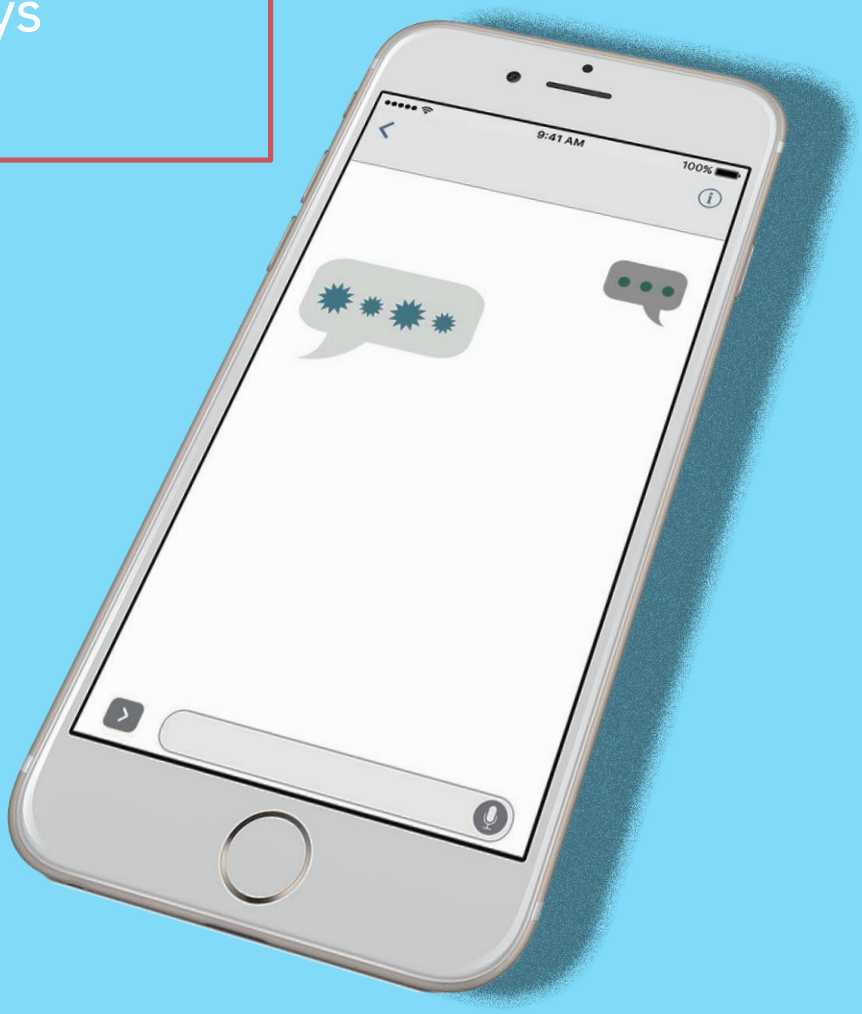
Perspectives on Harmful Speech Online

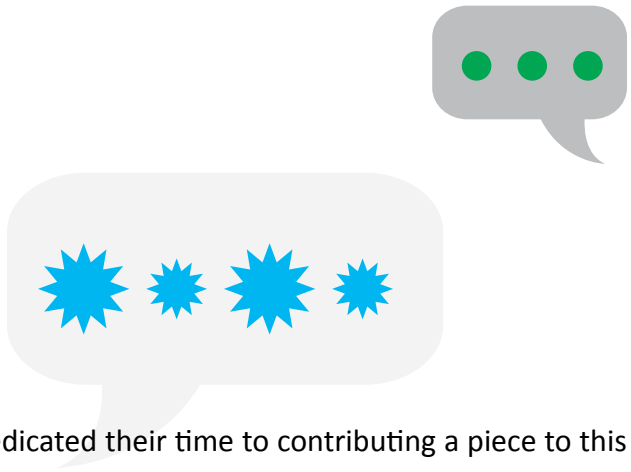
HARMFUL SPEECH ONLINE

Perspectives on Harmful Speech Online

a collection of essays

August 2017





Acknowledgments

I am deeply grateful to the writers and thinkers who dedicated their time to contributing a piece to this collection and shared their ideas and research with us.

Special thanks to Susan Benesch, who gave thoughtful and insightful feedback into the collection's construction and served as a guiding light throughout its assembly, and to Sandra Cortesi, who shared sage advice and offered recommendations that shaped the collection's development.

I thank the members of the Harmful Speech Online project at Berkman Klein, whose work is helping to illuminate the many challenges, problems, ideas, and solutions related to harmful speech online, and knitting together a common understanding of these. Rob Faris, Berkman Klein's research director, drove the collection's overall vision and direction. Senior researcher Amar Ashar imparted invaluable editorial experience and expertise on the process, and Nikki Bourassa managed its overall development and provided research and editorial support.

Lastly, I would like to give special thanks to the John D. and Catherine T. MacArthur Foundation for its support of our work on harmful speech online.

The views and opinions expressed in the articles are those of the authors.

Urs Gasser
Executive Director
the Berkman Klein Center for Internet & Society at Harvard University

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax)
cyber.harvard.edu • hello@cyber.harvard.edu

Permanent link: http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746096

About This Collection

This collection of essays includes perspectives on and approaches to harmful speech online from a wide range of voices within the Berkman Klein Center community. Recognizing that harmful speech online is an increasingly prevalent issue within society, we intend for the collection to highlight diverse views and strands of thought and to make them available to a wide range of audiences.

We issued an open call to our community for short pieces that respond to issues related to harmful speech online. Through this collection, we sought to highlight ongoing research and thinking within our extended community that would be available to readers in a way that is more accessible than traditional academic research. The 16 short essays compiled in this collection are authored by a global group of friends, colleagues, and collaborators. We hope that this diverse mix of perspectives, viewpoints, and data points provokes thought and debate, and inspires further exploration.

Evidence of the complexity of the issue is that no two writers sought to cover the same topic from a similar point of view; from legal perspectives to research results to paradigm-shifting provocations, a multitude of topics, opinions, and approaches are included. Many pieces draw from research, while others are more opinion-based, indicating that discourse around this topic can be inherently opinionated and passionate as well as scholarly and academic. Some pieces are written in a style evocative of advocacy, whereas others are written with scholarly communities in mind. The range of perspectives and opinions found here—and the lack of consensus on some topics—highlight the dynamic complexity of the issues and how competing values are frequently entangled.

We organized the pieces into three categories: Framing the Problem, International Perspectives, and Approaches, Interventions, and Solutions. The first and last sections include essays that build upon our understanding of their categories, and the section on International Perspectives addresses specific geopolitical contexts and ways in which the regulation of harmful speech may or may not be serving the citizens of a particular country or region.

To quote the old adage, To move a mountain, it takes a village. Although many pieces of the puzzle related to harmful speech online require our collective attention, the experts featured in this collection offer a variety of responses. We hope that their insights engender conversation, prompt reflection, and inspire action around the world.

About the Harmful Speech Online Project

The Berkman Klein Center for Internet & Society is undertaking 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 and 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. Our efforts build upon many complementary projects and initiatives, including the Berkman Klein Center's ongoing work related to youth-oriented online hate speech, as well as the activities of various individuals and institutions within our networks.

Contents

Introduction

From cyberbullying to violent extremism to gender-based or racial harassment, harmful speech permeates online space in a variety of forms, despite society's best attempts to prevent it. Although its various forms are linked in the sense that they cause harm to a targeted individual, harmful speech is notoriously difficult to define, and thus studying and regulating it becomes exceptionally challenging as well.[1]

Related to this challenge is the rapid rate of technological innovation that our global society is currently experiencing. Internet penetration is at an all-time high, with 81% of the populations of developed countries, 40% in developing countries, and 15% in the least developed countries having the capacity to connect to the internet.[2] These numbers indicate there is still room for internet growth as prices continue to drop, and, as an extension, digital platforms and content are welcoming growing audiences through continued development and expansion. As the internet consumes larger and larger chunks of our attention and as engagement with it increases, the effort to mitigate detrimental effects of harmful speech online becomes urgently necessary.

This changing online world requires those working on issues related to harmful speech to constantly upgrade and refresh their outlooks; with every fun new technology comes the potential for unanticipated cases of abuse.[3] Indeed, a central theme that emerged from this collection was the role of platforms in the harmful speech ecosystem. Platforms abound, ready to connect people to their peers and interests in new and innovative ways, but they are also unintentionally equipped to be used as tools of harm and hurt, ready to capitalize on the connected masses to torment a targeted individual just as easily as to crowdfund for a morally worthy cause.

There is hope yet. With the internet's perpetual growth and evolution comes not just new problems but also the opportunity for new approaches and strategies to prevent or otherwise mitigate instances of harmful speech online. Many authors in this collection wrestled with concepts of governance from legislative, platform-level, and user-level perspectives. Others identified possibilities for technical interventions, research initiatives, and grassroots civil society approaches, speaking to the many opportunities for

Encouraging freedom of expression while mitigating harms to users is no easy task. However, the possibilities are endless and ideas abundant. This collection captures that frontier spirit.

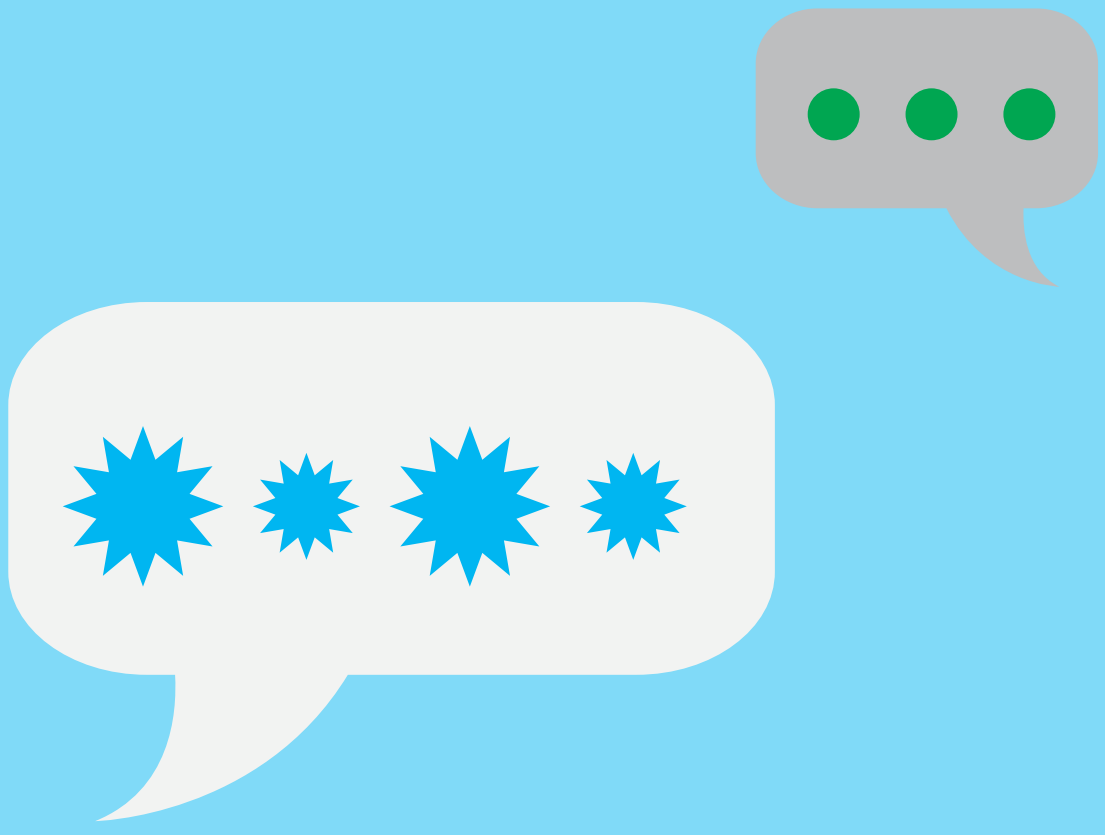
References

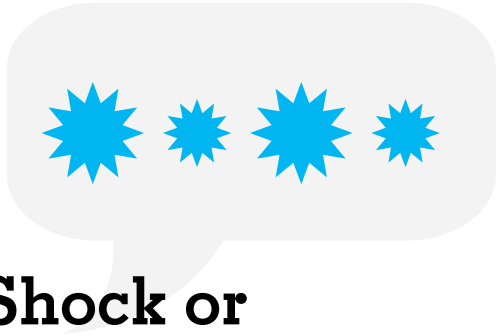
1 - Rob Faris, Amar Ashar, Urs Gasser, and Daisy Joo, "Understanding Harmful Speech Online," Berkman Klein Research Center, December 2016, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882824.

2 - "ITU releases 2016 ICT figures," International Telecommunications Union, July 22, 2016, http://www.itu.int/en/mediacentre/Pages/2016-PR30.aspx.

3 - Cherlynn Low, "Facebook Will Hire 3,000 Moderators to Prevent Livestreamed Violence," engadget, May 3, 2017, https://www.engadget.com/2017/05/03/facebook-will-hire-3-000-moderators-to-prevent-livestreamed-viol/.

Framing the Problem





The Right to 'Offend, Shock or Disturb,' or The Importance of Protecting Unpleasant Speech

Nani Jansen Reventlow

Free speech, a fundamental component of democracy, has been the subject of increasing debate as harmful speech, both online and offline, has emerged as a topic of public attention. While a recent *New York Times* article calls into question whether fixing problems such as online harassment is even possible,[1] the serious threat that online abuse, especially of women,[2] poses to free speech is widely acknowledged.

With such harms in mind, this brief essay does not dispute the importance of striving toward an internet that is a safe and open space for all to exchange views and ideas, regardless of gender, race, religion, or affiliation. Instead, it intends to underline the importance of devising measures to combat harmful speech online that leave sufficient space for the right to "offend, shock or disturb," as the European Court of Human Rights aptly stated in the Handyside case.[3] While stopping short of arguing for the creation of a "right to insult," as was recently proposed,[4] this essay does argue that we should take care in safeguarding a space in which unpleasant, unpopular, and offensive ideas and views can be freely shared, to ensure that free speech indeed remains the cornerstone of a democratic society.

The First Amendment to the United States Constitution probably offers the most robust protection to unfavorable speech. Except for so-called "fighting words," true threats, and incitement, a wide array of offensive speech is protected in the United States.[5,6] This broad scope of protection was recently brought to the fore when the ACLU[7] expressed support for Milo Yiannopoulos, editor of the far-right website Breitbart, after his scheduled talk at Berkeley University was canceled following violent protest.[8] This support drew severe criticisms, including from ACLU supporters. Lee Rowland, senior staff attorney at the ACLU, commented in response: "There's no question that the things that Mr. Yiannopoulos says are unbelievably hateful in nature. But the phrase hate speech is a form of free speech. ... we must all reach out and protect the speech that we most disagree with or else the First Amendment is just reduced to a popularity contest and has no meaning."[9]

The landscape outside the United States looks different. The main treaty regulating speech internationally is the International Covenant on Civil and Political Rights (ICCPR),[10] which sets out the parameters of the right to free speech in Article 19, including the permissible limitations to the right. Limitations to the right to free expression can be permissible if, in short, they pursue a legitimate aim (the rights and reputations of others, public order, public morals, and national security); have a basis in a law that is of sufficient quality (the law should be clear enough for citizens to regulate their conduct and not allow for authorities to exercise unfettered discretion in its application); are necessary and proportionate (there is no overriding public interest in the expression and the limitation is the least invasive measure that could be applied). Similar standards can be found in the European, African, and Inter-American regional treaties protecting human rights. Article 20, which the U.N. Human Rights Committee—the body that oversees the ICCPR's implementation—indicated is a lex specialis to Article 19, obliges States Parties to the treaty to prohibit by law, first, any propaganda for war and, second, "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[11]

The myriad of questions that these standards raise (what type of speech should be criminalized? should speech be criminalized in the first place? if so, how?) and the considerable gray area of what does and does not qualify as "hate speech" (does a racial slur fall within the remit of Article 20? what about an offensive comment about gay people?) merit a detailed discussion that falls beyond the scope of this essay. It is worthwhile, however, to emphasize the importance of drawing the line between acceptable and illicit speech very carefully and keeping very clearly in mind what offensive speech means for a democratic society.

