they define harm to the victim.

There is also the question of whose speech is worth identifying as hate speech since there are many instances of people identifying others' hate speech on social media having their content flagged and removed since it is falsely identified as hate speech itself. But it also raises a more complex question of the role of the press/media in amplifying hate speech which Lori Tharps raises in her article *"Reprint Reporting" and Race* (see my summary of that article), since drawing attention to hate speech can make people think it is more common and acceptable than it might actually be, resulting in greater hate speech, and repeating the hate speech—whether critical or not—can still harm the targeted group.

Academic Attempts

Academic definitions vary as some are trying to formulate a definition that can be applied to legal sanction whereas others simply want to define hate speech in order to better understand the phenomenon.

Richard Delgado's influential "Words that Wound" he focuses on racist hate speech and crafts a definition of hate speech for tort law regulation. His definition avoids criteria for content and focuses on intent, impact, and objective perception by a "reasonable person":

1. that "[language was addressed to him or her by the defendant that was intended to demean through reference to race;"
2. "that the plaintiff understood as intended to demean through reference to race; and"
3. "That a reasonable person would recognize as a racial insult."

Mari J Matsuda's definition draws on her structural analysis of law and inequality. Her definition pays more attention to content (racial inferiority) of the hate speech than Delgado's definition:

1. the message is "of racial inferiority;"
2. the message is "directed against a historically oppressed group;" and treat all members as alike and inferior
3. the message is "prosecutorial, hateful, and degrading," which Matsuda later clarifies has an intent-element within it.

Calvin Massey is establishing a theoretical definition rather than a prosecutable one and looks at the outcomes and effects of the speech more than the intentions of the speaker. By avoiding intention/subjectivity of the speaker, Massey's definition includes both polite and vulgar racists:

> "hate speech is any form of speech that produces the harms which advocates for suppression ascribe to hate speech: loss of self-esteem, economic and social subordination, physical and mental stress, silencing of the victim, and effective exclusion from the political arena." (16)

Mayo Moran's definition defines hate speech as ""speech that is intended to promote hatred against traditionally disadvantaged groups" (16). Her definition differs from others as it points to intention to promote (rather than intention to incite) hatred and she avoids defining the content of the speech and explicitly extends her definition to include coded speech.

One of the debates around the definition of hate speech is related to the "marketplace of ideas" theory that holds that more speech is a good in itself and that allowing the maximum amount of speech allows for the survival/uncovering of truth. Some people like Kenneth Ward incorporate into their definition the criteria that hate speech has "no redeeming purpose" and that the "attacks are so virulent that an observer would have a great difficulty separating the message delivered from the attack against the victim" (17). This form of speech is deemed excludable from the "marketplace of ideas" since it contributes nothing and also can even exclude or silence the voices of victims.

Susan Benesch's research deals specifically with **dangerous speech** which is more directly linked to the incitement of mass violence. Rather than providing an exact definition, Benesch identifies five variables that help determine the severity of the dangerous speech:

1. there is a "powerful speaker with a high degree of influence;"
2. there is a receptive audience with "grievances and fear that the speaker can cultivate;"
3. a speech act "that is clearly under-stood as a call to violence;"
4. a social or historical context that is "propitious for violence, for any of a variety of reasons;"and
5. An "influential means of dissemination."" (17)

Alice Marwick and Ross Miller identify three general elements used to define hate speech:

1. Content-based element
2. Intent-based element
3. Harms-based element

Legal definitions

Article 20 of the International Covenant on Civil and Political Rights (ICCPR) says that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law" (20). The UNHCR creates the "Rabat Plan" which sets out a six part test to assess the severity of hate speech:

(1) the social and political context in which the statement is made;

(2) the position or status of the speaker in society;
(3) the specific intent to cause harm;
(4) the degree to which the content of the speech was "provocative and direct," and the "nature of the arguments deployed in the speech";
(5) the extent and reach of the speech and the size

The European Union drafted the "Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law." This is the framework used for the recent cooperative agreement reached between private online platforms and the European Commission to police hate speech on the websites' platforms.[1]
The framework defines hate speech as one of three things:
(1)"Public incitement to violence or hatred directed against a group of persons or a member of such group defined on the basis of race, [color], descent, religion or belief, or national or ethnic origin;"
(2) The same, when done through "public dissemination or distribution of tracts, pictures, or other material;"
(3)"publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity, and war crimes [as defined in EU law], when the conduct is carried out in a manner likely to incite violence or hatred against such group or a member of such group."

