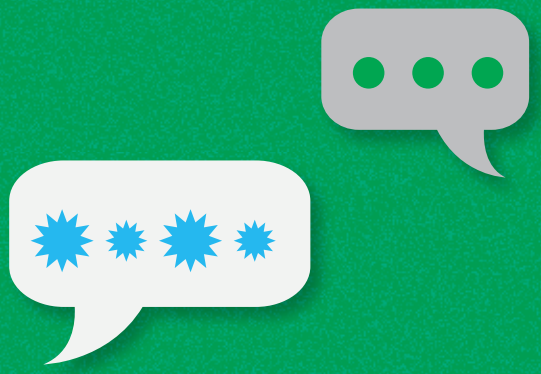
Understanding Harmful Speech Online

Networked Policy Series

December 2016

Translating Research for Action:
Ideas and Examples for
Informing Digital Policy



Research Note

Understanding Harmful Speech Online

Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo

for more from this series visit
cyber.harvard.edu

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Understanding Harmful Speech Online

Research Note

Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo

Suggested Citation: Faris, Robert, Ashar, Amar, Gasser, Urs, and Joo, Daisy. Understanding Harmful Speech Online (December 8, 2016). Networked Policy Series, Berkman Klein Center Publication Series No. 2016-18. Available at: https://cyber.harvard.edu/publications/2016/UnderstandingHarmfulSpeech

Harmful Speech Online Project
23 Everett Street | Second floor | Cambridge, Massachusetts 02138
+1 617.495.7547 | +1 617.495.7641 (fax)
http://cyber.harvard.edu

harmfulspeech@cyber.harvard.edu

Acknowledgements

We are grateful to the John D. and Catherine T. MacArthur Foundation for their support of the Berkman Klein Center's Harmful Speech Online Project. Berkman community members Susan Benesch, Andy Sellars, Chinmayi Arun, Niousha Roushani, Sandra Cortesi, Nathan Matias, Ellery Biddle, and Nikki Bourassa generously shared with us their expertise, feedback, and contributions. Ofra Klein, Prashant Bhat, and Daisy Joo provided valuable research assistance to the project and this paper. The authors also wish to thank the Berkman Klein Center's core and communications teams for their support.

About the Harmful Speech Online Project

The Berkman Klein Center for Internet & Society has launched a research, policy analysis, and network building effort devoted to the study of harmful speech, in close collaboration with the Center for Communication Governance at National Law University in New Delhi, the Digitally Connected network, and in conjunction with the Global Network of Internet & Society Centers. This effort aims to develop research methods and protocols to enable and support robust cross-country comparisons; study and document country experiences, including the policies and practices of governments and private companies, as well as civil society initiatives and responses; and build and expand research, advocacy, and support networks. See https://cyber.harvard.edu/research/harmfulspeech for more information.

Introduction

This paper offers reflections and observations on the state of research related to harmful speech online. The perspectives outlined here are grounded in the lessons from a year of exploratory work in the field by researchers at the Berkman Klein Center and collaborating researchers and institutions. Our review includes an assessment of the efforts of civil society organizations to address racist speech in Brazil and Colombia; a study of the legal foundations of harmful speech regulation in India; a mixed methods look at discourse among white identity groups in the United States; an attempt to track offensive speech online in Tunisia; and a paper that explores the definitional and framing questions that complicate efforts to study and address harmful speech online. We also highlight a small selection of other recent efforts in the field.

A key element of this initiative is to explore different approaches to the study of harmful speech and to draw lessons from comparative analysis. We chose to pursue this diverse set of research efforts in our first year in order to better understand the strengths and limitations of various research strategies and to assess why different types of interventions exist in some contexts and are missing in others. We hope to accumulate enough experiences to begin to answer what has worked and what has not, to define what constitutes success, whether in government, private sector, or community responses such as counter speech.

We come away from this review with a plethora of questions that are worthy of further exploration. Our work over the past year leaves us with a greater appreciation of the complexity of the topic covering a wide range of social phenomena that are manifest in distinctly different ways across different groups and contexts. Each of the methodological approaches described here have strengths and weaknesses and are positioned to help answer different subsets of the many policy questions facing policymakers, companies, and civil society organizations. The continuation and extension of this multifaceted research approach applied to additional countries and topics will help to further refine these methods and provide a basis for robust comparative assessments.

Networked Policymaking & Harmful Speech Online

The Berkman Klein Center for Internet & Society at Harvard University has prepared this research briefing on harmful speech online as a guide for researchers, as well as decision makers in public, private, and civil society organizations seeking to better understand and make informed decisions in this area. This briefing document is part of the Berkman Klein Center's Networked Policymaking Series,[1] which seeks to build knowledge and capacity among peers and across fields, as well as across sectors and among diverse stakeholders, in order for scholarship and evidence-based approaches to have greater impact.

Building on deep institutional knowledge of issues of freedom of expression, as well as collaborative research efforts with stakeholders from across the globe, the BKCIS team seeks to summarize selected research findings related to issues of harmful speech online into practical considerations and takeaways to inform the current state of research. Much of the work conducted through the project points to an evolving state of understanding of the phenomenon which currently limits the ability of researchers to make concrete recommendations for non-academic actors.