In these pursuits, I suggest three fundamental considerations. First, views on what is considered offensive or acceptable speech will inevitably change according to who is judging. This is exactly why it is dangerous to put any policing powers of this sort into anyone's hands—let alone the hands of private actors such as intermediaries, who will have a clear interest in erring on the side of caution out of self-preservation. Before you know it, the distribution of human rights material could be prohibited as "propaganda."[12]

Second, allowing offensive ideas to be expressed verbally serves as an important safety valve against the expression of such ideas by means of physical violence. If we consider expression the middle stage between thought and action, this is also the stage at which correction can take place: by means of vigorous debate.

Third, and most important, we can't get closer to a functioning "marketplace of ideas" if the only ideas allowed into that marketplace consist of speech everyone agrees with or feels neutral toward. Not much would be left of the vigorous debate that provides the lifeblood of a democracy. In order to move forward as a society, we need dissenting voices; even ones that express their views in a way that may be offensive or shocking to others, however unpleasant that might be.

References

1 - Jenna Wortham, "Why Can't Silicon Valley Fix Online Harassment?," *The New York Times Magazine*, April 4, 2007, https://www.nytimes.com/2017/04/04/magazine/why-cant-silicon-valley-fix-online-harassment.html.

2 - Dunja Milatovic, "A Threat to Free Speech: The Online Abuse of Female Journalists," *Georgetown Journal of International Affairs,* February 29, 2016, http://journal.georgetown.edu/a-threat-to-free-speech-the-online-abuse-of-female-journalists/.

3 - Article 19, Handyside v. United Kingdom, *Columbia Global Freedom of Expression*, https://globalfreedomofexpression.columbia.edu/cases/handyside-v-uk/.

4 - Amal Clooney and Philippa Webb, "The Right to Insult in International Law," *Columbia Human Rights Law Review*, 48, no. 2 (2017), http://hrlr.law.columbia.edu/wp-content/uploads/sites/10/2017/03/Clooney-Webb.pdf.

5 - David L. Hudson, Jr., "Fighting words," *First Amendment Center*, November 5, 2003, http://www.firstamendmentcenter.org/fighting-words/.

6 - David L. Hudson, Jr., "True threats," *First Amendment Center*, May 12, 2008, http://www.firstamendmentcenter.org/true-threats/.

7 - American Civil Liberties Union, https://www.aclu.org/.

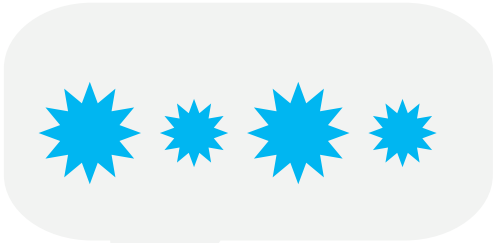
8 - Public Affairs, "Milo Yiannopoulos event canceled after violence erupts," UC Berkeley, February 1, 2017, http://news.berkeley.edu/2017/02/01/yiannopoulos-event-canceled/.

9 - Lee Rowland, interview by Lulu Garcia-Navarro, *Weekend Edition Sunday*, NPR, February 12, 2017, http://www.npr.org/2017/02/12/514785623/the-aclu-explains-why-theyre-supporting-the-rights-of-milo-yiannopoulos.

10 - U.N. General Assembly, *International Covenant on Civil and Political Rights,* adopted December 16, 1966, entered into force March 23, 1976, http://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx.

11 - U.N. Human Rights Committee, General comment No. 34: Article 19: Freedoms of opinion and expression, *International Covenant on Civil and Political Rights*, UN Doc CCPR/C/GC/34, http://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf.

12 - Miriam Elder, "Russia passes law banning gay 'propaganda,'" *The Guardian*, June 11, 2013, https://www.theguardian.com/world/2013/jun/11/russia-law-banning-gay-propaganda.



Can Cyber Harassment Laws Encourage Online Speech?

Jonathon W. Penney

Do laws criminalizing online harassment and cyberbullying "chill" online speech? My new study suggests, perhaps counter-intuitively, that such legal interventions may actually facilitate and encourage more speech, expression, and sharing by those who are most often the targets of online harassment: women.[1]

The study involves a first-of-its-kind online survey administered to 1,212 U.S.-based adult internet users that examines multiple dimensions of chilling effects online. It does so by comparing and analyzing responses to hypothetical scenarios that involve different kinds of regulatory actions—including an online harassment law, public/private sector surveillance, and an online regulatory scheme based on the Digital Millennium Copyright Act (DMCA) and enforced through personally received legal notices. The survey sample was roughly representative of the U.S. internet-using population, with a few biases, mainly being somewhat younger and having slightly lower incomes than the overall U.S. internet population.[2] Responses to each scenario were compiled, compared, and statistically analyzed.

Findings from the scenario involving the online harassment/cyberbullying law, which criminalized online speech intending to "harass or intimidate another person," and the scenario involving a personally received legal threat, each have implications for our understanding of online harassment and laws criminalizing it. This is important, for while these laws have proliferated across the United States and internationally,[3] little is known about the impact and efficacy of these legal interventions.[4]

First, the study found not only that online harassment and cyberbullying statutes may have far less overall chill on different online activities, at least compared with other forms of regulatory actions (like online surveillance),[5] but that these laws had a statistically significant salutary impact on women's willingness to share personal content online. This gender effect likely evidences that if women are aware of a law that penalizes or criminalizes online harassment and bullying, they feel less likely to be attacked or harassed and are thus more secure and willing to share, speak, and engage online. In other words, these statutes may actually lead to more speech, expression, and sharing online among adult women online, not less.

This is noteworthy because many question the effectiveness of these legislative efforts and their constitutionality—inasmuch as these laws often criminalize online speech, it is argued they prohibit or have a "chilling effect" on First Amendment protected speech.[6,7] And, indeed, courts have previously stricken down such laws on First Amendment grounds on "numerous occasions."[8] The analysis may change, however, if these laws actually lead to more speech and sharing online—especially from women, the traditional victims of these malicious activities online—while only minimally impacting other forms of speech and expression.

The study also found statistically significant gender effects in the hypothetical scenario in which respondents receive a personalized legal notice that contains a legal threat. Here, women were more likely to be chilled from engagement in a range of internet activities (online speech, search, and personal sharing) after receiving the personalized legal threat. Results suggested they were also more likely to be chilled in a scenario in which not they but a friend received a similar personal legal threat, and were less likely to take steps to defend themselves from the threatening legal notice that they had received. It is difficult to say, from the results, why women were more negatively affected in these scenarios, but the results no doubt suggest women were more cautious and chilled once they were personally targeted.

Besides being more often the victims of online harassment,[9] these findings suggest women may also be more affected by these harmful activities. This is consistent with recent findings by Lenhart et al. that women are more likely to be negatively affected—more likely to become angry, worried, and scared—as a result of online harassment and abuse.[10]

Of course, these findings do not obviate other important concerns about statutes criminalizing online harassment and cyberbullying, such as a lack of enforcement or disparate effects on other internet speech.[11] More research must be done on these counts to achieve a clearer picture. Still, the study's findings on these laws' potential salutary effect is consistent with what advocates like Danielle Citron have argued—that such legal measures can help preserve the "expressive autonomy" of internet users by facilitating more speech, participation, and sharing by frequent targets of such abuse.[12] And in light of the gendered impact of online harassment, these laws may also have the egalitarian impact of improving women's experience online generally.

References

1 - Jonathon Penney, "Internet surveillance, regulation, and chilling effects online: a comparative case study," *Internet Policy Review*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959611.

2 - Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis, "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making* 5, no. 1 (2010): 411-419. In short, the sample was very similar to previous ones recruited through Amazon's Mechanical Turk site.

3 - Jieun Baek and Lyndal M. Bullock, "Cyberbullying: A Cross-Cultural Perspective," *Emotional and Behavioural Difficulties* 19, no. 2 (2014): 226-238, http://www.tandfonline.com/doi/pdf/10.1080/13632752.2013.849028?needAccess=true.

4 - Aiman El Asam and Muthanna Samara, "Cyberbullying and the Law: A Review of Psychological and Legal Challenges," *Computers in Human Behavior* 65, (Dec. 2016): 138-139, http://www.sciencedirect.com/science/article/pii/S0747563216305775; Steven D. Hazelwood and Sarah Koon-Magnin, "Cyber Stalking and Cyber Harassment Legislation in the United States: A Qualitative Analysis," *International Journal of Cyber Criminology* 7, no. 2 (2013): 155-168, http://www.cybercrimejournal.com/hazelwoodkoonmagninijcc2013vol7issue2.pdf; Nicola Henry and Anastasia Powell, "Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research," *Trauma, Violence, and Abuse* 14, no. 1, (2016), http://journals.sagepub.com/doi/pdf/10.1177/1524838016650189.

5 - Penney, "Understanding the Comparative Dimensions of Regulatory Chilling Effects Online." For example, 62% of respondents indicated that the statute would either have no impact or render them more likely to speak/write online.

6 - See, e.g., Alice E. Marwick and Ross W. Miller, "Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape," *Fordham Center on Law and Information Policy Report*, 2014, http://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1002&context=clip; Dia Kayyali and Danny O'Brien, "Facing the Challenge of Online Harassment," *EFF Deeplinks Blog*, January 8, 2015, https://www.eff.org/deeplinks/2015/01/facing-challenge-online-harassment.

7 - See, e.g., Eugene Volokh, "Challenge to Maryland Law Banning Speech That Intentionally Seriously Distresses Minors," Volkh Conspiracy Blog, *Washington Post*, June 29, 2016, https://www.washingtonpost.com/news/volokh-conspiracy/wp/2016/06/29/challenge-to-maryland-law-banning-speech-that-intentionally-seriously-distresses-minors/?utm_term=.5c491cbf2b39; Fernando L. Diaz, "Trolling & the First Amendment: Protecting Internet Speech in the Era of Cyberbullies and Internet Defamation," *Journal of Law, Technology, and Policy*, (2016); Michal Buchhandler-Raphael, "Overcriminalizing Speech," *Cardozo Law Review* 36 (2015); Alison Virginia King, "Constitutionality of Cyberbullying Laws: Keeping the Online Playground Safe for Both Teens and Free Speech," *Vanderbilt Law Review* 63 (2010), https://education.ohio.gov/getattachment/Topics/Other-Resources/School-Safety/Safe-and-Supportive-Learning/Anti-Harassment-Intimidation-and-Bullying-Resource/Educator-s-Guide-Cyber-Safety.pdf.aspx. For a discussion of regulatory chilling effects theory and literature see Jonathon Penney, "Chilling Effects: Online Surveillance and Wikipedia Use," *Berkeley Technology Law Journal* 31, no. 1 (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2769645.

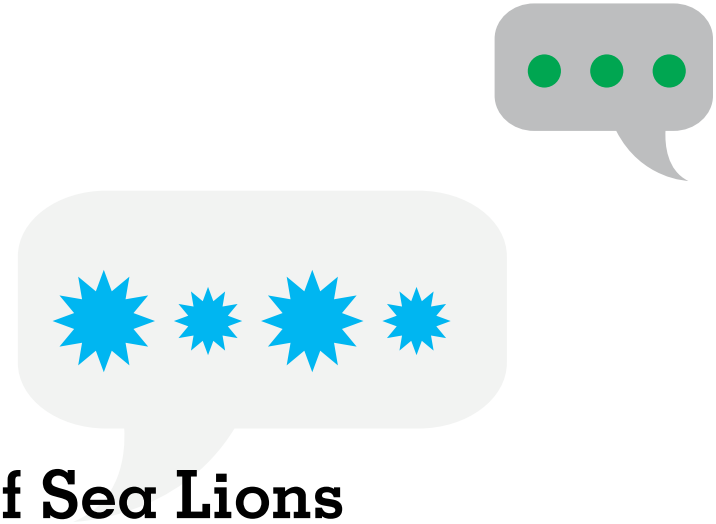
8 - Diaz, "Trolling & the First Amendment."

9 - Lenhart et al., "Online Harassment, Digital Abuse and Cyberstalking in America," *Data & Society Research Institute*, November 21, 2016, https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf; Marwick and Miller, "Online Harassment, Defamation, and Hateful Speech."

10 - Lenhart et al., "Online Harassment, Digital Abuse and Cyberstalking in America."

11 - See, e.g., Marwick and Miller, "Online Harassment, Defamation, and Hateful Speech"; Kayyali and O'Brien, "Facing the Challenge of Online Harassment."

12 - Danielle Keats Citron, *Hate Crimes in Cyberspace* (Cambridge: Harvard University Press, 2014).



The Multiple Harms of Sea Lions

Amy Johnson

"Where is the evidence for that opinion?"
…
"But doesn't [x] really mean [y]?"
…
"What about [other issue]—how do you explain that?"
…
"What's wrong with a polite question?"
…
"I'm just trying to engage in civil debate."

This series of questions may seem like a well-intentioned search for answers. It's not—it's a simplified example of a rhetorical strategy called sealioning. Sealioning is an intentional, combative performance of cluelessness. Rhetorically, sealioning fuses persistent questioning—often about basic information, information easily found elsewhere, or unrelated or tangential points—with a loudly-insisted-upon commitment to reasonable debate. It disguises itself as a sincere attempt to learn and communicate. Sealioning thus works both to exhaust a target's patience, attention, and communicative effort, and to portray the target as unreasonable. While the questions of the "sea lion" may seem innocent, they're intended maliciously and have harmful consequences. The ellipses in the sequence above stand in for multiple possible responses from targets, from lengthy explanations to pointing to logical fallacies in the questions themselves, from calling out the sealioning to ignoring it. It is these responses that the sea lion seeks to shape—and it is here that multiple harms occur.

Let me say first: we need a better term. "Sealioning" is both opaque and obscure. The term comes from a Wondermark cartoon by David Malki, entitled "The Terrible Sea Lion," published on September 19, 2014.[1] The cartoon captured interactions prevalent at the time in the context of Gamergate on platforms like Twitter. For those long familiar with the term, it thus immediately evokes Gamergate. The rhetorical strategy it describes, though, appears in multiple contexts, from Twitter to face-to-face conversations—and is hardly new.[2] Sealioning is a classic strategy of early trolling, but the meaning of "troll" itself con-

tinues to shift, so that's not much help. For the moment, I will stick with the term "sealioning," although I offer an alternative below.

Sealioning, which can be performed by a single user or as a tag-team effort, may feel familiar: it evokes the toddler who incessantly asks why, the adolescent who has just discovered philosophy, the condescending family member who disapproves of your life choices. This familiarity is part of its power. These interaction patterns summon a set of responses geared toward well-intentioned questioning. Sealioning also fits into a larger set of rhetorical marginalization practices. Refusals to understand can be subtle forms of erasure. Questions—shaped by explicit or implicit expectations about who has the right to question and who can be questioned about what—impose labor by demanding the questioned party either answer or appear indifferent; providing explanations and maintaining patience takes time and effort.