Online Platforms
Technological responses to online hate speech: deleting content, modifying content, blocking users, temporary bans or internal quasi-judicial resolution among users. There is a gap between "the public declaration of the rule, which has a vague or ceremonial rule, and the actual operational document hidden from public view." Youtube, twitter, and facebook all have a bifurcated structure whereby they have a formal terms of service on the one hand which rejects liability for the offensive speech of their users, and they have a less formal community standards page that goes into more detail about the sites values and a more defines what sort of speech is not tolerated.
Youtube (this language is from 2015):
"Terms of service": mentions "offensive content" and makes clear that the platform is not liable for the offensive content of its users
"Community Guidelines" : "we don't 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 bases of these core characteristics."
Twitter:
Terms of service: disclaims liability for offensive content
"Twitter Rules": prohibits abusive behavior and hateful conduct defined as "promoting violence against or directly attacking or threatening other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease"
Facebook
"Statement of Rights and Responsibilities": disclaims liability

[1] https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code
https://www.theguardian.com/media/2017/jun/30/germany-approves-plans-to-fine-social-media-firms-up-to-50m

“Community Standards”: Facebook will remove content that “directly attacks people based on their race; ethnicity; national origin; religious affiliation; sexual orientation; sex, gender, or gender identity; or serious disabilities or diseases”

Sellars identifies 8 emergent themes from reviewing various definitions of hate speech:

1. Targeting of a Group, or Individual as a Member of a Group
2. Content in the Message that Expresses Hatred
    1. Content is tricky because any form of expression can be hate speech if it's made with intent to incite hatred—this points to the tension between content and context
3. The Speech Causes a Harm
    1. Extrinsic harm/physical violence versus structural and psychic violence which includes silencing the victim
4. The Speaker Intends Harm or Bad Activity
    1. Different types of intention: intent to vilify, humiliate, incite hatred or promote hatred or promote violence etc.
5. The Speech Incites Bad Actions Beyond the Speech Itself
6. The Speech is Either Public or Directed at a Member of the Group
    1. This includes private messages sent to a victim, but the definition does not include private speech that isn’t targeted at a member of the targeted group. This means that in most definitions it is acceptable for two racists to send hate speech to each other about another group because this speech is considered by many theorists to be closer to thought than it is to speech/action.
7. The Context Makes Violent Response Possible
8. The Speech Has No Redeeming Purpose

The paper concludes by discussing the relative power differences between different groups. Asserting that promoting the power of a traditionally disadvantaged group shouldn’t be thought of as hate speech, whereas things such as men’s rights or white power often should be. Furthermore, “a victim's struggle for self-identity in response to racism” shouldn’t be considered hate speech (even though facebook has taken down posts by anti-racists).

Related readings:
https://www.article19.org/data/files/medialibrary/38231/%27Hate-Speech%27-Explained---A-Toolkit-%282015-Edition%29.pdf

https://cyber.harvard.edu/publications/2017/08/harmfulspeech

https://medium.com/berkman-klein-center/exploring-the-role-of-algorithms-in-online-harmful-speech-1b804936f279

https://cyber.harvard.edu/publications/2016/UnderstandingHarmfulSpeech

https://cdt.org/files/pdfs/Report_on_Account_Deactivation_and_Content_Removal.pdf

https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/

https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html

****

Notes on:
Counterspeech on Twitter: A Field Study, by Susan Benesch, Derek Ruths, Kelly Dillon, Haji Mohammad Saleem, and Lucas Wright
&
Considerations for Successful Counterspeech, by Susan Benesch, Derek Ruths, Kelly Dillon, Haji Mohammad Saleem, and Lucas Wright

Counter Speech can address the perpetrator of hate speech or the cyber bystanders whose discourse norms are shaped by seeing hate speech go uncontested. For hateful speakers with deeply ingrained hate, counter speech is less effective, but it can still alter their discourse to be less overtly hateful which in itself can also positively affect the cyber bystanders.