1 Translating Research for Action: Ideas and Examples for Informing Digital Policy. Berkman Klein Center Research Publication, September 2016. https://cyber.harvard.edu/node/99639

The Role of Research in Addressing Harmful Speech

Harmful speech is proliferating online, and calls to restrain it come from virtually every country and community. Yet it is in no sense a new phenomenon and has been a recognized problem in online communities for well over two decades. While the primary focus of this paper is harmful speech online, is it broadly observant of the long history of issues surrounding offline harmful speech. Over the past several years, harmful speech online has received much more public and media attention and has emerged as one of the central challenges for Internet policy experts, often pitting protections for freedom of expression online against the rights and interests of those that are subject to online harassment. For some, it draws into question the viability of maintaining the current levels of open participation online and suggests a need to substantially increase moderation and regulatory oversight. Others frame it as a reasonable trade-off and an inevitable price to pay for the benefits of an inclusive and open Internet. A more optimistic view is that there may be ways to significantly reduce the incidence and impact of harmful speech online without unduly restricting freedom of speech. The jury on this is still out.

Data and evidence to inform decision making are scarce, even as governments, Internet platforms, and civil society actors attempt—more vigorously than ever—to diminish harmful speech and its impact. There is an important role for research to better understand the motives, mechanisms, and propagation of harmful speech online, and to support the evaluation and design of potential interventions by government, civil society and private sector organizations. Despite increasing attention to the topic, we still lack a full understanding of the reach and impact of harmful speech online and know relatively little about the efficacy of different interventions. Moreover, our understanding of the collateral costs of various interventions is rudimentary. Research will also help to guide normative judgments that underlie any policy interventions, especially where there are trade-offs between protecting the interests of vulnerable populations and victims of harmful speech online on one hand, and more broadly protecting freedom of expression on the other.

In this paper, we highlight research in two key areas: research that seeks to document and understand the phenomenon of harmful speech online; and research that focuses on the benefits, costs, and efficacy of different approaches to addressing harmful speech online.

The Definitional Quagmire

Whether studying the broad phenomenon or assessing interventions to address harmful speech, one needs to be able to describe and identify what constitutes harmful speech and what does not.

Harmful speech consists of a range of phenomenon that often overlap and intersect, and includes a variety of types of speech that cause different harms. The most familiar type is hate speech, which commonly refers to speech which demeans or attacks a person or people as members of a group with shared characteristics such as race, gender, religion, sexual orientation, or disability. In a companion paper, Sellars reviews prior efforts to define hate speech, offers an in-depth examination of the theoretical context of hate speech and summarizes emerging themes in the discussion and scholarship of hate speech online.[2] An alternative framing—online harassment—is defined by Lenhart, et al as "unwanted contact that is used to create an intimidating, annoying, frightening, or even hostile environment for the victim and that uses digital means to reach the

2 Sellars, Andrew "Defining Hate Speech." Berkman Klein Center Research Publication, 2016. https://cyber.harvard.edu/publications/2016/DefiningHateSpeech

victim."[3] The Women's Media Center describes how harassment online may include a variety of tactics—from doxxing to revenge porn to gender-based harassment and beyond—that impact targets in legal, physical, emotional, and in other consequential ways.[4] Focusing on a narrower subset of harmful speech, Benesch defines dangerous speech as that which increases the risk of violence through a range of rhetorical techniques (e.g. instilling fear by warning of impending threats) and may contain explicit threats or incitement to violence.[5]

The research that we summarize here primarily focuses on harmful speech that is motivated by personal enmity, but the conceptual framing and discussion may also apply to broader conceptions of harmful speech covering for example incursions on privacy, violent extremism online, or financially motivated attacks.

A principal source of confusion and complexity in the debate over harmful speech is the multiplicity of phenomena that are often lumped together spanning a range of instigators, targets, motives, tactics, and media. Incidents of objectionable or potentially harmful speech range anywhere from individual assailants who seemingly target random individuals to orchestrated attacks again individuals or groups and wider movements that link together thousands of participants. In some cases, assailants know their targets. In many others, they do not. Speech that incites violence is markedly different from speech that is 'merely' offensive. And for any given level of vitriol, a single attacker is not the same as a mob.

Conceptually, harmful speech can be defined from a number of different perspectives. One view is outcome-based with a focus on the harm to groups or individuals. An alternative approach looks instead to the intent of the speaker. A third perspective emphasizes the content of the speech. Others have emphasized the need to factor in the context when evaluating intent, content, and harm. There is validity to each of these approaches, and a working definition need not be restricted to only one.

Developing a crisp definition and criteria as a guide for identifying the speech of interest is helpful. However, definitions that allow one to cleanly and reliably identify different forms of speech also tend not to satisfactorily reflect the complexity of harmful speech. For example, defining dangerous speech to be any speech that includes the words 'kill' or 'murder' is simple to apply, but leads to errors of under and over inclusion. Defining dangerous speech as speech that incites violence is conceptually stronger, but leaves a high degree of subjectivity in deciding whether speech fits that standard.