The harms of sealioning can seem relatively small: short-term annoyance when the practice is recognized, wasted energy when the practice goes unrecognized and the respondent gives sincere answers, the opportunity cost of the time spent. This assessment, though, comes from looking at each single instance or person targeted in isolation. Repeated experiences add up to larger social harms. The person targeted now doubts the sincerity of future questioners and becomes less inclined to engage in informal teaching. She or he is more likely to be curt and ignore others, to focus on broadcast communication rather than conversation. As a result, in future iterations the person who joins a conversation thread after someone has been sealioned will either recognize the sealioning or observe what appears to be hostility to rational, if misguided, discussion. Neither is likely to encourage trust in the ability to learn from one another. Meanwhile, the kind of person the sea lion was pretending to be—the sincere learner—no longer has as many opportunities to learn. Small harms affect personal habits, which in turn reshape larger social practices.

Informal teaching undergirds mediated communication. Informal teaching is an unacknowledged foundation of techno-utopian dreams from telegraphy to the present: by learning through interactions with each other, we will achieve universal understanding and eliminate conflict. And to some extent, this happens. At any one moment, informal teaching—about everything from platform norms and literacies to life experiences—bridges the hugely diverse skill sets and histories of people online. Who is online and the spaces they choose to inhabit perpetually change, yielding complex mixtures of experts and novices. Much informal teaching is necessarily repetitive and depends on good will. Informal teaching is vital, and sealioning attacks it.

As an alternative to the term "sealioning," we could simply refer to this practice as a different type of denial of service (DoS) attack—one aimed at humans rather than servers. Sealioning integrates social and technological manipulation to overload response capacity, tricking people into making an extensive, expensive effort that simultaneously prevents them from engaging elsewhere. On the one hand, sealioning capitalizes on civility and conversation norms to demand debate and labor. On the other, sealioning exploits threading capabilities and often launches through search. A social media platform with comprehensive search functions is a database in which every word is indexed, and every public word retrievable. Keyword searches thus become scouting tools for attacks. Even three years or so after Gamergate began, Twitter users still intentionally misspell Gamergate to avoid appearing in searches.

Even when sealioning is recognized, responding suitably can be difficult. There are no clear norms for handling it—advice tends to simply suggest "Don't feed the troll." While this may allow an individual to navigate the moment, it doesn't address broader effects on trust and learning.

In many ways, sealioning resembles the Gish gallop,[3] a rhetorical strategy that creationists deployed when debating evolutionists in the late twentieth century. The Gish gallop—named for Duane Gish, a biochemist who became a famed creationist debater—careens through topics, rattling through half-truth after half-truth. It aims both to overwhelm opponents' ability to respond and to introduce doubt into the minds of audiences. As a result, Eugenie C. Scott, anthropologist and former executive director for the

National Center for Science Education, advised evolutionists to avoid free-form debates. If debates simply had to be undertaken, Scott said, formal televised debates offered better spaces for argument—the focused structure reined in the Gish gallop.[4]

While superficially sealioning may seem but a mild annoyance, it undermines important social practices of trust and informal teaching. It's a complicated problem to address. As sealioning mixes the social and the technological, so too might solutions to it. Perhaps solutions can draw from Scott's advice, building or repurposing structures for focused interaction—something along the lines of Reddit's Ask Me Anything. Users might establish social norms of directing potential sea lions to such opportunities. From a linguistic perspective, sealioning is noticeably patterned—perhaps platforms could automate discovery of likely sealioning attempts and interrupt them. Better still, users and platforms could work together to redirect potential sea lions to structured learning opportunities—because sealioning attempts, after all, are teachable moments.

Many thanks to Lan Li and David Singerman for their thoughtful advice on this essay.

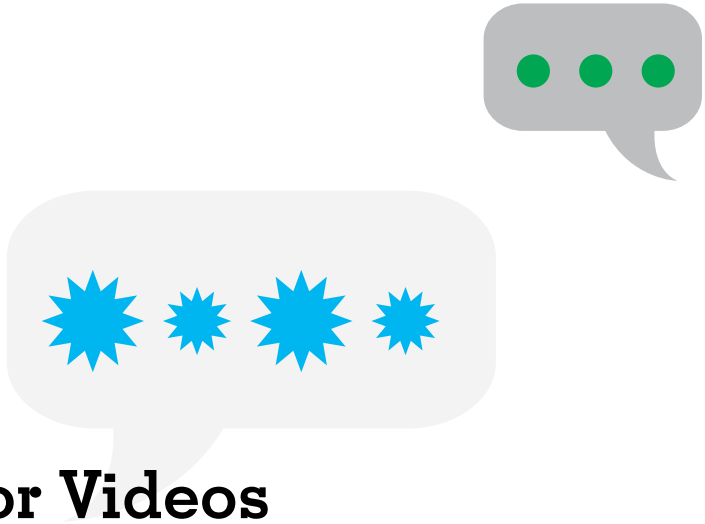
References

1 - David Malki, "#1062; The Terrible Sea Lion," *Wondermark*, September 19, 2014, http://wondermark.com/1k62/.

2 - See, for example, Tim Ferriss's best seller, *The 4-Hour Workweek: Escape 9-5, Live Anywhere, and Join the New Rich*, in which Ferriss describes intentionally seeking to exhaust the explanatory energy and patience of teaching assistants while he was a college student, in hopes that they would be more inclined to give him good grades in order not to have to respond to his excessive questions in the future. This example, which comes from the 1990s, lacks the element of performance to a larger audience.

3 - Eugenie Scott, "Debates and the Globetrotters," *The TalkOrigins Archive*, July 7, 1994, http://www.talkorigins.org/faqs/debating/globetrotters.html.

4 - Scott, "Debates and the Globetrotters."



Resharing of Images or Videos Without Consent: A Form of Relationship Violence and Harassment

Reynol Junco

The sharing of sexually explicit images, videos, and messages ("sexting") is a relatively common practice among people of different age groups.[1] People report that sexting can increase sexual communication and intimacy with a partner (especially in long-distance relationships), that it provides an outlet for sexual self-expression that can help people overcome inhibitions, and that many find it exciting, arousing, and fun.[2] Contrary to many media messages, adults aged 25 to 34 are more likely to sext than 18- to 24-year-olds.[3]

Can it go wrong?

You've probably read this story in the media: person A sends person B a sexually explicit image or video of herself/himself. Person B then reshares that image with a larger audience (e.g., because of anger and hurt after a breakup, to embarrass the person, or to express sexual prowess), possibly yielding catastrophic consequences for person A. Person A, the victim, may suffer psychological distress and social consequences (e.g., public embarrassment, being shamed, losing a job, having to drop out of school). Sometimes the consequences are even worse. There have been a number of cases in which a victim committed suicide because of reshared images and the resultant bullying.[4] While not directly suggesting blame, the traditional narrative tends to highlight primarily the terrible consequences of sexting, often arguing that if the person hadn't sent the sexually explicit image in the first place, he or she would have never been exploited.[5] Furthermore, the narrative rarely focuses on the person who reshared the images.

How does sexting relate to harmful speech?

The traditional narrative not only is wrong but can be seen as victim-shaming. Instead of focusing primarily on the victim's behavior, I believe we should devote much more attention to the behavior of the person who reshared the image or video without consent. Like other nonconsensual sexual acts, such resharing may be considered a form of relationship violence,[6] and it can include elements of harassment. According to the Women's Media Center, harassment may include a variety of tactics—including the resharing of images and videos without consent—that impact victims in legal, physical, emotional, and other consequential ways.[7] Moreover, there is research suggesting that victims of nonconsensual resharing experience offline threats, blackmail, and a host of negative psychological consequences, including an increased prevalence of anxiety and depression.[8]

Moving forward

It is essential that as a society we understand that sexting—although a risky online practice (see Palfrey & Gasser for challenges related to young people and sexting[9])—is relatively common and healthy, particularly among people who are dating.[10] Nonconsensual resharing, on the other hand, is I believe harassment. Adopting this narrative is no easy task, as it may challenge some of our society's beliefs about sexuality, especially women's sexuality. On the other hand, like other forms of abstinence-only education, suggesting people stop sexting to minimize the risks seems like a bad approach and will not work. Yet if we allow ourselves to focus on resharing as the problem, then we can begin to develop ways to tackle the challenges. For instance, relatively little research to date has focused on the characteristics of people who nonconsensually reshare sexually explicit images. Perhaps these people aren't as empathic as the ones who don't reshare. Perhaps they have more negative views about the other gender; perhaps these views might later predict physically abusive relationship behavior. There is so much we still do not know about who will nonconsensually reshare images, and until we know more, we will be unable to plan and implement appropriate strategies for mitigating and solving this form of relationship violence and harassment.

References

1 - Amanda Lenhart and Maeve Duggan, "Couples, the Internet, and social media: How American couples use digital technology to manage life, logistics, and emotional intimacy within their relationships," *Pew Research Center*, 2014, accessed April 13, 2017, http://www.pewinternet.org/2014/02/11/main-report-30/.

2 - Amy Adele Hasinoff, *Sexting panic: Rethinking criminalization, privacy, and consent* (Chicago, IL: University of Illinois Press, 2015).

3 - Lenhart and Duggan, "Couples, the Internet, and social media: How American couples use digital technology to manage life, logistics, and emotional intimacy within their relationships."

4 - Nina Burleigh, "Sexting, shame and suicide: A shocking tale of sexual assault in the digital age," *Rolling Stone*, September 17, 2013, accessed June 27, 2017, http://www.rollingstone.com/culture/news/sexting-shame-and-suicide-20130917; Randi Kaye, "How a cell phone picture led to girl's suicide," *CNN*, October 7, 2010, accessed June 27, 2017, http://www.cnn.com/2010/LIVING/10/07/hope.witsells.story/index.html.

5 - Hasinoff, *Sexting panic: Rethinking criminalization, privacy, and consent*.

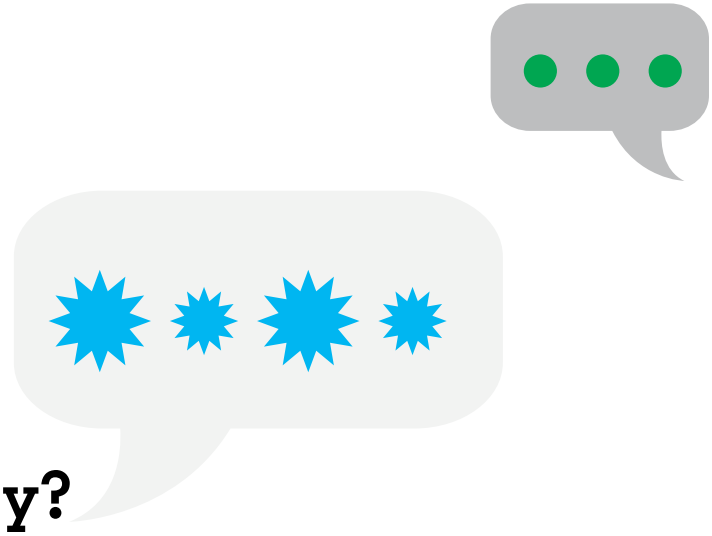
6 - Hasinoff, *Sexting panic: Rethinking criminalization, privacy, and consent*.

7 - "Online Abuse 101," *Women's Media Center*, accessed June 1, 2017, http://wmcspeechproject.com/online-abuse-101/.

8 - Karen Cooper, Ethel Quayle, Linda Jonsson, Carl Göran Svedin, "Adolescents and Self-Taken Sexual Images: A Review of the Literature," *Computers in Human Behavior*, 55 (2016): 706-716.

9 - John Palfrey and Urs Gasser, *Born digital: How children grow up in a digital age* (New York, NY: Basic Books, 2016).

10 - Cooper, Quayle, Jonsson and Svedin, "Adolescents and Self-Taken Sexual Images: A Review of the Literature." Hasinoff, *Sexting panic: Rethinking criminalization, privacy, and consent*. Lenhart and Duggan, "Couples, the Internet, and social media: How American couples use digital technology to manage life, logistics, and emotional intimacy within their relationships."



Goodbye to Anonymity? A New Era of Online Comment Sections

Casey Tilton

In May 2016, FoxNews.com ran a short article about Malia Obama's decision to attend Harvard University in 2017.[1] To say that the anonymous comments in response to the article were racist and hate-filled would be an understatement; some of the more civil comments accused the president's daughter of being a lucky recipient of affirmative action, while other comments were so shockingly toxic that the site soon decided to remove the comment section for the article altogether.

However, the article remained posted on Fox News' Facebook page in the days following publication. Presumably due to Facebook's real-name policy, the Facebook commenters used more civilized language—but in my opinion there was still a racist tone running through many of the top comments. The lifting of anonymity may have improved the civility of the discourse, but it had little effect on the sentiment of some of the commenters.

Fox News is just one of many websites that have taken steps to deal with the harmful speech that is all too common in anonymous comment sections. In an attempt to curtail the onslaught of harmful speech, some high-profile websites including CNN, Popular Science, and Reuters have in recent years removed their online comment sections.[2,3,4] Others require users to identify themselves before posting. USA Today, ESPN, and the Huffington Post, for example, require readers to log in to their Facebook accounts before they can contribute to a discussion.[5,6,7]

I believe it is unfortunate, at least from a free speech perspective, but understandable that news websites would rather remove their comment sections or integrate with Facebook than choose to spend the necessary resources to moderate their anonymous comment sections. After all, human moderation of online discussion sections is time-consuming and expensive.[8,9] Yet the decisions of these individual news outlets could have broader and potentially unintended implications for the future of online discourse. It is easy to imagine the real possibility of a future Internet largely devoid of anonymous discussions, where the great majority of online conversations take place in social media bubbles of like-minded communities.

While the move to social media and non-anonymous discourse can improve online behavior in many cases, some industry experts warn of the consequences of this societal shift.[10] A March 2017 study by the Pew Research Center,[11] which surveyed 1,500 industry stakeholders about the future of online conversations, discovered that many experts are wary of the migration to social media and the accompanying societal purge of user anonymity. Although most of the experts agree that non-anonymous online spaces are more inclusive and have generally more civil discourse, they worry that moving online conversations to non-anonymous social media will allow governments or other dominant institutions—like the social media platforms themselves—to pervasively monitor citizens, suppress free speech, and shape the social debate.

The decisions of individual organizations to remove anonymity from their comment sections are perfectly reasonable when considered independently, although it is unfortunate that an inexhaustible stream of anonymous trolls have forced news websites to make these decisions. However, I worry that we do not yet fully understand the aggregate impact these individual decisions will have on the future of online discourse. As mentioned above, centralizing conversations on non-anonymous social media platforms like Facebook could give governments or other dominant institutions even more power to surveil citizens and shape the social debate. But purging anonymity from portions of the public sphere may do more good than harm if reducing the level of harmful speech provides a less hostile and less exclusionary environment. Ultimately, I hope news websites and social media platforms alike continue to look for solutions that will rid the Internet of harmful speech while simultaneously preserving anonymous, colorful debate.

References

1 - "Malia Obama to Attend Harvard in 2017 after Taking Year Off," Associated Press, May 1, 2016, http://www.foxnews.com/politics/2016/05/01/malia-obama-to-attend-harvard-in-2017-after-taking-year-off.html.