The guiding principle of counterspeech is the liberal ideal that more speech is the best remedy to harmful speech. [2] This is partially informed by the idea that if you delete hateful content from one platform that there is always somewhere else to go (I'm not sure this is totally convincing).

Recommended Strategies

- Warning of consequences
    - Remind speaker of harm done by speech
    - Remind of offline consequences and the permanence of online communication
    - Remind of online consequences (blocking, reporting, suspended account)
    - Mainly effective at getting hate speech deleted and doesn't necessarily change speakers POV
- Shaming and Labeling
    - Labeling the speech (not speaker) as bigoted, misogynist, etc
    - Helpful to cyberbystanders
    - Speaker 'may not have known'
- Empathy and Affiliation
    - Change the tone to friendly, empathetic or peaceful
    - Affiliate with speaker and establish a connection (ex: I am also a conservative, but…)
    - Affiliate with targeted group (ex: what you said hurt me as an asian…)
    - Changing the tone is more effective when the speaker affiliates with counterspeaker

---

[2] However others argue that no matter how much good speech or counter speech there is, since it does not reverse the initial harm caused by the hate speech itself against the targeted group:
"Sticks and stones can break my bones," we are taught to chant as children, "but words can never hurt me." Americans "are taught this view by about the fourth grade, and continue to absorb it through osmosis from everything around them for the rest of their lives," Catherine MacKinnon writes with no little asperity in *Only Words*, her latest and most accessible book, "to the point that those who embrace it think it is their own personal faith, their own original view, and trot it out like something learned form their own personal lives every time a problem is denominated one of 'speech,' whether it really fits or not."
– Henry Louis Gates, Jr., *War of Words: Critical Race Theory and the First Amendment* (pg. 18)

- Humor
    - Neutralize hateful speech that is seen as dangerous and intimidating
    - Attract larger audience to the counterspeech's message
    - Use humor to soften the message of counterspeech that could otherwise come off as hostile or aggressive (ex: it's time to stop posting cat)
- Images
    - Can make counterspeech more viral
    - Counterspeech is generally more effective when it is emotive rather than rational/logical so images can be a good way to "send people along emotive pathways"

Discouraged Strategies

- Hostile or aggressive tone and insults
    - Can cause backfire or speaker to dig in their heels
- Fact-checking
    - Fact checking may sway cyber bystanders, but it is unlikely to influence original speaker
    - Speaker will find a way to fit the new facts presented to the conclusions they are already committed to
    - Social psychologists call this the **backfire effect** where challenging someone's views with facts leads them to hold those views even more firmly
    - Pointing out hypocrisy can be good for bystanders but usually not for speaker
- Harassment and Silencing

Successful counterspeech is indicated by:

- Speaker shifts their discourse if not also their beliefs
- Speaker apologizes, recants, or deletes original hate speech
- Discourse norms of the cyberbystanders are positively affected
- Hate speech narratives delegitimized (even if speaker is not swayed)
- More counter speech is elicited from the audience (this is good until it turns into harassment/dogpiling)

Other frameworks

- Institute for Strategic Dialogue:
    - Erod original speakers' intellectual framework
    - Mock or ridicule
    - Highlight the negative impact of extremist speech
    - Demonstrate inconsistencies in the extremists argument
    - Question effectiveness of extremists at achieving their goals
- Anti-Defamation League
    - Responding to the original speaker
    - Using comedy or satire
    - Correcting falsehoods.

****

(((Echoes))), Exposed: The Secret Symbol Neo-Nazis Use to Target Jews Online
https://mic.com/articles/144228/echoes-exposed-the-secret-symbol-neo-nazis-use-to-target-jews-online#.0c2VP2PZ9
The three parentheses around names has become a common ways for neo-nazis and other far-right twitter users to push anti-semitic conspiracy theories such as jews running the media. It is particularly difficult form of hate speech for concerned users to detect since most search engines strip searches of punctuation and searching just "((()))" yields nothing on twitter.

****

https://www.theverge.com/2016/4/25/11503512/twitter-abuse-report-multiple-abuse
In April 2016 Twitter created the option for users to report multiple abusive tweets at once rather than having to flag each tweet individually. This reduced the clunkiness of the process of reporting harassment, and also gives moderators a better understanding of the context of abuse.