The focus of inquiry or intervention also dictates markedly different approaches to defining and identifying harmful speech. The thresholds for illegal speech are rightly much stricter than the standards for acceptable speech on social media sites or those that motivate civil society organizations to engage in counter speech. As described by Sellars, there are a dizzying array of alternative definitions for hate speech that are found in laws across different jurisdictions and terms of service agreements on social media sites, in addition to those offered by scholars.[6]

---

3 Lenhart, Amanda, Michelle Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. "Online Harassment, Digital Abuse, and Cyberstalking in America." Data & Society Institute, November 21, 2016. https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/

4 Women's Media Center. "Online Abuse 101." Accessed November 30, 2016. http://wmcspeechproject.com/online-abuse-101/#first_what

5 Benesch, Susan. "Proposed Guidelines for Dangerous Speech." Dangerous Speech Project. February 23, 2013. http://dangerousspeech.org/guidelines/

6 Sellars, Andrew. "Defining Hate Speech." Berkman Klein Center Research Publication, 2016. https://cyber.harvard.edu/publications/2016/DefiningHateSpeech

A growing body of literature and legal precedents serves as a critical guide. Although there are no easy answers, the challenges inherent in studying and regulating harmful speech can be addressed by greater clarity, specificity, and focus.

Studying the Incidence, Prevalence, & Impact of Harmful Speech Online

An important area for research in the field is directed at improving our understanding of the phenomenon. For researchers and policy makers, a key transition in the making is the leap from reports of specific incidents of online attacks to a broader understanding of the prevalence of different types of events. At the same time, a wide range of contextual factors—a small sample of which include language, country, medium, as well as instigators, targets, and other actors—are important to consider when assessing the impact of a given statement.

The types of questions to be addressed include:

- How widespread is the phenomenon, who participates, who is harmed, and how?
- Is it increasing or decreasing, and how does it vary over time? Or is there evidence that the prevalence of harmful speech is steady and only receiving increased attention online?
- Are there signs of the normalization of harmful speech online? How are the actors that participate in harmful speech organized? How are they influenced by leaders, governments, public figures, and the media?
- What is the social network structure of groups that engage in harmful speech and what is the role of key influencers? How can we better understand the interplay between in-group and out-group interactions?
- What contextual factors are associated with the incidence, intensity, and impact of harmful speech online?

Research conducted in Tunisia explored a mixed methods approach to studying harmful speech online.[7] Within online spaces in Tunisia, researchers tracked the incidence of a list of inflammatory keywords compiled by the research team across digital media and social media over the past five years. The researchers found that instances of offensive speech (i.e. the subject of the speech may find it insulting, humiliating, derogatory) are mostly concentrated on Facebook, which is also by far the most popular social media platform there. Researchers also found that many spikes in offensive speech are linked to political events on the ground. However, some of the spikes in offensive language are more difficult to explain. Some may be driven by several smaller events and with an underlying variation that might not be explained except by natural fluctuations or stochastic changes.

This analysis suggests that broad scale monitoring of offensive speech through automated means is possible and may serve multiple ends. As a monitoring device, it could be used to alert officials charged with public safety to anticipate flash points and time periods where extra vigilance is needs. From a research perspective, compiling longitudinal data would help to quantitatively test

7 Innova Tunisia. "Hate Speech in Tunisia," 2017 *(forthcoming)*

theories that attempt to explain the nature and evolution of different outbreaks of harmful speech online.

In Tunisia, harmful and offensive speech online appears to be a relatively small proportion of online speech, in contrast to the widespread perception that the Internet is filled with vitriol. Researchers found that approximately 0.5% of social media posts included language associated with offensive speech. This is consistent with findings by researchers that estimated the incidence of such speech in Ethiopia based on a detailed study that took place in 2015.[8]

Research focused on the discursive practices of white identity groups in the United States on Twitter attempted to identify and track different types of offensive speech online, dividing instances into speech that calls for violence, speech that promotes discriminatory actions again certain groups, speech that is offensive to target groups, and speech oriented towards bonding among groups but that is not overtly offensive to other groups.[9] Like others before us, we had a difficult time distinguishing among these different categories of speech. While we found examples of speech that would clearly be construed as offensive, there are many examples where the threshold for speech that might be identified as racist, misogynistic or religiously biased is fuzzy. The use of coded language and apparent dog-whistle references among the white identify groups we examined suggests that clearly and consistently identifying speech that acts to offend, discriminate, or threaten may be impossible. We also found very little dangerous speech and little discriminatory speech. While there are myriad well-documented incidents of highly charged and potentially injurious speech online, we found little evidence of it in the everyday discourse in social media, even among groups reputed to engage in such behavior.

In the appendix, we list several studies that have taken on different parts of the issue.

Assessing Strategies & Policies to Address Harmful Speech

Researchers have a valuable role in better understanding the efficacy of different policies and approaches to address harmful speech. While there are many possible points of intervention, success stories are still few and far between. Options for addressing harmful speech can be divided into two strategies: those that aim to reduce the incidence of such speech and those meant to mitigate the impact where it does occur (See Table 1).

	Legal	**Content curation & filters**	**Normative**
Reduce incidence and prevalence	Pursuing action against instigators of illegal speech	Terms of service enforcement Taking down posts Blocking users	Education/Literacy Counter-speech Public leadership
Mitigate impact	Validation Punitive awards	Individually controlled blocking features	Counter narratives Media representation Community support

Table 1. Strategies for addressing harmful speech online

8 "Mechachal: Online Debates and Elections in Ethiopia. Report One: A Preliminary Assessment of Online Debates in Ethiopia." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2782070

9 "Discursive practices of white identity groups on Twitter," 2017 (*Forthcoming).*

As summarized in Figure 1 below, there are a range of actors that might be involved in addressing harmful speech. Reducing the incidence of harmful speech might be achieved by censoring harmful speech, discouraging harmful speech through legal means, and employing strategies designed to shape behavioral norms, for example through education or counter speech. In discouraging harmful speech, initiatives might attempt to change the core attitudes and ideologies that feed harmful speech or, short of that, convince instigators to abstain from expressing their views in a manner that is detrimental to others. Limiting the impact of such speech is another strategy, either by reducing the exposure of targeted individuals and groups from such speech or providing a counter balancing stream of positive messages and expressions of community support. While the categories of actors and points of intervention are few, there are great number of complex human, technical, and institutional interactions embedded within each of these avenues.