2 - Doug Gross, "Online comments are being phased out," CNN, Nov. 21, 2014, http://www.cnn.com/2014/11/21/tech/web/online-comment-sections/.

3 - Suzanne LaBarre. "Why We're Shutting Off Our Comments," Popular Science, Sept. 24, 2013, http://www.popsci.com/science/article/2013-09/why-were-shutting-our-comments.

4 - "Editor's note: Reader comments in the age of social media," Reuters, Nov. 7, 2014, http://blogs.reuters.com/great-debate/2014/11/07/editors-note-reader-comments-in-the-age-of-social-media.

5 - "USA TODAY Network Conversation Guidelines," USA Today, http://static.usatoday.com/conversation-guidelines/.

6 - Marie Shanahan, "More News Organizations Try Civilizing Online Comments with the Help of Social Media," Poynter, July 16, 2013, http://www.poynter.org/2013/more-news-organizations-try-civilizing-online-comments-with-the-help-of-social-media/218284/.

7 - Joseph Lichterman, "Want to comment on a Huffington Post article? You'll need to use Facebook now," NiemanLab, June 2, 2014, http://www.niemanlab.org/2014/06/want-to-comment-on-a-huffington-post-article-youll-need-to-use-facebook-now/.

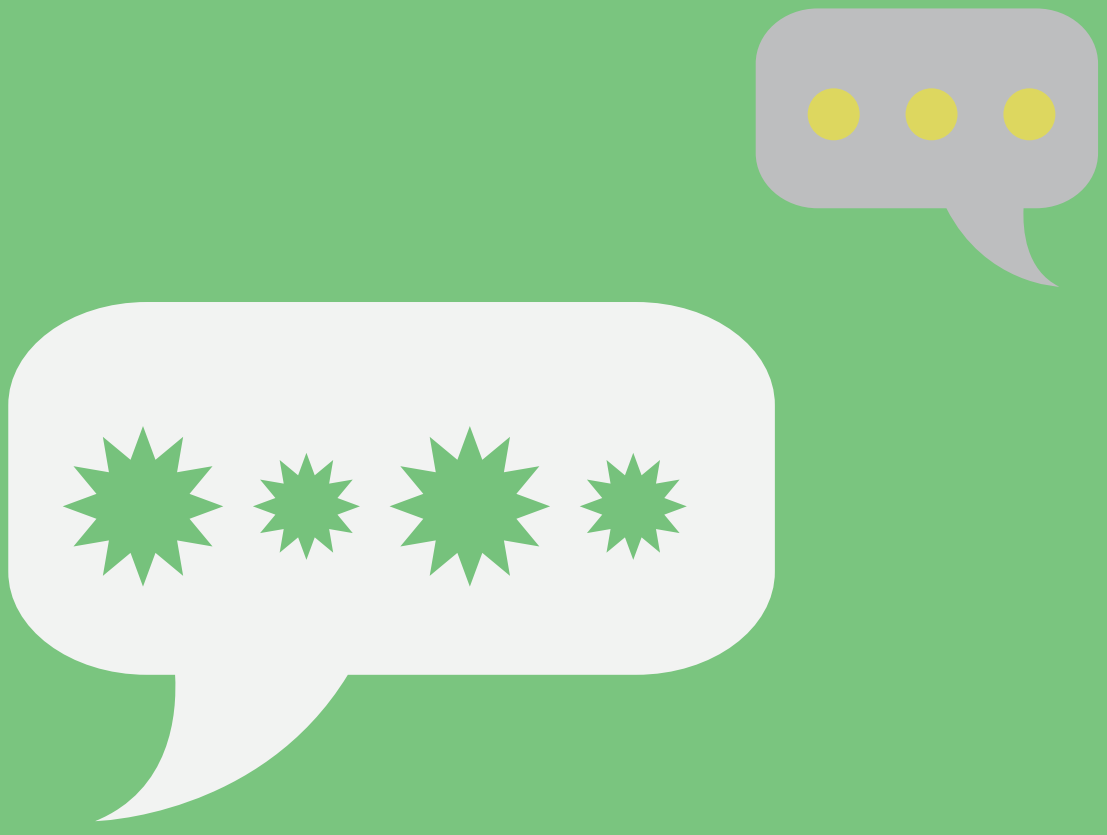
8 - "The Times is Partnering with Jigsaw to Expand Comment Capabilities," New York Times, Sept. 20, 2016, http://www.nytco.com/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/.

9 - Julia Franz, "Human moderators do the dirty work of keeping disturbing content off the internet," PRI, March 26, 2017, https://www.pri.org/stories/2017-03-26/human-moderators-do-dirty-work-keeping-disturbing-content-internet.

10 - Lee Raine, Janna Anderson, and Jonathan Albright, "The Future of Free Speech, Trolls, Anonymity and Fake News Online," Pew Research Center, March 29, 2017, http://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/.

11 - Raine, Anderson, and Albright, "The Future of Free Speech, Trolls, Anonymity and Fake News Online."

International Perspectives



Pakistan's Blasphemy Law: Using Hate Speech Laws to Limit Rights Online and Offline

Nighat Dad and Adnan Chaudhri

Among Muslim nations, Pakistan has a blasphemy law that is regarded as one of the most strict, going so far as to carry the death sentence should the state determine that certain actions or words on the part of the accused meet the criteria for blasphemy, whether by "innuendo, or insinuation, directly or indirectly."[1] Through such broad and ambiguous wording, the law forbids anti-religious speech both offline and online, the latter owing to the August 2016 passage of the Prevention of Electronic Crimes Act, which contains provisions to tackle the "glorification" of loosely defined "hate speech" online.[2] As a result, these ambiguous laws and the ways in which they are applied limit personal and digital freedoms.

Often, implementation of the blasphemy law is carried out extrajudicially, with critics of the law having long regarded the legislation as providing justification or carte blanche for horrific grassroots attacks.[3] The government of Pakistan claims that no one has yet been executed for blasphemy in Pakistan, and yet increasing religious conservatism and intolerance in Pakistan have ensured that merely to be accused of blasphemy is to be at risk of grave physical danger beyond the purview of the institutional justice system. At the most severe end of the spectrum, accusations alone can sometimes lead to death at the hands of a mob, especially if the accused is a member of a religious minority.

For example, on January 4, 2011, Salmaan Taseer, the governor of Punjab, was assassinated by Mumtaz Qadri, one of his bodyguards and a member of a counterterrorist division of the Punjab police force. Qadri said that he had killed Taseer because of his criticisms of Pakistan's blasphemy legislation, and because Taseer had supported Asia Bibi, a Christian Pakistani who was convicted of blasphemy in 2010. In the wake of the news of the murder, there was widespread horror and condemnation both in Pakistan and overseas, but in some corners of Pakistan Qadri was celebrated and showered with flower petals. Some religious leaders in Pakistan even went on record that they "salute the bravery, valor and faith of

Mumtaz Qadri."[4] Executed on February 29, 2016, Qadri remains popular in death. His grave is visited by hundreds per day, and a mosque that was built to honor him raised funds in 2014 in order to "double its capacity."[5]

Such extreme responses to accusations of blasphemy also occur in the digital world. Social media platforms are popular with Pakistanis as a means of keeping in touch with friends and family, but they are also a vital space for dissenting voices, ethnic and religious minorities, and the LGBT community, especially now that physical spaces for free speech are shrinking. However, in 2015 and 2016 at least three people were given 13-year prison sentences for allegedly posting material of a blasphemous nature on Facebook. According to others, their only crime was to "Like" posts of a critical nature. Further, the law gives rise to platform censorship; news reports claim that Facebook has extensively blocked "illegal blasphemous content on Pakistan's request"[6] from its site, with "85 per cent of such material" being removed permanently. What's more, members of Pakistan's government have called for a clampdown on "blasphemy gangs" that post blasphemous content online[7] and have demanded that Pakistanis involved in these gangs overseas be extradited back to Pakistan to face blasphemy charges.[8]

The following examples suggest that, as a result of the ambiguous nature of laws and their one-sided application and enforcement, open and free discourse is under threat. The government of Pakistan continues to seek out perpetrators of broadly defined and thus ambiguously "blasphemous" online content, shrinking space available for non-traditional ideas and content, and leaving many vulnerable groups in possible danger. Specifically, critics of the blasphemy law assert that it has been used to intimidate and stifle legitimate criticism. In January 2017 at least five Pakistani activists were picked up by plainclothes intelligence officers. The officers had received a complaint that alleged the activists, who had been critical of the government on social media, had spread blasphemous content on social media.[9] As online and offline news outlets started circulating the story, protests calling for the return of the activists were launched. The activists were soon freed, but the events had a chilling effect on public efforts and activism. Further, on April 11, 2017, Pakistan's Federal Investigation Agency launched a probe into Pakistani and international nongovernmental and civil society organizations that it claimed are "funding, supporting and disseminating blasphemous content and related activities in the country;"[10] this will likely also chill free speech.

The irregularity with which the government applies the blasphemy law is beginning to encourage emulation by the public. On April 13, 2017, a university student was murdered by a mob made up of fellow students after he was accused of blasphemy. The student, Mashal Khan, had been vocal regarding claims of corruption by university staff and faculty. After his death, it was discovered not only that the allegations of blasphemy against Khan were false[11] but that staff members at the university may have been responsible for the blasphemy claim.[12]

The conflicting values and social divide over Pakistan's blasphemy law are a prime example of the struggle to protect freedom of expression in the face of legislation that supposedly seeks to do exactly that by preventing "hate speech." The restrictions that warrant removal of content that many find deeply offensive also are used as a pretense for restricting human rights, open discourse, and dissent. A progressive balance that proactively protects freedom of speech while taking a practical and evidence-based approach toward defining and implementing harmful speech legislation is not only needed but vitally necessary, to ensure that democratic discourse online and in the real world is not silenced.

References

1 - Pakistan Penal Code (Act XLV of 1860), C.295.

2 - Danny O'Brien, "Global Ambitions of Pakistan's New Cyber-Crime Act," *Electronic Frontier Foundation*, August 18, 2016, https://www.eff.org/deeplinks/2016/08/global-ambitions-pakistans-new-cyber-crime-act.

3 - "Pakistan: How the blasphemy laws enable abuse," *Amnesty International*, December 21, 2016, https://www.amnesty.org/en/latest/news/2016/12/pakistan-how-the-blasphemy-laws-enable-abuse/.

4 - Robert Mackey, "Pakistani Lawyers Shower Murder Suspect With Roses," The Lede, *The New York Times*, January 5, 2011, https://thelede.blogs.nytimes.com/2011/01/05/pakistani-lawyers-shower-murder-suspect-with-roses/.

5 - Jon Boone, "Pakistan mosque built to honour politician's killer to double in size," *The Guardian*, April 30, 2014, https://www.theguardian.com/world/2014/apr/30/pakistan-mosque-killer-mumtaz-qadri-salaman-taseer.

6 - Aamir Jami, "Facebook removed 85% of blasphemous material on Pakistan's request, high court told," *Dawn*, March 27, 2017, https://www.dawn.com/news/1323131.

7 - "FIA traces '11-member gang of blaphemers,'" *The Nation*, March 16, 2017, http://nation.com.pk/national/16-Mar-2017/fia-traces-11-member-gang-of-blasphemers.

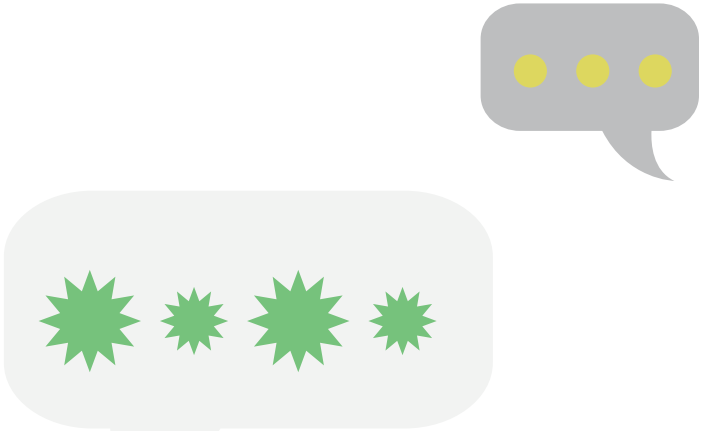
8 - Alex Hern and agencies, "Pakistan asks Facebook and Twitter to help identify blasphemers," *The Guardian*, March 17, 2017, https://www.theguardian.com/world/2017/mar/17/pakistan-asks-facebook-twitter-help-identify-blasphemers.

9 - "Pakistan: Bloggers Feared Abducted," *Human Rights Watch*, 10 January 2017, https://www.hrw.org/news/2017/01/10/pakistan-bloggers-feared-abducted.

10 - Riazul Haq, "FIA kicks off probe against NGOs funding blasphemous content in Pakistan," *The Express Tribune*, April 11, 2017, https://tribune.com.pk/story/1380163/fia-kicks-off-probe-ngos-promoting-sacrilege/.

11 - Mohammad Ashfaq, Inamullah Khattak, Ali Akbar and Hassan Farhan. "No evidence found to suggest Mashal Khan committed blasphemy: CM Khattak," *Dawn*, April 15, 2017, https://www.dawn.com/news/1327151.

12 - Ali Akbar. "University employee arrested in Mashal lynching case," *Dawn*, April 23, 2017, https://www.dawn.com/news/1328747.



State Power and Extremism in Europe: The Uneasy Relationship Between Governments and Social Media Companies

Kate Coyer

Terrorism and extremism are too often used as justifications for governments to expand state surveillance and content regulation powers in the name of national security. These restrictions reveal themselves more overtly in repressive regimes like Turkey, Egypt, and Russia, but trends across the European Union demonstrate a range of governmental efforts taking advantage of the current climate to seize far-reaching powers with troubling implications for civil liberties and free expression online.

One of British Prime Minister Theresa May's immediate responses to recent terror attacks in the United Kingdom was to call for tighter internet regulations. However, the 2016 U.K. Investigatory Powers Act extends state surveillance over online communications by unprecedented measures, granting government expansive unchecked powers without judicial oversight and undermining fundamental source protections for journalists in the name of national security.[1] The act authorizes bulk data collection, requires internet and phone companies to collect and store browsing histories for 12 months—including lists of every website each internet user has visited—and gives law enforcement broad access to the data.

One month after the attacks at Charlie Hebdo, the French government issued a decree granting the state powers to block websites accused of promoting terrorism without a court order, bypassing the court system and thus subverting due process and the rule of law, instead placing police in the position of content regulators.[2] The regulations had been under consideration since 2011 but were swiftly introduced follow-

ing the Paris attacks. The government argued that companies like Facebook and Google were otherwise "accomplices" to terrorism. Digital rights group La Quadrature du Net argued: "The measure only gives the illusion that the State is acting for our safety … while going one step further in undermining fundamental rights online."[3] Since then, state orders for content removal have doubled in France.[4] Last year, the French government also passed a law criminalizing visiting a "terrorist website," legislation criticized because it presupposes intent.

The relationship between government and social media companies is particularly fraught in authoritarian countries. In Russia, anti-extremism and internet laws introduced over the last few years have given the state far-reaching powers to control online communication and curb political speech. Andrei Bubeyev was sentenced to over two years in prison under Russia's extremism laws for posting an innocuous image on Russian-based social media platform VKontakte with the caption "Crimea Is Ukraine."[5] The increasingly authoritarian Turkish government, following the failed coup in July 2016, has come down heavily on dissent and oppositional voices online. The Turkish Ministry of the Interior reports that from August 2016 to January 2017, 10,000 people were under suspicion and 1,656 social media users were arrested on suspicion of spreading terrorist propaganda and insulting state officials. Pianist and supporter of the 2013 Gezi Park protests Dengin Ceyhan was arrested in winter 2017 for posts on social media that allegedly insulted the president. President Erdoğan has regularly blocked access to social media platforms and communication apps in the name of public safety, including access during the arrest of opposition members of Parliament.