****

Alt-right trolls are using these code words for racial slurs online
https://qz.com/798305/alt-right-trolls-are-using-googles-yahoos-skittles-and-skypes-as-code-words-for-racial-slurs-on-twitter/
http://knowyourmeme.com/memes/events/operation-google

"Now a new type of hateful internet code appears to be emerging: The systematic use of innocuous words to stand in for offensive racial slurs. Search Twitter for "googles," "skypes," or "yahoos," and you will encounter some shocking results, like this tweet: "If welfare state is a given it must go towards our own who needs. No Skypes, googles, or yahoos." Or this one, reading "Chain the googles / Gas the yahoos."
What does this mean? Nothing good. In this lexicon, "googles" means the n-word; "skypes" means Jews; and "yahoo's" means "spic." The word "skittles" has come to refer to Muslims, an obvious reference to Donald Trump Jr.'s comparing of refugees with candy that "would kill you."

****

Lori Tharps 9/25/13 'Reprint reporting' and race
archives.cjr.org/minority_reports/reprint_reporting_and_race.php

Tharps piece argues that simply reprinting or pointing out racist speech online is an insufficient form of intervention, and it also makes a serious ethical claim about the role journalists can have in amplifying hateful speech in a way that continues to hurt the targeted individual/group. Instances of racism in America are not surprising or newsworthy, so articles that simply reprint racist tweets do not constitute good journalism.

Tharps argues that reprint reporting is not only lazy journalism, but it can skew the public's dialogue around events. She points to Olympic Gold medalist Gabby Douglas and Miss America Nina Davuluri as examples of people whose web presence is now largely defined by articles that amplify fringe racist tweets that were not representative of the majority's opinion of them. Listicles that generated a lot of views were accompanied by little thoughtful, intelligent commentary on racism in America and resulted in:

> “Miss America, Divuluri’s crowning moment has been forever eclipsed by negativity, and she has to spend her 15 minutes of fame talking about the morons who think she’s an Arab instead of the issues that are important to her, like STEM education and eating disorders in young women. And because reputable news outlets including The Daily Beast, *USA Today*, and *The Washington Post* ran stories based on a group of statistically insignificant, unsubstantiated, racist tweets, anyone searching in the future for how Americans reacted to Divuluri’s win would surmise that America wasn’t ready for an Indian American Miss America in 2013. This is just plain embarrassing, and it paints all Americans with broad brush strokes as intolerant racists.”

****

https://www.theringer.com/2017/1/17/16039094/curbing-terrorist-social-media-activity-facebook-twitter-google-601ff9684068

Extreme Moderation

By: Katie Knibbs

Hany Farid in 2009 created PhotoDNA which helps identify and remove child pornography from the internet (by comparing photos with those kept in a centralized database of banned images held by the National Center for Missing and Exploited Children. Farid then moved on to work with the Counter Extremism Project (CEP) to develop similar software that flags terrorist photos

CEP’s proposed software however has been criticized for being closely tied to bush administration and republican establishment, and this combined with their lack of transparency on criteria for extremist content causes many tech companies to be weary.

****

The Center for Democracy and Technology created the Digital Decision tool to help programmers throughout the phases of developing an algorithm to ensure that the digital decisions made by algorithms reflect values such as equality, democracy, and justice

https://cdt.info/ddtool/

https://cdt.org/issue/privacy-data/digital-decisions/

****

Quartz has “A running list of websites and apps that have banned, blocked, deleted, and otherwise dropped white supremacists”

https://qz.com/1055141/what-websites-and-apps-have-banned-neo-nazis-and-white-supremacists/

****

http://newsfeed.time.com/2013/05/20/the-geography-of-u-s-hate-mapped-using-twitter/

Humboldt State University has a “Hate Map” that maps out geotagged tweets that have been identified as hateful. They categorize hate speech as either homophobic, racist, or ableist, which excludes a lot of other forms of hate speech. They are only relying on a small subset of tweets by users who allow geotagging, and only using a small sample of keywords that students identified as being used negatively. The map is also skewed because it shows hate speech as

more clustered in the east coast which is mainly a result of counties being closer together on the east coast. Overall the map is not that useful.