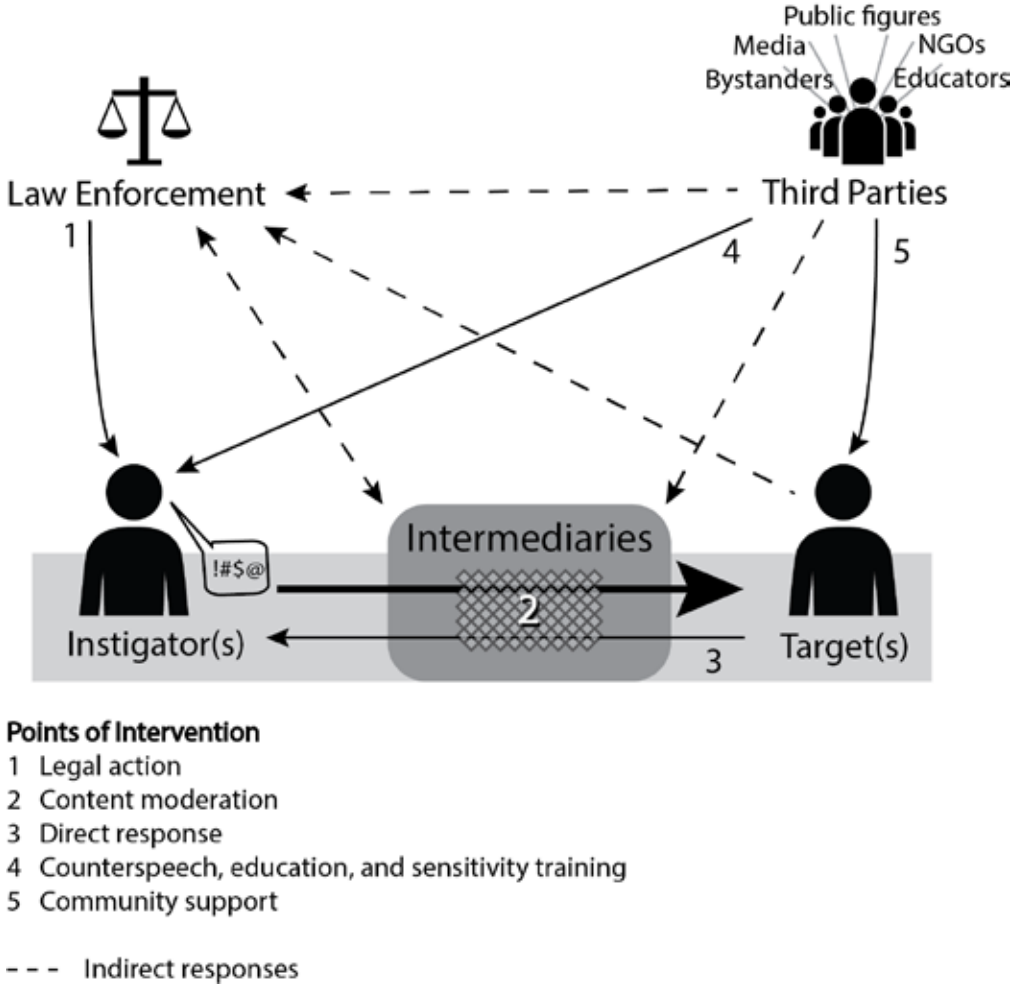


Figure 1. Mapping Dynamics Between Actors and Points of Intervention

Legal Remedies

A typical starting point in assessing options for addressing harmful speech is the role of government action, either in the laws and regulations that govern online speech or the stridency in which law are enforced. Where a relatively small number of individuals engage in illegal speech online, a law enforcement-centered response makes sense. However, by most accounts, the extent of the problem is considerably greater than enforcing existing laws and combating illegal speech; a sizable proportion of deeply objectionable and harmful speech either is not illegal or falls outside of the reach of law enforcement.

Research conducted in India over the past year, with particular emphasis on the legal foundations for regulating harmful speech, offers a view of the complexities of the legal landscape within the country, along with several promising angles for future work.[10] One point that stands out is the overlapping layers of potential legal instruments that form the laws and regulations and facilitate and constrain formal legal action again harmful speech. There are several laws that cover speech commonly included under the hate speech umbrella forming a patchwork of overlapping laws, although the term "hate speech" is not used in any of statues. The legal basis for restricting harmful speech in India is rooted in the Constitution, which contains a 'public order' exception to freedom of speech protections.

Of particular concern is excessive criminalization of speech in India, including harmful speech, infringing upon legitimate speech and freedom of expression. Arun and co-authors report that laws are "often imprecise and overbroad, covering vast swathes of legitimate speech."

The research in India also highlights challenges to the implementation of existing laws, primarily the combination of anonymous speech on social media platforms outside of Indian jurisdiction which makes it difficult to identify those engaging in illegal speech online. Partly as the result of this, Indian authorities in certain regions have resorted to shutting down the Internet entirely.