In an effort to stave off potentially overreaching regulations, the technology companies Facebook, Google, Twitter, and Microsoft agreed upon a Code of Conduct on Illegal Online Hate Speech with the European Commission. The companies committed to take the lead on countering the spread of illegal hate speech online and to review removal requests within 24 hours. In the EU's announcement, EU Commissioner Věra Jourová said, "The recent terror attacks have reminded us of the urgent need to address illegal online hate speech."[6] Numerous digital rights organizations voiced strong concerns about the agreement, in no small part because the code was developed with no public input and without adequate involvement from civil society.[7]

The code raises several concerns: the implication that states are delegating enforcement to companies, and the likelihood of corrosive impacts on proportionality and due process. Without assurances that companies will increase their content review staff to manage the process, this will likely result in over-compliance and removal of legitimate speech. It's also debatable whether or not an EU initiative that encourages private companies to restrict freedom of expression without sufficient safeguards would stand under the scrutiny of the European Charter on Fundamental Rights, through which it could be labeled a type of "state interference by proxy."[8]

Also of concern is that companies are being asked to define extremism while there is no internationally recognized definition or common standards across social media platforms. Even the category of "hate speech" lacks consistency across the European Union.[9] A handful of European countries (Italy, Austria, and Denmark) still have blasphemy laws (and Ireland only recently enacted one), despite international human rights efforts to end such statutes.

Countering violent extremism and hate speech are closely linked, but companies need to resist government pressures to broadly define and overregulate hate speech; regulatory responses should be bal-

anced with the protection of civil liberties; and the public should become more vigilant to recognize and object to both corporate and governmental overreach. We need to be sure our conversations about extremism look inclusively at where the greatest threats to safety and security lie, including right-wing extremism and white supremacy, and the emergent intentional and unintentional harms.

References

1 - UK Government, "Investigatory Powers Act 2016," accessed on April 11, 2017, http://www.legislation.gov.uk/ukpga/2016/25/contents/enacted/data.htm.

2 - Republique Francaise, "Décret n° 2015-125 du 5 février 2015 relatif au blocage des sites provoquant à des actes de terrorisme ou en faisant l'apologie et des sites diffusant des images et représentations de mineurs à caractère pornographique," accessed April 11, 2017, https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000030195477&dateTexte=&categorieLien=id.

3 - Amar Toor, "France Can Now Block Suspected Terrorism Websites Without a Court Order," *The Verge*, Feb 9, 2015, accessed April 11, 2017, https://www.theverge.com/2015/2/9/8003907/france-terrorist-child-pornography-website-law-censorship.

4 - Suzanne, "France Sees Sharp Rise in Blocked and De-listed Websites," *Global Voices Advox*, March 6, 2017, accessed April 11, 2017, https://advox.globalvoices.org/2017/03/06/france-sees-sharp-rise-in-blocked-and-de-listed-websites/.

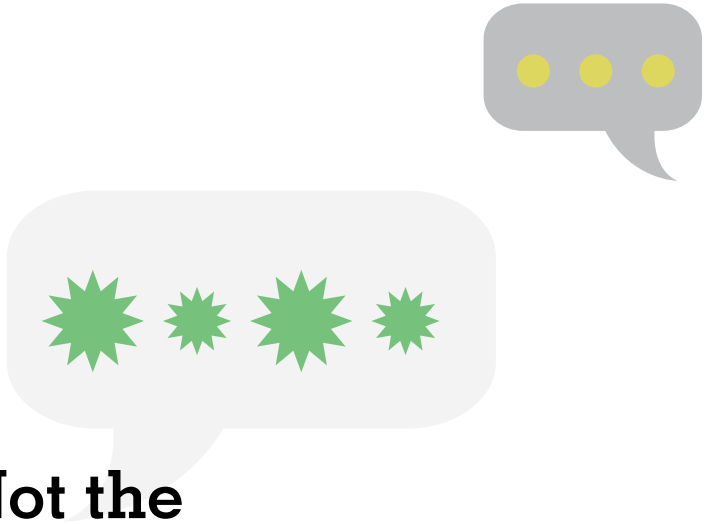
5 - Stephen Ennis, "Russian web activists face increased jail threat," *BBC News*, February 25, 2016, accessed April 19, 2017, http://www.bbc.com/news/world-europe-35651119.

6 - European Commission, "European Commission and IT Companies announce Code of Conduct on illegal online hate speech," May 31, 2016, accessed April 11, 2017, http://europa.eu/rapid/press-release_IP-16-1937_en.htm.

7 - EDRI, "EDRi and Access Now withdraw from the EU Commission IT Forum discussions," May 31, 2016, accessed April 11, 2017, https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions/.

8 - Aleksandra Kuczerawy, "The Code of Conduct on Online Hate Speech: an example of state interference by proxy?," *KU Leuven*, July 20, 2016, accessed April 11, 2017, https://www.law.kuleuven.be/citip/blog/the-code-of-conduct-on-online-hate-speech-an-example-of-state-interference-by-proxy/.

9 - Mandola Project, "Monitoring and Detecting Online Hate Speech," *Rights, Equality and Citizenship (REC) Programme of the EU Commission*, March 31, 2016, accessed on April 5, 2017, http://mandola-project.eu/m/filer_public/7b/8f/7b8f3f88-2270-47ed-8791-8fbfb320b755/mandola-d21.pdf.



Internet Shutdowns: Not the Answer to Harmful Speech Online

Grace Mutung'u

Although the Kenyan Constitution enshrines freedoms of conscience, expression, media, and access to information, freedom of expression is limited. It does not "extend to propaganda for war, incitement to violence, hate speech or advocacy of hatred that constitutes ethnic incitement, vilification of others, incitement to cause harm or discrimination."[1] Hate speech is defined in the National Cohesion and Integration Act as the use or display of threatening, abusive, or insulting words or behavior with the intent of stirring ethnic hatred.[2] There is also a Code of Conduct under the Elections Act (2011) that, beyond forbidding parties and candidates from acts of violence or intimidation, requires them to "condemn, avoid and take steps to prevent violence and intimidation."[3]

Kenya will hold its second General Election under the new Constitution on 8 August 2017. Traditionally, elections are highly contested for many reasons; among them, leaders wield considerable power over who has access to economic opportunities, and voters are mobilized along ethnic lines. Since the reintroduction of multipartyism in 1991, there has been violence with every election.[4]

Aided by social media tools such as Whatsapp, Twitter, and Facebook, Kenyans have vibrant, but also sometimes abrasive, conversations online during election periods. The government, wary of the country's deep history of violence surrounding election time, considers such dialogue as hate speech online, and responds to it in two ways: several agencies monitor online communications, and, should social media become "unmanageable," the sector regulator has warned of its willingness to induce an Internet shutdown.[5] This essay discusses three aspects of hate speech online and argues that enforcement of existing law would be more effective than stifling rights and shutting down the Internet.

Considering Hate Speech Online in Kenya

Three aspects of hate speech online in Kenya must be considered to fully grasp the problem's intricacies. First, academics, activists, artists, and thought leaders use the internet to speak in ways that could be considered provocative, especially regarding societal issues such as governance and historical injustices.

While such speech is not intended to cause harm, it often receives emotional comments that ultimately degenerate into tribal disparagement. For example, David Ndii, a writer, has been the target of several social media campaigns dubbed #DavidNdiiExposed for his articles, such as one that interrogates the notion of Kenyan nationhood.[6] Such campaigns generate heated debates that have tribal undertones.

Second is the nexus among political messaging, ethnicity, and hate speech online. Kenyans carry antagonism built up during campaigning into the "governance" period,[7] and this often surfaces online. After 2013 elections, the presidency set up the Presidential Strategic Communications Unit (PSCU), which actively engages the opposition politically on social media.[8] The conversations are sometimes laced with vilifying statements that encourage a culture of reckless discourse and increase polarization.

A third consideration is the role of law enforcement in cases involving hate speech online. Two concerns arise: the first is the weak prosecution of high-profile offenders under existing laws, and the second is increased online surveillance. Since the run-up to the 2013 election, the public has been keen to call out content deemed hate speech,[9] particularly from politicians, resulting in a few high-profile prosecutions. The most notorious is "Pangani 6," in which eight parliamentarians from both coalitions were arrested for hate-mongering during political rallies.[10] Despite video evidence of their utterances, which have been shared widely online, these and other such cases have been collapsing for lack of evidence or other technicalities. While Susan Benesch argues that the definition of hate speech in the Cohesion Act is ambiguous,[11] weak prosecution has denied the judiciary an opportunity to interpret aspects of hate speech online.[12]

Alongside weak prosecution, law enforcement has also increased surveillance and monitoring, reminiscent of the 1980s.[13] The National Cohesion and Integration Commission hired social media monitors, while the Communications Authority procured social media monitoring tools.[14,15] There are no official reports on the nature and benefits of the surveillance but Privacy International links communication surveillance to extrajudicial killings.[16]

Toward 2017 Elections

Societal issues that arise due to hate speech require multilayered interventions, most of which are part of constitutional implementation. The role of state agencies is to protect space for speech—not diminish it. For instance, in April 2017 the main political parties nominated candidates, but the process was laden with logistical and functional inefficiencies.[17] With contested results in the nominations, the availability of communications infrastructure gave the public an outlet to express their views about democracy and the flaws in the current system. This demonstrates the necessity for online connectivity during periods such as elections (an unrelated three-hour national blackout of the main mobile network operator, Safaricom, was reported).[18]

Non-state actors must continue to aid the public in contributing meaningfully and civilly to political discussions. This could be done by creating processes of accountability for political speech, as well as peace-building efforts such as those in the 2013 election.[19] These include civic education and opening debates on issues such as land and corruption as healthy outlets for civil discussion. Self-regulation in community spaces, the media, and among service providers must also be encouraged.[20, 21]

Combating hate speech online is a collective effort, shutting down the internet should not be the answer to this problem. On the issue of enforcement of the law, the need for better coordination among agencies in the anti-hate-speech space cannot be overstated, as overall there is enough space within the law and judicial mechanisms to pursue perpetrators of hate speech online. When Kenyans play their part by identifying instances of incitement, state agencies should equally perform their role within the law.

Further, the Kenyan government has had a history of being overbearing and stifling freedoms. The nation is undergoing a rebirth through the new Constitution. Rebuilding the nation is a progressive task, but it also requires safeguarding the gains under the new dispensation. In the case of hate speech online, recourse to the law has not been tested enough to justify measures as drastic as shutting down the internet in response.

References

1 - Republic of Kenya, *Constitution of Kenya*, 2010.

2 - Republic of Kenya, *National Cohesion and Integration Act*, no. 12 of 2008, revised 2012.

3 - Z. Elisha Ongoya and Willis E. Otieno, *Handbook on Kenya's Electoral Laws and System*, (Nairobi: Electoral Institute for Sustainable Democracy in Africa, 2012), http://aceproject.org/ero-en/regions/africa/KE/kenya-handbook-on-kenyas-electoral-laws-and-system.

4 - Maina Kiai, "Speech, Power and Violence: Hate Speech and Political Crisis in Kenya," *United States Holocaust Memorial Museum*, April 23, 2010, https://www.ushmm.org/m/pdfs/20100423-speech-power-violence-kiai.pdf.

5 - "Kenya's Communication Authority May 'Block Internet' During Elections," *PC Tech Magazine*, January 13, 2017, http://pctechmag.com/2017/01/kenyas-communication-authority-may-block-internet-during-elections/.

6 - David Ndii, "Kenya Is a Cruel Marriage, It's Time We Talk Divorce," *Daily Nation*, March 26, 2016, http://www.nation.co.ke/oped/Opinion/Kenya-is-a-cruel-marriage--it-s-time-we-talk-divorce/440808-3134132-2i7ea3/index.html.

7 - Peter Kagwanja, "Africa Is Suffering a New Bout of Populism," *Daily Nation*, March 7, 2016, http://www.nation.co.ke/oped/Opinion/africa-is-suffering-a-new-bout-of-populism/440808-3465976-udmn59z/index.html.

8 - For instance, the Director of PSCU, aggressively engages with the public and opposition coalition leaders through Facebook and Twitter posts, sometimes abrasively. It is rumoured that the ruling coalition has a team of 36 bloggers commissioned to tackle negative posts against the presidency. Conversely, the opposition also has dedicated "bloggers" who defend negative press against their leaders.

9 - E.g., Ushahidi and iHub Research, *Umati Final Report*, Prevent Violent Extremism Research Portal, June 28, 2013, https://preventviolentextremism.info/sites/default/files/Umati%20Final%20Report.pdf.

10 - Vincent Agoya, "Court Detains Eight Politicians in Hate Speech Probe," *Daily Nation*, June 14, 2016, http://www.nation.co.ke/news/Court-detains-Cord-Jubilee-politicians-in-hate-speech-probe/1056-3249748-v5h34tz/index.html.

11 - Susan Benesch, *Countering Dangerous Speech to Prevent Mass Violence during Kenya's 2013 Elections*, February 9, 2014, https://www.ushmm.org/m/pdfs/20140212-benesch-kenya.pdf.

12 - For instance, R v Kioi (2013) where a musician was accused of inflammatory lyrics and R v Ngunyi (2016) where a popular political analyst was accused of publishing ethnic contempt. Both these cases would have provided an opportunity to interpret artistic and academic freedom against the limits to freedom of speech.

13 - Kevin Conboy, "Detention Without Trial in Kenya," *Georgia Journal of International & Comparative Law*, 8 (1978): 441-461, http://digitalcommons.law.uga.edu/cgi/viewcontent.cgi?article=2083&context=gjicl.

14 - Lynet Igadwah, "Agency Hiring Social Media Monitors to Track Hate Mongers," *Business Daily*, January 11, 2017, http://www.businessdailyafrica.com/news/Agency-hiring-social-media-monitors-to-track-hate-mongers/539546-3514258-11ffr3jz/index.html.

15 - Joseph Wangui, "Agency to Monitor Social Media Use," *Daily Nation*, January 29, 2017, http://www.nation.co.ke/news/social-media-use/1056-3790946-4kundt/.

16 - "Track, Capture, Kill: Inside Communications Surveillance and Counterterrorism in Kenya," *Privacy International*, March 15, 2017, https://www.privacyinternational.org/node/1366.

17 - Nic Cheeseman, "What Party Primaries Mean for Kenya's General Election," *Daily Nation*, April 30, 2017, http://www.nation.co.ke/oped/Opinion/-What-party-primaries-mean-for-Kenya-s-General-Election/440808-3908344-asokklz/index.html.