****

Kwok, Irene, and Yuzhou Wang. "Locate the Hate: Detecting Tweets against Blacks." *AAAI*. 2013.
https://pdfs.semanticscholar.org/db55/11e90b2f4d650067ebf934294617eff81eca.pdf

This study uses the bag-of-words approach to determine if a set of tweets related to obama were racist. Their method was 76% percent accurate on average, but had no way of avoiding classifying tweets with words like black, white, or filthy when used outside of a racist context.

The main interesting finding of the study was "We compiled a hundred tweets that contained keywords or sentiments generally found in hate speech, and asked three students of different races (but of the same age and gender) to classify whether a tweet was offensive or not, and if classified as offensive, to rate how offensive it was on a scale of one through five (with five being the most offensive). The calculated percentage of overall agreement was only 33%,indicating that this classification would be even more difficult for machines to do accurately, which is consistent with previous research" (1621). This demonstrates how hard it is to distinguish between hate speech and offensive speech, especially across different identities and perspectives.

****

How Right-Wing Extremists Stalk, Dox, and Harass Their Enemies
Micah Lee
September 6 2017
https://theintercept.com/2017/09/06/how-right-wing-extremists-stalk-dox-and-harass-their-enemies/?comments=1#comments

In this article Micah Lee goes through a set of screenshots of leaked chats on the discord chat app. The specific channel he was looking at involved some 50 users seeking to dox (making available personal information such as phone number, address etc for the sake of harassment) left wing and anti-racist activists. The information that the members of the chat room collected included photographs, social media profiles, home address, phone numbers, email addresses, date of birth, driver license numbers, vehicle information, place of employment, and in one instance, a social security number.

For the most part those targeted were people who identified with “antifa” tactics. However, they also doxxed individuals deemed sympathetic to antifa or anti-racism, such as people that RSVP'd to anti-nazi protests on facebook. They also discussed targeting the group Safety Pin Box, an anti-racist group for white allies, as well as other groups such as Southern Poverty Law and left-leaning university professors and journalists. One 22 year old college student who simply had a photo of herself wearing a punch nazis shirt in her cover photo was doxxed even though she's not active in any antifa groups, and her address, social media accounts, and usernames were revealed.

One tactic employed in doxxing was sending victims a malicious link to collect their IP address. . “What happens is the person goes through our link to an actual website, and from there this website logs the IP as it redirects the person without them knowing through their IP tracking website.” Once an IP address is obtained, it can be fairly easy to estimate someone’s geographic location.

Lee argues that there is a significant difference between the far-rights doxing and that of antifa activists, since the far-right often doxes people who simply disagree with their world view, whereas “Antifa activists only target members of hate groups, a small but growing subset of American society that President Trump refuses to condemn, responsible for mounting terrorist attacks against mosques, black churches, trans women, and people of color.” Furthermore, left-wing doxxing typically seeks to get enough information to contact employers or the police, but does not generally disseminate personal information for the sake of general harassment.

****

https://www.theringer.com/tech/2017/8/22/16180026/charlottesville-politics-hate-speech-internet

The Problems With Internet Platforms Policing Hate
By: Kate Knibbs

After the violent white nationalist rally in charlottesville, a litany of websites and other internet services finally responded to calls to cease providing platforms to white nationalist hate speech. Many saw this as long overdue, as social media platforms and domain name and hosting services had pushed back against prior calls to remove clients disseminating hate speech. However, other groups such as the EFF are concerned that many of the recent decisions have been made on an ad hoc basis and are a dangerous and insufficient substitute for an actual framework for content regulation and a transparent process for removing hate groups.

This was concern reflected in Cloudflare CEO Matthew Prince’s statement on why he decided denied service to *The Daily Stormer*, where he emphasizes the importance of his decision to stem the *Stormer*’s hate speech, but that moving forward it is important for their to be an established framework that will guide these sorts of decisions in the future. Otherwise “In a not-so-distant future, if we're not there already, it may be that if you're going to put content on the Internet you'll need to use a company with a giant network like Cloudflare, Google, Microsoft, Facebook, Amazon, or Alibaba...Without a clear framework as a guide for content regulation, a small number of companies will largely determine what can and cannot be online.” The danger of leaving decisions up to whims of these large companies is that groups fighting hate speech such as Black Lives Matter could be classified as hate groups as happened to the NAACP in the civil rights era.