The authors of the case study conclude that hate speech law in India is outdated and is ineffective at preventing violence, while powerful speakers who cause real harm are able to avoid punishment. They also note the rise of social media labs to monitor harmful speech and the initiation of counter-speech projects. For example, they reference a police initiative seeking to quell rumors that might incite violence by engaging in social media.

While legal remedies clearly have their place, they also have limitations. When considering a stronger legal response as a part of the solution, an obvious point of concern is whether expanding legal tools and enforcement capacity does more harm than good, or even harms those groups that are targets of harmful speech. Further research is critical towards building a better understanding of these dynamics, which will help to induce and inform legal reform.

Civil Society Responses

Where harmful speech permeates broad reaches of societal public discourse and harms people in ways that are within the boundaries of protected speech, the nature of the issue and potential interventions are quite different, and this shifts the burden from government action to civil society groups and the private sector. Research conducted in Brazil and Colombia documented the work of civil society groups there to influence public discourse with particular attention to countering

10 Arun , Chinmayi and Nayak, Nakul. "Preliminary Findings on Online Hate Speech and the Law in India." Berkman Klein Center Research Publication, 2016. https://cyber.harvard.edu/publications/2016/HateSpeechIndia

racism online. [11]

Online speech has opened up new venues for expressing racist ideology. Although it represents a particularly worrisome manifestation of these sentiments, racist speech online is seen as part of much larger problem, as it is embedded in a broader context and long history of discrimination.

The prevailing opinion of experts in the region is that, if unaddressed, the presence of racist speech online will translate into greater social harms and exacerbate the impact and damage of racism. However, the emergence of racist speech online also has brought these deeply seated issues to the surface and made racist ideologies more visible to a broader swath and society. This phenomenon makes it more difficult for individuals and societies to avoid the issue and may force more people to come to terms with the longstanding problem. A second factor that may shift dynamics is how online media offers the targets of racist speech a mechanism to craft their own response and to tell the story in the way that they want it told. Online avenues for generating counter narratives have given voice to a growing number of civil society activists and organizations, and for a substantial proportion of the population, an opportunity to define their plight on their terms. Qualitative evidence suggests that this has had a significant positive influence on Afro-descendent communities, although practitioners also asserted that their efforts would have greater impact in conjunction with increased efforts by private companies and government to address online racist speech.

Understanding and measuring the efficacy of civil society efforts to reframe online narratives and to lessen the damage to minority communities from online speech is an area in which researchers could add tremendous value in both informing ongoing policy debates and in supporting ongoing efforts to address harmful speech online. This attempt to study civil society through interviews and qualitative methods demonstrates both the strengths and limitations of this approach. An ethnographic approach is able to consider a broad range of complex factors and draw upon the knowledge and perspectives of those that best understand the dynamics at play. A weakness of this approach is that it is difficult to measure impacts with precision and to compare trends over time.

Emerging attempts to understand and measure the efficacy of interventions to harmful speech, particularly in the area of counter speech, offer a glimpse of promising areas of inquiry for researchers. Based on an analysis of counter speech practices on Twitter, Benesch and Ruths characterize successful counter speech as having "favorable impact on the original (hateful) Twitter user, shifting his or her discourse if not also his or her beliefs" or "positively affect[ing] the discourse norms of the 'audience' of a counterspeech conversation: all of the other Twitter users or 'cyberbystanders' who read one or more of the relevant exchange of tweets."[12] The authors also suggest that other indicators, such as whether significant numbers of Twitter users join a campaign, may be used to explore whether counter speech had an effect on users' motivations or perspectives. In addition, Munger also recently conducted an experiment among groups of users on Twitter considered harassers on the platform and found that counter speech using automated bots can impact and reduce instances of racist speech if "that subjects… were sanctioned by

11 Roushani, Niousha. "Deconstructing and Reconstructing Representations of Afro-descendants: Hate Speech, Race, and Inequality in Colombia and Brazil." Berkman Klein Center Research Publication, 2016. https://cyber.harvard.edu/publications/2016/GrassrootsPerspectives

12 Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. "Counterspeech on Twitter: A Field Study." A report for Public Safety Canada under the Kanishka Project, https://www.scribd.com/document/327586365/Counterspeech-on-Twitter-A-Field-Study

a high-follower white male."[13] Bartlett and Krasodomski-Jones of Demos express the difficulty of capturing what it means to have a successful outcome in responding to harmful speech, but offer quantitative metrics (such as engagement, volume, reach), content metrics (using sentiment analysis or natural language processing techniques), or real world metrics (examining effects of counter speech in offline contexts) as starting points for measurement and analysis.[14]

The Role of Intermediaries

Intermediaries and platforms occupy a powerful position as hosts of content, gatekeepers and enforcement agents, and architects and designers of online environments. Researchers have contributed to a number of efforts to document the actions of intermediaries, and in some instances, have played a more active role advocating to and advising companies on policies. Matias has explored the practices and governance structure of volunteer moderators of platforms that are centered on online communities which are actively encountering sub-communities that are organized around topics that may be considered harmful or hateful.[15] A number of academics who have studied harmful speech online also sit on Twitter's Trust and Safety Council, announced in early 2016, which works with representatives from across sectors to prevent abuse on the platform.[16] The Wikimedia Foundation has also begun to collaborate with industry groups and academics to explore how machine learning techniques may help users and the platform deal with "toxic" speech, under a project called Detox.[17] These projects represent a small window into the many efforts researchers across sectors are undertaking in order to better understand the phenomenon, dynamics, and difficult questions posed by harmful speech online.