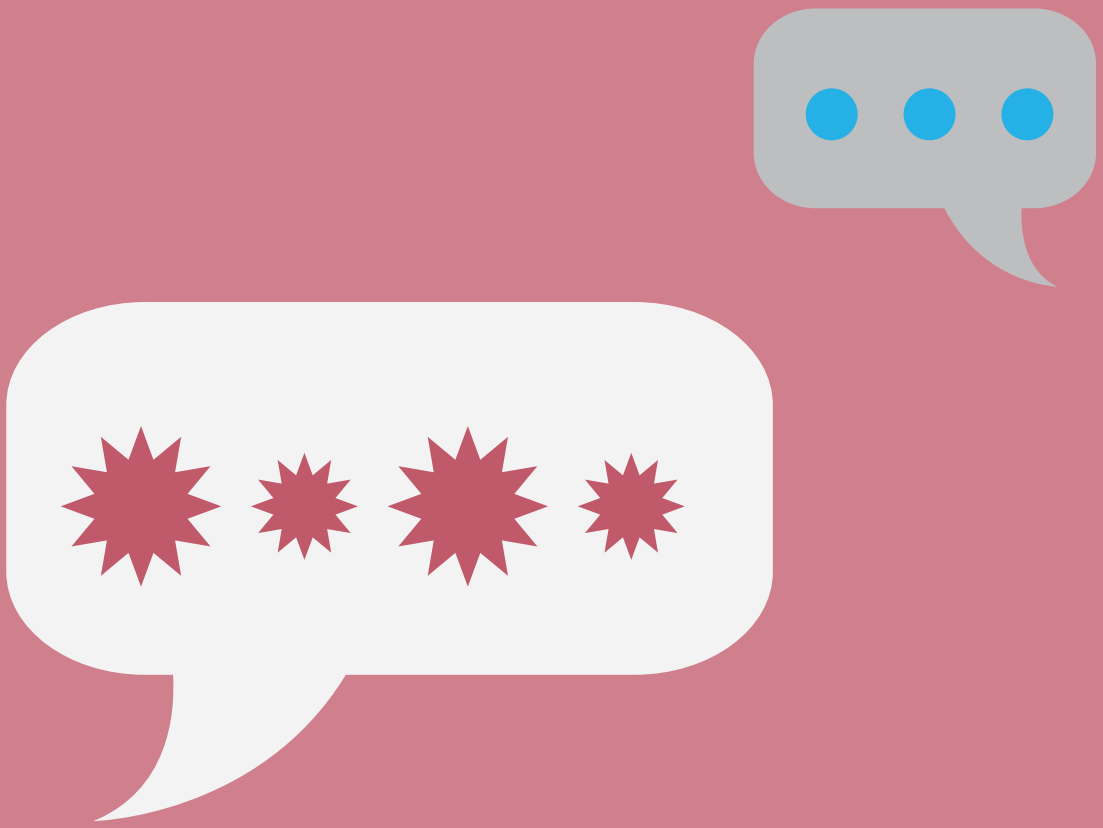
18 - Muthoki Mumo, "Safaricom Network Failure Paralyses Millions," *Business Daily,* April 24, 2017, accessed May 8, 2017, http://www.businessdailyafrica.com/corporate/Safaricom-outage-communication-blackout/539550-3901452-i30g2gz/.

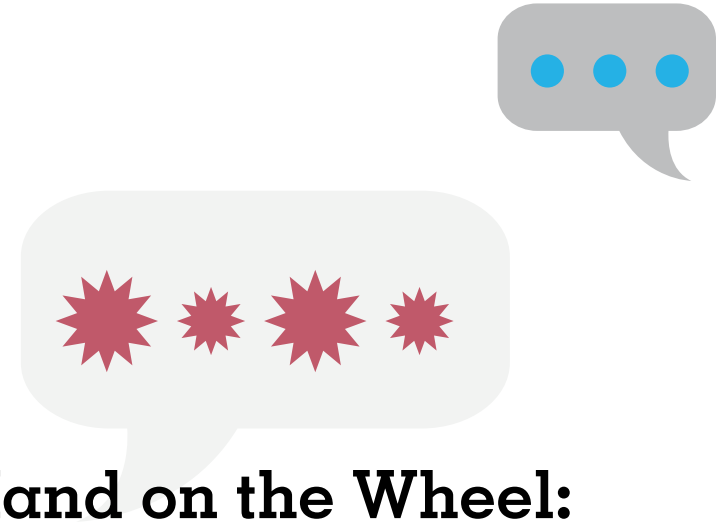
19 - For example, the Uraia Campaign, which is a joint civil society effort; Uraia Trust, http://uraia.or.ke/about-uraia-trust/.

20 - "How Media Covered 2013 Elections," Kenya Union of Journalists, May 12, 2012, http://www.kenyaunionofjournalists.org/how-media-covered-2013-elecions/.

21 - Charles Gichane, "Safaricom Issues Tough Rules on Political Messaging," *Capital News*, June 16, 2012, http://www.capitalfm.co.ke/news/2012/06/safaricom-issues-tough-rules-on-political-messanging/.

Approaches, Interventions, & Solutions





Civil Society Puts a Hand on the Wheel: Diverse Responses to Harmful Speech

Susan Benesch

A single response to harmful speech online—deletion, or "takedown" in industry parlance—is by far the most discussed and demanded, but other responses deserve notice, especially those that convince people to post less harmful speech.

Takedown focuses on the offending content alone—not on those who post it nor those who are harmed by it—so it doesn't do much to persuade people to stop thinking, speaking, or re-posting in harmful ways, and it doesn't remedy the harm done to people who are exposed to the content before it's removed. Takedown is an ever-expanding game of Whac-a-Mole[1]: it can't keep up with the staggering rate at which new content appears online, except perhaps if takedown becomes automated, algorithmic prior censorship, which tends to be overbroad and would infringe on freedom of speech.[2] Also, takedown is a method that can be practiced only by internet companies, making and applying their own internal rules for it, while governments pressure them to take down more content.[3]

Meanwhile, alternative responses to harmful speech online are being invented and tested by internet users and nonprofit organizations. Some of these efforts have persuaded people, albeit on a small scale so far, to stop posting harmful content. Other civil society responses sidestep both content and content producers, to succor the targets of harmful speech instead. Still others use humor to defuse harmful speech. Of the intriguing new civil society responses to harmful speech, a few of the most promising are described below.

A common thread among the methods is to take action offline, calling out harmful speech where people may be more susceptible to social pressure against it. If a boy or young man threatens to rape you, for example, you may get a speedy apology by telling (or merely threatening to tell) his mother. When Oliver Rawlings, then 20 years old, sent a highly offensive sexualized tweet to the University of Cambridge classics scholar Mary Beard, one of her Twitter followers offered Beard the mailing address of his mother.[4]

Rawlings recanted instead, and Beard later invited him to lunch in Cambridge. Alanah Pearce, an Australian game critic, found the mothers of several of her trolls on Facebook and forwarded rape threats their sons had sent her. One of the mothers required her son to handwrite a letter of apology and pressed his school to teach online safety.[5]

Not everyone relies on mothers: many outraged internet users contact employers instead, demanding that people be fired. Such demands have succeeded—and this type of effort often spills over into vigilantism and excessive punishment. In many cases, the response to harmful speech is as vitriolic and relentless as the speech it denounces.[6]

In light of these dangers, some campaigns against harmful speech have been thoughtfully calibrated to teach, while also protecting people from angry outrage. In Brazil, for example, the selection of Maria Julia Coutinho as the first black weather forecaster for the popular television news program *Jornal Nacional* in 2015 was met with a surge of online racism. The black women's rights organization Criola responded with a campaign called Mirrors of Racism, reproducing racist comments from Facebook on large billboards, with the slogan "Virtual Racism. Real consequences."[7]

Criola's director, Jurema Werneck, said the organization geolocated the authors of the racist comments by studying their social media presences and put the billboards in their own neighborhoods, so they would see their own words called out near their homes. But Criola chose not to name them. "We omitted names and faces of the authors—we had no intention of exposing them. We just wanted to raise awareness and start a discussion, in order to make people think about the consequences before posting this kind of comments on the internet," Criola wrote.[8] The authors of racist posts were protected from public attack, therefore, but were still exposed to the silent shaming of seeing their own words emblazoned in large letters—and called out as harmful speech.

As a result of the campaign, 83 percent of the commenters deleted their accounts, according to Criola—and one commenter came forward to apologize. "I could see just how racist I had really been, even though that wasn't really my intention," said Lucas Arruda, in front of his post on a billboard: "Cheguei a casa fedendo a preto" (I got home stinking of black people).

In some cases, the goal of a response to harmful speech is not to discourage that speech per se but instead to alleviate the harm felt by people at whom it is directly aimed. HeartMob is a web platform that allows the targets of harassment to recruit online "bystanders"—other people who are online at the same time—to help them during an attack by giving particular types of aid, such as sending the target supportive messages, documenting the abuse, and/or reporting the content to an internet platform. Emily May, who created HeartMob after she was inspired by activism to diminish street harassment of women, points out that on the street bystanders are not always present, but bystanders are always available online. Trollbusters, a platform similar to HeartMob, recruits "virtual S.O.S. teams" to send messages of support to the social media feed of a target of online harassment, in an attempt to drown it out.

Blockbots are one more tool that targets of harassment use to protect themselves against it, in this case by preventing themselves from seeing it at all. Blockbots are applications that allow Twitter users to block lists of other users—and to share those lists easily. When a user subscribes to a list, the bot uses the Twitter API to block the accounts on the list automatically, avoiding the tedious process of blocking each account individually. This innovation proved so useful for targets of harassment that Twitter built into its platform the option to import and export block lists in 2015.

These efforts use fundamentally different methods to diminish harm: from magnifying racist messages and displaying them, literally, in the sunlight to trying to leave them hidden in the dark corners of Twitter. Each of them deserves study, to determine its actual capacity to reduce harmful speech or the damage it does. Researchers must be careful, though, to protect themselves from targeted abuse in the course of their work.[9]

References

1 - Kate David Jones, "How to Fight Trolls Online," *Vice.com*, July 10, 2015, https://www.vice.com/en_us/article/how-to-fight-trolls-online-235.

2 - Haji Mohammad Saleem et al., "A Web of Hate: Tackling Hateful Speech in Online Social Spaces," (paper presented at the First Workshop on Text Analytics for Cybersecurity and Online Safety, Portorož, Slovenia, May 2016), http://www.ta-cos.org/sites/ta-cos.org/files/tacos2016_SaleemDillionBeneschRuths.pdf. This article explains that automated detection of harmful speech produces abundant false positives.

3 - See, e.g. European Commission, Code of Conduct on Countering Illegal Hate Speech Online, May 29, 2016, http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf. See also Automattic, Russian Censorship Details, https://transparency.automattic.com/russian-censorship-demands.

4 - Rebecca Mead, "The Troll Slayer,' *New Yorker,* September 1, 2014, http://www.newyorker.com/magazine/2014/09/01/troll-slayer; Sam Marsden, "Internet Troll Who Abused Mary Beard Apologises After Threat to Tell His Mother," *Telegraph*, July 29, 2013, http://www.telegraph.co.uk/news/uknews/law-and-order/10209643/Internet-troll-who-abused-Mary-Beard-apologises-after-threat-to-tell-his-mother.html.

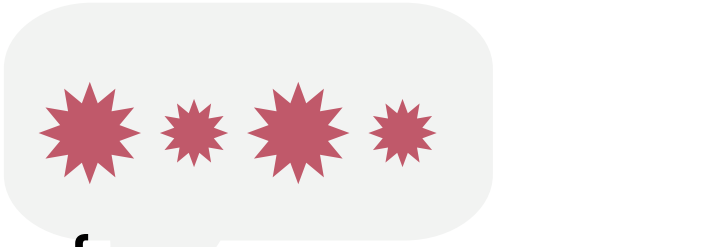
5 - Alanah Pearce, "How I Told On Kids Who Sent Me Rape Threats," *YouTube.com*, November 19, 2016, https://www.youtube.com/watch?v=b-J11JDm_XY.

6 - Jon Ronson, "How One Stupid Tweet Blew Up Justine Sacco's Life," *New York Times*, February 12, 2015, https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html.

7 - bossanovagroup, "Criola :: Mirrors of Racism," *vimeo.com*, June 30, 2016, https://vimeo.com/172953927.

8 - "Mirrors of Racism," Internet Archive, https://web.archive.org/web/20151209140417/http://www.racismovirtual.com.br/virtual-racism.

9 - Alice E. Marwick, Lindsay Blackwell and Katherine Lo, "Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment," *Data & Society Research Institute*, 2016, https://datasociety.net/pubs/res/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.



Moderation and Sense of Community in a Youth-Oriented Online Platform: Scratch's Governance Strategy for Addressing Harmful Speech

Andres Lombana-Bermudez

Online platforms and virtual worlds have become important spaces for youth development, socialization, and learning. Children and youth are growing up in a networked communication environment in which they are leveraging digital tools for expressing their creativity, seeking information, and building relationships. They are participating in "networked publics" of different sizes and themes, where they communicate with peers and mentors, share content they create, and engage in communal activities such as playing games and exchanging information about specific topics.[1,2,3,4]

Although engaging in online platforms and networked publics presents opportunities for learning, identity development, and networking, doing so also poses certain risks. Parents and other adults have raised concerns about the presence of harmful speech in these digital spaces, particularly cyberbullying. An extension of bullying behaviors in offline spaces, cyberbullying consists of the use of digital tools to harm others, and it has driven much of the discourse around child safety in mainstream media.[5,6,7,8]

Youth-oriented online platforms have approached harmful speech in different ways. While some heavily moderated platforms try to minimize risks by limiting opportunities for creative expression (e.g., pre-written chat messages as in Kart Kingdom and Club Penguin), other platforms take a "no holds barred" approach even if it results in mean and rude content (e.g., 4chan, MemeGenerator). In between these competing perspectives is Scratch, a youth-oriented nonprofit platform launched in 2007 by the Lifelong Kindergarten research group at the MIT Media Lab. This platform allows users to create and share inter-

active multimedia projects and to publish text-based messages (a form of asynchronous communication) across several spaces such as projects and studio comments, and discussion forums.

Scratch is a successful example of how an online community can reduce the incidence of harmful speech and foster civil dialogue while at the same time scaling up, and fostering youth's agency and creative expression.[9] The platform has implemented a governance strategy that combines proactive and reactive moderation (through content curation and filtering) with the cultivation of socially beneficial norms and a sense of community. This hybrid strategy has allowed Scratch to address harmful speech successfully, decreasing its incidence and prevalence. Particularly, it has allowed adult moderators and young community members to regulate uncivil behaviors such as spamming, harassment, and publishing mean, rude, and inappropriate or profane content.

Scratch Guiding Principles

Establishing clear, brief, and youth-friendly Community Guidelines has been key to cultivating a supportive and safe community. The guidelines lay out a set of core values or guiding principles that all members of the community share and follow. As one of the Scratch moderators explained to me during an interview for the Coding for All project, the Community Guidelines are easy to "absorb" and to "own" by youth of all ages.[10]

The guidelines encompass six short guiding principles: (1) be respectful; (2) be constructive; (3) share; (4) keep personal info private; (5) be honest; and (6) help keep the site friendly. All new users of the platform are encouraged to read the Community Guidelines when they join, and they receive automated messages that remind them of "commenting respectfully" when they start publishing their first comments and posts. If new users try to create a comment that uses language the system detects as unconstructive or inappropriate, they get an automated message that prevents them from posting; tells them that their comment "may be mean or disrespectful" and that they need to read the Community Guidelines; and includes a reminder to "be nice." The guidelines are available across the platform, accessible through a link that appears at the bottom of all pages.