Right now companies are functioning along the “i know it when i see it” rule or when there is enough public pressure, as there was recently due to the violent white nationalist rally in charlottesville

****

Exploring the Role of Algorithms in Online Harmful Speech
https://medium.com/berkman-klein-center/exploring-the-role-of-algorithms-in-online-harmful-speech-1b804936f279

This article summarizes some of the themes discussed at the berkman klein's conference on the role of algorithms in online harmful speech. These are quotes that i found interesting:
"Amanda Lenhart, a senior research scientist at the Associated Press-NORC Center for Public Affairs Research, said that the problem of harmful speech online is pervasive; in a survey she helped conduct while at another organization, Data & Society, 47 percent of Americans over 15 said they had experienced at least one instance of online harassment, and 72 percent said they had witnessed online harassment or abuse. (A more recent survey produced similar findings.) Lenhart also found substantial differences by gender: while men and women were equally likely to have experienced some form of harassment, women generally faced a wider variety of online abuse, including more serious violations such as sexual harassment and long-term harassment, while men were more likely to experience abusive name-calling or physical threats. Men and women also experienced harassment differently, with women being almost three times as likely to say that an experience of harassment made them feel scared. These different experiences of harassment underscore the importance of considering diverse perspectives when defining and addressing harassment online."

"Aarti Shahani, a technology reporter at NPR, described a case in which Facebook decided to take down a photo of a noose accompanied by a sign saying (we delete the racial slur here) "[n-word] swing set." The photo was initially flagged by users as violating Facebook's terms of service. But content moderators didn't take it down, because it didn't clearly depict a human victim. After it was flagged a second time, the content was eventually removed. The reason was the use of the "n-word." The implication was that the platform could have continued to host an image containing a violent and frightening post promoting lynching if only the caption had been slightly different."

"Zeynep Tufekci, a Berkman Klein faculty associate and professor at the School of Information and Library Science at the University of North Carolina at Chapel Hill, argued that the Facebook approach of "moderating from first principles" is problematic. Deciding what constitutes harmful speech is subtle, local, and context-dependent, and is therefore fundamentally in tension with technology business models that emphasize scale, she said. Tufekci asserted that it is "absurd that a platform of 2 billion people is moderated by a team of several thousand" and called on Facebook to dramatically increase the size of its content moderation team."

"Camille François, a principal researcher at Jigsaw — a think tank and technology incubator within Google — discussed her group's recent partnership with the *New York Times* to develop a machine learning tool to help the Times moderate its comment sections online. The new tool, called Moderator, can automatically prioritize comments that are likely to be in need of review or removal, easing the job of content moderation. The tool was trained on more than 16 million moderated Times comments, and has allowed the Times to substantially increase the volume of commenting it allows. François emphasized the importance of transparency and collaboration in developing automated tools of this sort, and she also highlighted the value of data and experimentation."

"Tarleton Gillespie, a principal researcher at Microsoft asked in one session, "What role have our algorithms played in *calling forth* harmful speech online?" Gillespie explained that the

sorting, filtering, and recommendation algorithms that structure many online spaces — such as Twitter's "Trending" or Facebook's News Feed — often form the "terrain" on which harmful speech occurs...Gillespie also raised questions about the second-order effects of using algorithms to moderate online content. For example, he raised concerns about the impact on constituents' sense of autonomy if algorithmic decision-making goes unexplained. Will users still feel in control if Twitter, say, automatically removes a post or comment without explanation?"

****

www.huffingtonpost.com/sean-mcelwee/hate-speech-online_b_3620270.html

The Case for Censoring Hate Speech

By Sean McElwee

"American free speech jurisprudence relies upon the assumption that speech is merely the extension of a thought, and not an action. If we consider it an action, then saying that we should combat hate speech with more positive speech is an absurd proposition; the speech has already done the harm, and no amount of support will defray the victim's impression that they are not truly secure in this society. We don't simply tell the victim of a robbery, "Hey, it's okay, there are lots of other people who aren't going to rob you." Similarly, it isn't incredibly useful to tell someone who has just had their race/gender/sexuality defamed, "There are a lot of other nice people out there."