13 Munger, Kevin. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." Political Behavior, 2016. http://link.springer.com/article/10.1007/s11109-016-9373-5

14 Bartlett, Jamie and Krasodomski-Jones, Alex. "Counter Speech: Examining Content That Challenges Extremist Online." Demos, October 2015.http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

15 Matias, J. Nathan. "Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1138-1151. ACM, 2016.

16 Twitter Trust and Safety council. https://about.twitter.com/safety/council

17 Adams, CJ, Lucas Dixon, Patrick Earley, Haitham Shammaa, Dario Tababorelli, Nithum Thain, and Camille Francois. "Research: Detox." Accessed November 30, 2016. https://meta.wikimedia.org/wiki/Research:Detox

A Summary of Research Methods

Studying harmful speech online demands a cross-disciplinary approach and expertise in different methodologies. Each of these research approaches have different strengths and weaknesses (as summarized in the table below) and are thereby highly complementary. To make progress, continuing in a multi-method approach makes sense. Moreover, the value of any particular approach is strongly linked to the social and political context, and hence the application of different research designs must be tailored to given country and focus.

	Applications	**Limitations**
Digital media monitoring	Increasingly sophisticated data gathering and analysis tools able to monitor digital communication at a broad scale. May provide early warning systems.	Often lacks fine-tuned interpretation of content.
Social network analysis	Identify links between actors, key nodes, and community structure. Potential for tracking the spread of ideologies and frames.	Broader contextual information is necessary for interpretation. Difficulty in establishing causal links.
Content analysis	Distinguish between different types of speech, instigators, and targets. Offers perspectives on sentiment and intention. Able to combine and leverage human and automated approaches.	Human coding is time intensive. Frequent ambiguity in language.
Legal and policy analysis	Document legal and policy instruments. 'Law in action' analysis. Comparative assessments.	Often involves extensive detailed analysis. Defining common frameworks for comparative work can be challenging.
Interviews, surveys & focus groups	Evaluate prevalence and impact. Assess counter measures.	Resource intensity limits application. Potential for selection and reporting biases.
Experimental studies	Ability to test responses and impacts in different situations.	Ethical considerations if applied without consent. Applicability of results from controlled experiments may be limited.

Table 2. A Summary of Research Methods, Applications, and Limitations

Reflections and Next Steps

In this research note we describe three principal focal areas for further study and possible intervention: law and policy; civil society engagement and counter speech; and intermediaries and content moderation. These interfaces are part of a broader system and hence interconnected; efforts directed at one facet must be considered in the broader context. We believe that research will play a critical role in policy design—both in crafting sound policy and fending off misguided policy—and that efforts in each of the focal areas is warranted. There is much progress to be made on all fronts.

Harmful Speech Project Companion Papers

- Arun, Chinmayi, and Nayak, Nakul. Preliminary Findings on Online Hate Speech and the Law in India (December 8, 2016). Berkman Klein Center Research Publication No. 2016-19. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882238

- Roshani, Niousha. Grassroots Perspectives on Hate Speech, Race, and Inequality in Brazil and Colombia (December 8, 2016). Berkman Klein Center Research Publication No. 2016-18. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882234

- Sellars, Andrew F. Defining Hate Speech (December 8, 2016). Berkman Klein Center Research Publication No. 2016-20. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

- Innova Tunisia. "Hate Speech in Tunisia," 2017 (Forthcoming).

- "Discursive practices of white identity groups on Twitter," 2017 (Forthcoming).

Additional Reading

This section includes a small sample of papers, articles, and developments that focus on issues and approaches related to harmful speech online. It is by no means comprehensive, as these resources are meant to be a starting point for learning more about the phenomenon and research efforts underway to better understand it.

Research Papers

- Mohammad, Saleem H., K. P. Dillon, S. Benesch, and D. Ruths. "A Web of Hate: Tackling Hateful Speech in Online Social Spaces." Proceedings of First Workshop on Text Analytics for Cybersecurity and Online Safety, Portorož, Slovenia. May 20, 2016. http://www.ta-cos.org/sites/ta-cos.org/files/tacos2016_SaleemDillionBeneschRuths.pdf.

    » *This conference paper discusses the limitations of keyword-based methods to identifying hate speech on social platforms and suggests the importance of community-based data in better detection and understanding of hateful speech. The study proposes a detection technique that uses language models based on data from self-identifying hateful communities, which avoids the interpretive challenge involved in manual annotation of offensive terms and requires less effort than collecting training data for classifiers. Using reddit as the primary source for the hateful communities, the research paper employed the method to create language models for target-specific hateful*

speech that succeeded to outperform keyword-based classifiers.

- Benesch, Susan. "Proposed Guidelines for Dangerous Speech." Dangerous Speech Project. February 23, 2013. http://dangerousspeech.org/guidelines/.