Being respectful, keeping the site friendly, and being constructive, in particular, are core values that directly address harmful speech. Any content that breaks these values is taken down by a sophisticated moderation scheme that includes adult moderators, automated software filters, and young community members. According to the four Scratch moderators I interviewed, the most common instances of harmful speech on Scratch are spam comments, followed by mean and rude comments that are unconstructive and inappropriate (typically profanity or swearing). Although instances of hate speech and harassment are rare, if they appear they are censored right away by the moderation system. As one of the moderators explained to me, the few cases of cyberbullying that have appeared on Scratch are among users who know each other in real life and carry conflict from their school into the online platform.

Scratch's core values promote kindness and inclusion within a platform that is diverse in terms of users' age, ethnicity, sexual orientation, gender identity, and religion. They empower youth to engage in civil dialogue and to actively and responsibly participate in building a safe space. As the members of the community "absorb" the guiding principles, they actively engage in the dissemination of the guidelines on their own. Exercising their agency, Scratchers, as the members of the community call themselves, have designed multimedia projects that explain the Community Guidelines in creative ways (there are hundreds of projects dedicated to explaining the guidelines).

Tandem Moderation: Adult Moderators + Scratchers

In order to effectively manage growth and foster a sense of community, Scratch has deployed a governance and moderation system in which both adults and youth engage in regulating, monitoring, and enforcing the Community Guidelines while leveraging different sociotechnical tools, all against the backdrop of a high degree of transparency (all user-generated content is public). As the Scratch community manager noted during one of our interviews, "moderation is done in tandem with Scratchers."

The Scratch Team has 16 adult moderators, including one community manager and one community coordinator, all of whom actively moderate. These adults have full-time and part-time paid jobs with Scratch and moderate all the discourse that is generated on the platform. They do this with the help of automated software filters that detect harmful speech using a list of designated words and phrases, and also help from users who flag content that violates the Community Guidelines. As adult moderators review the content that has been flagged, they evaluate if it violates the Community Guidelines. If it does, the content is considered harmful speech, it is removed from the platform, and the moderators send a private alert to the user who created it. Moderators also ban users who repeatedly break the guidelines and, in rare cases, communicate with parents via email in order to restore an account that has been blocked.

Young members of the community contribute to the moderation scheme by flagging inappropriate content that is published on the platform. From project and studio comments to forum posts, all spaces where content is shared have buttons that Scratchers can use for reporting, and approximately 200-300 user reports are generated daily. Moreover, Scratchers also engage in moderation by using the Community Guidelines as tools for civil dialogue, referring to specific principles when commenting in peers' projects and studios, and citing guidelines when posting in discussion forums.

Conclusion

Scratch is a successful example of how governance strategies can foster safe, positive, and diverse youth-oriented online platforms, and reduce the incidence of harmful speech. The implementation of a hybrid strategy that combines active content curation and filtering with the cultivation of a sense of community has proven to be highly effective. Specifically, Scratch has thrived through two key dimensions: 1) establishing a moderation scheme in which both adult moderators and community members actively monitor the platform with the help of automated software filters, and 2) supporting community engagement through the adoption and championing of clear core values. This combined approach has been highly successful in reducing the incidence of harmful speech while simultaneously supporting youth agency, freedom of expression, learning, and creativity.

References

1 - Mimi Ito, et al, *Hanging Out, Messing Around, and Geeking Out* (Cambridge, MA: The MIT Press, 2009).

2 - danah boyd, *It's Complicated: The Social Lives of Networked Teens* (New Haven: Yale University Press, 2014).

3 - Yasmin Kafai and Deborah Fields, *Connected Play: Tweens in a Virtual World* (Cambridge, MA: MIT Press, 2013).

4 - Henry Jenkins, Mimi Ito and danah boyd, *Participatory Culture in a Networked Era: A Conversation on Youth, Learning, Commerce, and Politics* (Malden, MA: Polity, 2015).

5 - John Palfrey and Urs Gasser, *Born Digital: Understanding the First Generation of Digital Natives*, (New York, NY: Basic Books, 2008).

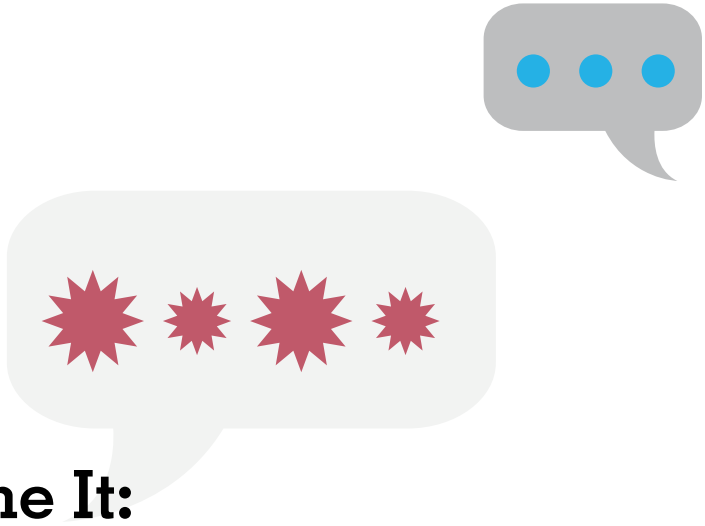
6 - Andrew Schrock and danah boyd, "Problematic Youth Interaction Online: Solicitation, Harassment, and Cyberbullying," *Computer-Mediated Communication in Personal Relationships*, eds. Kevin B. Wright and Lynn M. Webb (New York: Peter Lang, 2011).

7 - Sonia Livingstone, et al, "Children's Online Risks and Opportunities: Comparative Findings from EU Kids Online and Net Children Go Mobile," *EU Kids Online*, LSE (London, UK: 2014).

8 - Sameer Hinduja and Justin. W. Patchin, "What is Cyberbullying?" *Cyberbullying Research Center*, December 23, 2014, accessed February 2017, http://cyberbullying.org/what-is-cyberbullying.

9 - Today, 10 years after its launch, Scratch has grown to a user base of 17 million and is home to an enormous amount of text-based and multimedia content generated by children. The platform has been translated into over 40 languages and its users come from all around the world and are mostly between 8 and 16 years old. According to the Scratch Stats page (https://scratch.mit.edu/statistics/), during the month of February 2017, there were 822,667 New Projects, and 3,039,859 new comments. Since 2007, members of the community (17,777,432 registered users in total) have shared 21,594,894 projects, and posted 110,195,916 comments.

10 - Coding for All: Interest-Driven Trajectories to Computational Fluency, is a collaborative project between the MIT Media Lab, UC-Irvine's DML Research Hub, and the Berkman Klein Center for Internet & Society.



If We Own It, We Define It: The Dilemma of Self-Regulating Hate Speech

Helmi Noman

I examine in this essay the susceptibility of online communities to hosting harmful speech relative to their self-regulation dynamics. I review whether forum-based communities employ mechanisms to critically examine member contributions for potentially harmful speech and whether they introduce mitigation measures. I look into Arabic forum-based communities as an exploratory case study. I find that administrative ownership of the online space together with social norms and cultural prejudices implicate its self-regulation system in a regulatory dilemma: the administrators own both the forum and the definition of harmful speech. I conclude that forum-based communities need to introduce thoughtful deliberation of what constitutes harmful speech independent of their biases and apply cues of control to inhibit their in-group prejudices.

About the Forums

Arabic forum-based communities facilitate discourse on various issues. The forum's contributions are publicly accessible, but registration is required for contributors. Each forum has a thematic focus such as religion, politics, tribalism, technology, hacking, or entertainment. Forum owners register their forum's domain name, create the initial content and design, moderate member contributions, and continue to contribute content and shape the overall identity of their forum. They also assign administrative and moderation privileges to like-minded active members. Each forum has its own self-regulation guidelines, which reflect its community's understanding of what acceptable speech is, and are primed by the community's cultural inclination. Content that amounts to harmful speech appears in the form of sectarian and racial hatred, aggressive nationalism, tribal fanaticism and ethnocentrism, prejudice against non-Islamic faiths, and otherwise militant and extremist content. Many of the forums have a noticeably prejudiced outlook and in some cases a hostile attitude toward out-groups. This is particularly evident in religion-focused forums, especially those that are either Sunni or Shiite.

The Dilemma of Self-Regulation

Forum administrators have not formalized terms like harmful speech or hate speech in their self-regulation lexicon. Each forum's administrators provide broad guidelines on acceptable and offensive speech, and they neither cite independent attempts to define harmful speech nor consult national, regional, or international frames on the issue. The forums give their members the option to report offensive contributions to the administrators, but they do not present established mechanisms to critically examine member contributions for harmful speech. The space ownership and moderation dynamic seem to contribute to making each forum a host of a particular cultural inclination: The administrators contribute content and advance the forum cause or belief. At the same time, they assume the role of judges. They own the forum and hence the definition of harmful content; they draw the contours of what is admissible. This system seems to be behind the seemingly homogeneous behavior of each forum's members as voices challenging potentially harmful messages are rare and not audible enough to influence the policy on harmful speech. As a result, contributions do not go through a rigorous harmful speech test; rather, inclusion and exclusion of what amounts to harmful speech is subject to each community's interpretation.

The self-regulatory systems in these forums pose a dilemma because they do not articulate clear boundaries of inadmissible speech. The systems favor in-groups and the de facto enforcement policies condone posting of highly problematic content about out-groups. For example, religion-focused Sunni and Shiite forums mention in their contribution policies that they do not allow sectarian offenses, and in many cases they state exceptions to the limits of offensive speech such as evidence-based statements and arguments about out-groups. Despite these policies, inflammatory and harmful sectarian speech is common in these forums: members refer to the other sect using pejorative and demeaning terms, and report what they describe as scandals from the other sect's religious texts, historical symbols, and contemporary figures and practices.

In most cases, each forum's policy emphasizes in its guidelines respect to sanctity of life for all. And yet a religion-focused forum whose terms and conditions ban content that "violates human sanctuary" allows users to post remarks celebrating the murder of the Jordanian writer and political activist Nahed Hattar by an extremist in September 2016.[1,2] Hattar was killed on his way to attend a court hearing for sharing a caricature on Facebook that sparked controversy and was considered offensive to Islam. The killer was later charged with terrorism and put to death. Celebrating murder can be considered amoral in many societies, and yet this speech was permitted in this particular context.

In other examples, a forum that focuses mainly on historical religiously motivated "afflictions and fierce battles" contains incendiary remarks about religious minorities such as the Copts in Egypt.[3] A political discourse forum whose terms and conditions stress the use of appropriate terms in political dialogues allows users to use derogatory terms when referring to fellow Yemenis who are not from their region.[4] The terms, such as Dehbashi,[5] can damage the social fabric in the already civil-war-torn country. Voices with harmful messages attempt to influence hacking forum activities by encouraging hackers to target political networks and websites of those they describe as enemies of Islam and infidels; thus they create a hateful hacking subculture in a largely white hat hacking platform.[6]

This uncontested self-regulatory model is exacerbated by two contributing factors: Unlike social media, the forums receive limited external or independent attention that can invigorate debate on their harmful speech policies. For example, whereas the Jordanian authorities arrested and referred for legal prosecution social media users for posting hate speech after the killing of Hattar,[7] similar content appears on the forums and evades regulatory attention. Also, the forums work outside intermediary liability regimes. For example, there are no accessible mechanisms to flag potentially offensive content to the web services that host the forums.

Conclusion

Self-regulated forum-based communities are prone to promulgating harmful speech in a manner that is deeply embedded in self-regulatory systems. Forum administrators of homophilic communities risk creating a climate that permits and even normalizes harmful speech. The problem is likely to persist and hit more communities as forums increasingly host rhetoric on the region's ongoing political and religious conflicts. Harmful speech is rooted in cultural grounds—the "self" in "self-regulation." Because the self owns the space and shapes its policies, the priority is how to better moderate the self as that will subsequently influence the regulation.

References

1 - I am a Muslim Network for Islamic Dialogue Forum, "Qawaneen wa shurut shabakat ana muslim lelhewar alislami (Terms and Conditions)," accessed June 1, 2017, http://www.muslm.org/vb/misc.php?do=showrules.

2 - See for example discussion thread: I am a Muslim Network for Islamic Dialogue Forum, "Eghtiyal alkatib alshiyoi nahed hattar (Assassination of communist writer Nahed Hattar)," accessed June 1, 2017, http://www.muslm.org/vb/showthread.php?562066-%D8%A3%D8%BA%D8%AA%D9%8A%D8%A7%D9%84-%D8%A7%D9%84%D9%83%D8%A7%D8%AA%D8%A8-%D8%A7%D9%84%D8%B4%D9%8A%D9%88%D8%B9%D9%89-%D9%86%D8%A7%D9%87%D8%B6-%D8%AD%D8%AA%D8%B1.

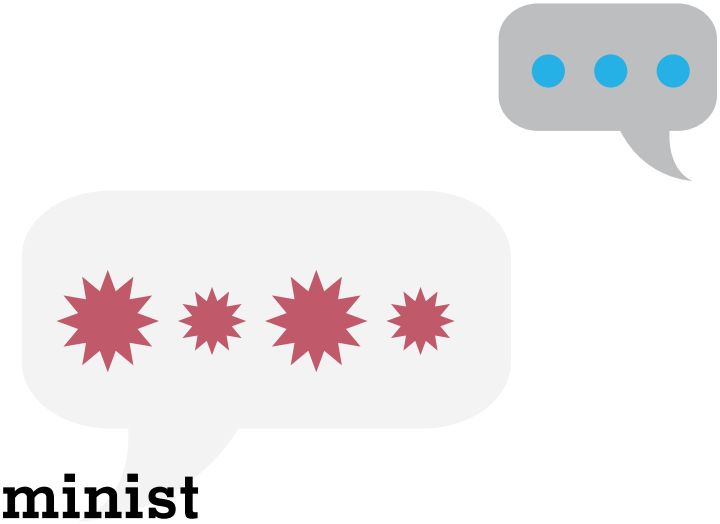
3 - See for example discussion thread: Afflictions and Fierce Battles Forum, "Allah Allah fi Aqbat Misr (Allah Allah, The Copts of Egypt)," accessed June 1, 2017, http://alfetn.net/vb3/showthread.php?t=34539.

4 - Yemeni Council Forum, Terms and Conditions, accessed June 1, 2017, www.ye1.org/forum/help/terms.

5 - Ali Aboluhom, "Nov. 30 Deadline Looms over Northeners in the South," *Yemen Times*, November 6, 2014, accessed June 1, 2017, http://www.yementimes.com/en/1831/report/4550/Nov-30-deadline-looms-over-northerners-in-the-south.htm.

6 - Aljiyyosh Hacking Forum, "Nasaeh qabl an takun hacker (Advice before you become a hacker)," accessed June 1, 2017, www.aljyyosh.com/vb/showthread.php?t=28990.

7 - Rana Husseini, "Social media users to be sued over hate speech in reaction to Hattar shooting," *The Jordan Times*, September 25, 2016, accessed June 1, 2017, http://jordantimes.com/news/local/social-media-users-be-sued-over-hate-speech-reaction-hattar-shooting%E2%80%99.



Difficult Speech in Feminist Communities

Kendra Albert

Many feminist communities online have developed sets of practices to accommodate, moderate, and regulate speech. As we consider the implications of hateful speech on our online communities, it is vital that we also reflect upon how communities deliberatively deal with wanted yet complicated topics, and whether these practices can provide models for dealing with formulating and regulating speech according to community-developed norms. This essay discusses one such set of models — a set of interventions against what I call “difficult speech.”