  » *Benesch proposes guidelines to isolate a subset of hate speech, termed "dangerous speech," for the purpose of locating key indicators of early warnings for mass violence, limiting violence by finding ways to limit identified speech, and holding speakers accountable for such speech that constitutes crime. The guidelines suggest the analysis of five variables in determining the dangerousness of a particular speech act: the speaker, the audience, the speech act itself, the socio-historical context, and the mode of dissemination.*

- Gagliardone, Iginio, Matti Pohjonen, Abdissa Zerai, Zenebe Beyene, Gerawork Aynekulu, Tewodros Gebrewolde, Michael Seifu, Nicole Stremlau, Jonathan Bright, Mesfin Bekalu, and Mulatu A. Moges. *Mechachal: Online Debates and Elections in Ethiopia. Report One: A Preliminary Assessment of Online Debates in Ethiopia*. October 2, 2015. https://ssrn.com/abstract=2782070

  » *During Ethiopia's 2015 general election, researchers found that only 0.7% of more than 13,000 statements made by Ethiopians on Facebook could be classified as "hate speech." This research study suggests that actual levels of online hate speech may be lower than anecdotal evidence, highlighting the demand for research that can detect and monitor online speech activity. The research proposes an alternative view that social media can serve as a space for tolerance and acceptance.*

- Marwick, Alice E., and Ross W. Miller. "Online harassment, defamation, and hateful speech: A primer of the legal landscape." Fordham Center on Law and Information Policy Report 2 (2014).

  » *Warwick and Miller's article describes the legal remedies available to victims of online harassment, hate speech, and defamation in the United States, in addition to the current legal protections afforded to such speech under the Constitution of the United States, and discusses related drawbacks and complications. The research suggests that legal remedies are most useful in specific cases such as harassment that constitutes a "true threat," since the laws consider First Amendment protection of most online speech. Legal remedies for online harassment are also complicated by the difficulty of identifying instigators and appropriate jurisdictions.*

- Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. UNESCO Publishing, 2015.

  » *UNESCO's report presents an overview of hate speech online and counteractive measures that have been adopted on a global level. Using an extensive literature review and other techniques for data collection and analyses of produced content, and semi-structured interviews, the study identifies four main tensions of contending with hate speech online: definition of the category of speech, jurisdiction for regulatory enforcement, comprehension of hate speech online in relation to offline speech and action, and intervention by varied and unrelated entities.*

- "'Hate Speech' Explained: A Toolkit." ARTICLE 19. December 23, 2015. https://www.article19.org/resources.php/resource/38231/en/'hate-speech'-explained:-a-toolkit.

  » *ARTICLE 19's guide addresses the following three questions related to hate speech: 1) how to identify hate speech that can be restricted and distinguished from protected speech?; 2) what positive measures can countries take to counter hate speech; and 3) which types of hate speech should be prohibited countries and under which circumstances? In discussing these questions, ARTICLE 19, highlights the importance that responses to hate speech comply with international human rights law.*

- Matias, J. Nathan, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. "Reporting, reviewing, and responding to harassment on Twitter." *Available at SSRN 2602018* (2015).

  » *This paper assesses a three-week project in November 2014 during which Twitter granted Women, Action, and the Media (WAM!) a special privilege to identify and report inappropriate content on behalf of others. Using both quantitative and qualitative methods, the study made findings on the people reporting and receiving harassment, the kinds of harassment reported, Twitter's response to harassment reports, the process of reviewing such reports, and challenges for related reporting processes.*

- Matias, J. Nathan, Camille Francois, Amy Johnson, Emilie Reiser, Susan Benesch, Lindsay Blackwell, Amy Bruckman, Jen Carter, Soraya Chemaly, Justin Cheng, Jason Coon, Lucas Dixon, Nicole Ellison, Eric Gilbert, Rey Junco, Cliff Lampe, Mariel Garcia M, Merry Mou, Katherine Lo, Alice Marwick, Kevin Munger, Sarah Otts, Derek Ruths, Andy Sellars, Sarah Sobieraj, T.L. Taylor, Nithum Thain, and Ellery Wulczyn. *High Impact Questions and Opportunities for Online Harassment Research and Action.* Report. MIT Center for Civic Media, Massachusetts Institute of Technology. August 2016. https://civic.mit.edu/sites/civic.mit.edu/files/OnlineHarassmentWorkshopReport-08.2016.pdf.

  » *This report is based on discussions held at a two-day Online Harassment workshop organized by Massachusetts Institute of Technology Media Lab and Jigsaw and relates key questions for progress on online harassment, information on infrastructures to support online harassment research, and updates on current high impact research projects, including estimating the chilling effects from online harassment.*

- Matias, J. Nathan. "Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1138-1151. ACM, 2016.

  » *This research paper examines the social factors that can lead to participation in mass collective action against a platform through an analysis of a 2015 protest by moderators of "subreddit" communities against the social news platform reddit where moderators collectively disabled their subreddits. This "blackout" of subreddits prevented millions of users from accessing parts of the platform and led reddit to negotiate with the moderators' demands. Matias utilizes mixed methods participatory hypothesis testing to find that the predictors of participation related to moderators' grievances: the work of moderators and their relationships with company policies. Matias also notes that community members were strong factors in pressuring moderator participation in the protest.*

- Munger, Kevin. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." Political Behavior, 2016. http://link.springer.com/article/10.1007/s11109-016-9373-5

  » *Munger's research measures the effect of specific interventions utilizing social sanctioning and promotion of norms on Twitter harassment. Through different Twitter accounts that varied in in-group/out-group identities and also in influence, Munger tweeted at Twitter harassers stating that their behavior was unacceptable. His research supports the hypothesis that such counter-messages sent by Twitter accounts that have both an in-group identity and high influence (white men with a large number of Twitter followers) caused the largest reduction of offensive behavior among a subject pool of white men. Munger notes that this method of performing experiments on social media subjects using experimenter controls can be important for examining online speech, and that an important extension would be a manipulation to reduce misogynist online harassment.*

- Bartlett, Jamie, Jeremy Reffin, Noelle Rumball, and Sarah Williamson. *Anti-social Media.* DEMOS Centre for Analysis of Social Media. February 2014. http://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf?1391774638.

  » *This report from the DEMOS Centre for Analysis of Social Media examines the manner in which racial, religious, and ethnic slurs are employed on Twitter, in an aim to better understand the pattern of hate speech online. The methodology involved the collection of public tweets containing one or more candidate slurs, crowd-sourced from Wikipedia, and categorization via automated machine classifiers and human analyses. The analyses provided an opportunity to estimate the prevalence of certain patterns of hate speech on Twitter (the authors estimate that roughly 2,000 English language tweets per day are directed racially or ethnically prejudicial).*

Framing, Identifying, and Addressing Issues of Harmful and Hate Speech Online

- Bazelon, Emily. "How to Stop the Bullies." *The Atlantic*, March 2013. http://www.theatlantic.com/magazine/archive/2013/03/how-to-stop-bullies/309217/.

  » *Bazelon's article describes the impact of social media on bullying and delves into social media platforms' policies on hate and harassment. Bazelon suggests the importance of citizen pressure on social media platforms to encourage platform responsibility for harmful posts.*

- Chaudhry, Irfan. "# Hashtagging hate: Using Twitter to track racism online." First Monday 20, no. 2 (2015).

  » *Chaudhry reviews three recent projects that have used Twitter as a data collection tool to track racist language in order to present the varied existing methods of analysis and each method's accompanying strengths and challenges. Projects reviewed include 1) Racist Tweets in Canada (the author's work); 2) Anti-social media (the 2014 study by DEMOS); and 3) The Geography of Hate Map (work by researchers at Humboldt Uni-*

versity). Chaudhry emphasizes the opportunity presented by Twitter as a data collection tool for researchers interested in studying race and racism, "to delve into the platform to see the ways Twitter users openly discuss race and racism – something which is not often seen in the 'off-line' world."

- Council of the European Union. *Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law.* November 28, 2008. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3Al33178

  » *This component to the broader European Union legislative framework requires Member States to penalize the most severe forms of hate speech and hate crime.*

- Dewey, Caitlin. "Robin Williams's daughter Zelda driven off Twitter by vicious trolls." The Washington Post, August 13, 2014. https://www.washingtonpost.com/news/the-intersect/wp/2014/08/13/robin-williamss-daughter-zelda-driven-off-twitter-by-vicious-trolls/

  » *This article described how a person can attacked without any reference to a group, such as vicious tweets sent to Robin Williams's daughter Zelda, immediately after his 2014 suicide, blaming her for his death. Doxxing, or publishing a person's private information, is also a form of harmful speech.*

- Nossel, Suzannne. "To Fight 'hate Speech,' Stop Talking about It." *The Washington Post*, June 3, 2016. https://www.washingtonpost.com/posteverything/wp/2016/06/03/we-dont-need-laws-banning-hate-speech-because-it-doesnt-exist/?utm_term=.8fd38e875a01.

  » *Nossel's article warns of the danger in using the umbrella of "hate speech," including the criminalizing effect of expression when the term "hate speech" is used in countries where free speech norms are not robust. Nossel observes that another problem of the concept of "hate speech" is that it does not distinguish between hateful intent and effect and suggests that certain kinds of "hate speech" be denoted and qualified as "dangerous speech" or "denigrating speech."*

- Oluo, Ijeoma. "Leslie Jones' Twitter Abuse Is a Deliberate Campaign of Hate." *The Guardian*, July 19, 2016. https://www.theguardian.com/commentisfree/2016/jul/19/leslie-jones-twitter-abuse-deliberate-campaign-hate.

  » *This article covers a recent case of online harassment on Twitter experienced by actor Leslie Jones and argues that the received tweets are not harmless pranks but better categorized as an incident within a "deliberate campaign of abuse" against women of color online. Oluo appeals to social media platforms to support their commitment to free speech by taking action against such online harassment, which has the effect of silencing members of minority groups online.*

- Silva, Leandro, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. "Analyzing the Targets of Hate in Online Social Media." *arXiv preprint arXiv:1603.07709* (2016).

  » *This study develops a methodology to identify hate speech on the social media systems of Whisper and Twitter and provides directions for possible prevention and detec-*

tion techniques. Instead of measuring hate speech with explicit hate words or targets, this study proposes and uses a method of collecting hate speech by sentence structure, then by filtering the results through templates created with the help of an online repository of hate speech, Hatebase.

- Thompson, Marcelo. "Beyond Gatekeeping: The Normative Responsibility of Internet Intermediaries." (2015).

  » *Thompson's article proposes a normative approach to Internet intermediary liability and suggests that the focus on the responsibility of Internet intermediaries should be in their decision-making processes and not on the outcomes of such decisions (similar to the standard that journalists are held to in publishing defamatory articles).*

- Umati Project Team. *Umati: Monitoring Online Dangerous Speech*. iHub Research. February 2013. https://ihub.co.ke/ihubresearch/uploads/2013/february/1361013008_819_929.pdf.

  » *This report provides information on efforts, methodologies, and findings of the Umati Project, which seeks to monitor the use of dangerous speech in Kenya on new media, including blogs, forums, Facebook, and Twitter, analyze the speech's effect on violence, and determine non-governmental methods of countering such speech.*