Difficult speech is speech that is wanted yet may also cause discomfort or harm in a community with a shared set of norms. For example, a trans person in a community aimed at trans folks might want to discuss their body as part of seeking advice on dysphoria (a psychological condition of distress stemming from one’s body not matching one’s gender). However, for other trans folks, a person’s recounting of their feelings about their body may be something that they cannot read without having suicidal thoughts. The issue is further complicated by the fact that what may cause difficulty for a person on their bad day might be perfectly fine a few days later. This variability, across both people and time, creates unique moderation needs. In writing this piece, I reviewed a small number (~5) of feminist sites, including both blogs with moderated comments sections and forums/private community spaces, to see how they deal with difficult speech. Content warnings and multiple channels with redirection are two options for handling this moderation that were common to multiple surveyed spaces.

Using Content Warnings to Offset the Impact of Difficult Speech

Perhaps the most obvious method of dealing with difficult speech is "content warnings" or “trigger warnings.” Content warnings are literal statements of the content of following text or images—for example, if a text contains the first-person narrative of sexual assault, a content warning might say “sexual assault.” (Generally, the term “content warning” is considered broader than “trigger warning” and thus I will use it.)

Content warnings are not unique to feminist communities but are probably more common in feminist spaces than elsewhere. Warnings can be used in a variety of circumstances, for content containing anything from depictions of rape to manifestations of white supremacy. In some communities, warnings are deployed along with tags that make the difficult material unreadable unless moused over ("spoiler tags"). Where material is not obscured, a content warning can be paired with a note about how much text the warning covers ("CN: police violence, next 4 paragraphs").

Communities often engage in discursive practices around what kinds of content requires a warning—allowing autonomy and discussion over shared values. Commonly chosen content warnings among some feminist communities surveyed include "sexual assault," "transphobia," "racism," "war on agency" (reproductive rights), and "Nazis." As demonstrated by this list, the potential options are broad and often depend on the needs and characteristics of the members of the community.

Using Multiple Channels to Respond to Difficult Speech

Some communities use a combination of multiple channels and conversation redirection to handle difficult speech. For example, there might be two channels for a particular issue: #bodyissues and #bodyissues-unfiltered. When someone wants to talk about something that others might find difficult, either as explicitly mentioned in guidelines or just understood as a sensitive topic, they might post in #bodyissues with a content warning and a pointer – "I want to talk about a dysmorphia thing in unfiltered. If you're up for listening meet me there." Users who are able to support can view #bodyissues-unfiltered to read and comment. Other users who may not be worried about potential triggers can view the unfiltered channel as part of their daily community interactions.

Finally, a user who is finding a conversation taking place in the #bodyissues tag difficult can ask other users to move to #bodyissues-unfiltered. This allows for more situational reactivity than a more traditional content warning system.

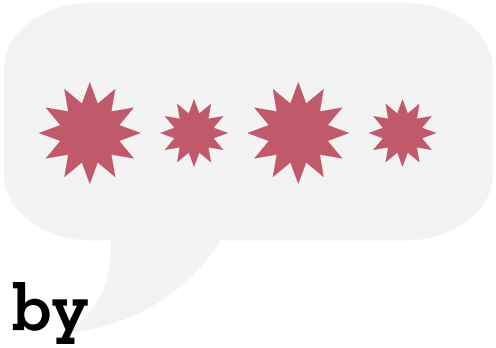
Platforms, Affordances, and Regulation

One notable characteristic of the aforementioned interventions that deal with difficult speech is that they rely on platforms having particular affordances, and on making these affordances accessible to moderators—the power to ban members, to create multiple channels, and to block out speech (for example, with spoiler tags). Thus, difficult speech interventions may not be possible in communities that work on platforms that lack these. For example, a community on Facebook couldn't use spoiler tags, as they are unsupported by the platform.

Additionally, difficult speech interventions can be undermined by more traditional moderation actions by platforms. For example, imagine a racial slur is used in the context of explaining a recent experience and asking for reassurance. If an appropriate content warning is used, the harmful effects on members of the targeted community may be mitigated. Nevertheless, the post containing the slur might trigger a "shadowban" or "time out" from the platform due to the language—resulting in fewer people seeing the post at all, the exact opposite of what the user may need.

As I write this essay, Mastodon, an alternative social network, has been rapidly gaining popularity. Mastodon supports content warnings, and users from different Mastodon servers have been engaged in robust debate over what content deserves warnings, from politics to porn. Whether Mastodon ends up going the way of forgotten social networks like Diaspora or Ello or becomes widely adopted, it is notable that content warnings are now increasingly integrated directly into platforms.

Since much regulation of speech is bound up in legal frameworks and debates over banned terms, community adaptations to difficult speech, like those taking place on feminist platforms or on Mastodon, suggest a new way forward for dealing with harmful speech online.



Comment Moderation by Algorithm: The Management of Online Comments at the German Newspaper 'Die Welt'

Anke Sterzing, Felix Oberholzer-Gee, and Holger Melas

Many media organizations refrain from publishing reader comments they deem inappropriate. In one recent high-profile case, National Public Radio stopped publishing comments altogether on NPR.org, arguing that social media such as Twitter and Facebook were better suited for listener comments and public debate.[1] In shutting down the entire comment section, NPR joined the ranks of the Chicago Sun-Times, Reuters, the Week, and even magazines such as Popular Science.[2] Media outlets that continue to publish reader comments invariably resort to heavy moderation in order to preserve civility and avoid hate speech, but also to save costs.

Media outlets that eliminate reader commentary give up a potentially attractive means to engage their readers. At the German newspaper Die Welt, a national daily and the object of this brief case study, the commentary section garners 10 percent of all page visits. Readers who visit the section exhibit increased retention and a greater likelihood of returning to Die Welt. Moreover, live chats with journalists seem to increase readers' willingness to pay and tie readers more effectively to the brand. Managing the flood of comments, however, is challenging. Initially, Die Welt checked each comment before publication, an expensive and time-consuming process that left the moderators with little opportunity to engage in the debates.[3]

To preserve the economic viability of online comments, Die Welt introduced a computational linguistic tool in October 2015. Its algorithms examine each comment prior to publication. Readers' contributions are classified as falling into one of three categories: publishable, to be rejected, and potentially unsafe. The moderation team therefore only needs to assess the comments that are classified as potentially unsafe.

Thanks to the linguistic tool, users who respect the law and the editorial rules of Die Welt can communicate almost in real time, even when contribution volumes are very high. In addition, manual moderation has decreased by about 70 percent. This frees up the moderators to answer questions, supervise live chats with authors and experts, and publish the most interesting reader comments to social media.

Similar to Google's tool Perspective, which is now used by the New York Times and the Economist,[4] the algorithm that Die Welt uses was trained on the 8.5 million comments the paper received from 2011 to 2015. Each of these had been classified by human editors. Building on human judgment, the tool now relies on semantic, syntactic, and morphological data analysis to distinguish publishable comments from problematic content. Each comment is compared with extensive blacklists and gray lists and a lexicon of offensive language. In the experience of Die Welt, electronic masking and meaningless sequences of symbols all heighten the likelihood that a comment cannot be published. Unusual punctuation and the heavy use of numbers, capitalized letters, and foreign words also raise suspicion.

As the language used online evolves quickly, Die Welt updates its algorithm constantly. The automatic moderation system has a feedback API that allows readers to flag inappropriate comments. If the moderators' judgment concurs with the view of readers, the reclassified comments help to further train the algorithm. Because Die Welt's algorithm is used by other media outlets as well, the paper can learn from the experience of online journalism more generally. At this time, the automated moderation system has an error rate of less than 4 percent.

Perhaps surprisingly, it now performs better than human moderators did in the past. While its average performance is remarkable, there are times when comments on specific articles cause waves of reader flags and a need for extensive re-moderation. In these circumstances, the paper switches back to human moderation of all comments on the article until its algorithm can be updated correspondingly.

Combining a learning algorithm with experienced moderators and critical readers allows Die Welt to continue the publication of reader comments at moderate cost and with little risk of publishing illegal and offensive speech.

References

1 - Elizabeth Jensen, "NPR Website To Get Rid Of Comments," *NPR*, August 17, 2016, http://www.npr.org/sections/ombudsman/2016/08/17/489516952/npr-website-to-get-rid-of-comments. In July 2017, readers left almost half a million comments on NPR.

2 - Justin Ellis, "What happened after 7 news sites got rid of reader comments," *NiemanLab*, September 16, 2015, http://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments/.

3 - Similar to many other EU countries, Germany has hate speech laws that criminalize certain types of speech, such as incitement to racial violence.

4 - See Perspective, https://www.perspectiveapi.com/.

Decoding Hate Speech in the Danish Public Online Debate

Lumi Zuleta

Social media is playing an integral part in today's public debate. On the one hand, social media has made it easier to use our freedom of expression and participate in the public debate. On the other hand, the *tone* online has been criticized for polarizing, spreading hate, and silencing people.

The rise of hate speech online is ascribed to social media, where hateful comments are easily shared and spread to a large audience. Specifically, in Denmark, Facebook is the most commonly accessed social media platform, used by 97 percent of 16- to 89-year-olds every week.[1]

This paper presents key findings from a study of hate speech on the Facebook pages of DR and TV 2 — the two major news outlets in Denmark.[2] The study is based on a questionnaire survey and a quantitative content analysis of 2,996 Facebook comments.[3]

The Chilling Effect

In a representative survey from 2016, 50 percent of the Danes said that the tone in social media debates keeps them from expressing their opinion and from participating in the debate.[4] The fact that a harsh tone scares people off points to a democratic problem with consequences for public debate.

Public debate is a cornerstone of a democratic society—not as a constitutional right or an institution but as something citizens and decision makers develop together, often using the media as an intermediary. In this light, it is interesting to study debate on one of the latest and most popular platforms—Facebook—and look at how it is moderated by news media.

Prevalence of Hate Speech

Content analysis found that one in seven comments (15 percent) contains hate speech (the definition covers both lawful and unlawful hate speech).[5] Hate speech most often appears in connection with news posts about religion, refugees, gender equality, politics, and integration.

The majority of the hateful comments derive from male debaters (76 percent). In most instances, hate speech targets people's political beliefs or specific politicians. Other areas that often draw hateful comments are religion and ethnicity. Particularly Islam and individuals of non-Danish descent are subjected to hate speech. Hate speech based on gender is more frequently targeted at women than at men.

Media Responsibility?

At a time where much of the public debate takes place online, it is important to discuss media responsibility with regard to combating hate speech. In a Danish context, the media's legal responsibility when hosting online debates on social media platforms remains unclear. However, when asking users,[6] a large majority (77 percent) believe that the media has a responsibility to remove offensive and derogatory comments. This indicates that users want the media to be more proactive in ensuring a civil tone in the debate. The question is how the media should approach this task, without violating freedom of expression.

The fact that one in two refrains from participating in the public online debate is an issue that news media should take seriously. In Denmark, this discussion is just beginning.

References

1 - "Social Medier Brug, Interesseområder og debatlyst grafikker og tabeller," *The Danish Agency for Culture*, 2015, http://slks.dk/fileadmin/user_upload/dokumenter/medier/Mediernes_udvikling/2015/Specialrapporter/Sociale_medier/PDF-filer_dokumenter/SAMLET_Rap_2_FIGURER_OG_TABELLER.pdf.

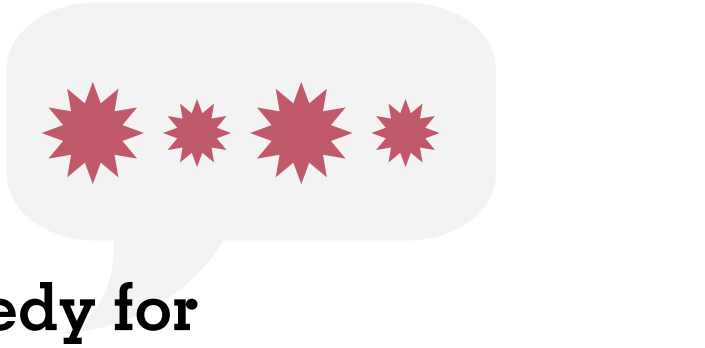
2 - Lumi Zuleta and Rasmus Burkal, "Hate speech in the public online debate," *The Danish Institute for Human Rights*, 2017, https://menneskeret.dk/udgivelser/hadefulde-ytringer-paa-facebook.

3 - The comments were collected in the period April-July 2016 and no sooner than 12 hours after publishing allowing the media time to edit, hide and possibly delete comments conflicting with their guidelines. Consequently, the data does not represent the total volume of hate speech comments actually posted by users, but shows the number remaining after allowing reasonable time to moderate.

4 - Question included on behalf of the Danish Institute for Human Rights by Statistics Denmark, and survey was conducted on 16 June 2016. Available (in Danish) at: http://www.dst.dk/da/Sites/folkemode/undersogelse.

5 - There is no universally accepted definition of hate speech in international human rights law. In the study, hate speech is defined as: "Publically voiced stigmatizing, derogatory, offensive, harassing and threatening statements that are directed at an individual or a group based on the individual's or the group's gender, ethnicity, religion, disability, sexual orientation, age, political beliefs or social status." The eight grounds are all protected grounds of discrimination in human rights law (See Article 2 of the U.N. Universal Declaration of Human Rights, Article 21, section 1 of the EU Charter of Fundamental Rights).

6 - 1,045 respondents were asked: "To what degree do you believe that the media (DR, TV 2, Politiken, etc.) are responsible for ensuring a civil debate and for deleting derogatory and offensive comments from their own Facebook pages?" on a Megafon survey.



Verification as a Remedy for Harmful Speech Online

Simin Kargar

In September 2016, the Instagram account of Shahin Najafi, an Iranian musician, was hacked and defaced. Instead of Shahin's profile picture, his over 500,000 followers saw the flag of the Islamic Republic of Iran. In addition, the account's bio caption was brazenly replaced with the attacker's contact information — a typical feature of state-aligned cyberattacks and intrusion.[1]

Najafi's songs address socially and politically sensitive issues such as theocracy, censorship, sexism, and homophobia. Following the publication of his controversial song about a Shiite saint in 2012, two leading Iranian clerics issued fatwas declaring Najafi guilty of apostasy.[2] He received multiple death threats from across social media, and a far right website offered a $100,000 bounty to anyone who killed Najafi.[3] He has remained a constant target of hate speech and different types of cyberattacks. Multiple fake accounts have impersonated Najafi and echoed negative speech about him. In addition, state-run media have repeatedly conducted smear campaigns against him.

Despite his celebrity status and clear need for more protection by platform operators, Najafi has been unable to verify and secure his Instagram and Twitter accounts against malicious impersonation and other forms of abuse.

Najafi's case has not been the only incident of this kind. For several years, Iranian civil and political dissidents have been top targets of state-sponsored cyberattacks and intrusion campaigns. More recently, these groups have become regular targets of coordinated online mobs that sometimes appear to have links to the state agencies. Many encounter content takedown and account suspension due to coordinated flagging and reporting of their posts on social media. In addition, they are often impersonated by fake accounts that disseminate misinformation about their private and public lives. With their privacy and integrity under attack, some end up deactivating their accounts. Others restrict the comment section of their profiles, and a few seek protection and support from the intermediaries that host their content.

On the other hand, social media platforms such as Facebook, Twitter, and Instagram strive to protect free expression, often for good reasons. By prioritizing expression, however, intermediaries may inadvertently undermine the basic human rights to privacy and free expression of some of the most vulnerable targets of harmful speech online. Human rights situations are dynamic, and so is the need for protection in the different environments in which intermediaries operate.

While not a solution to the problem, one technical remedy that has helped many public-facing artists, activists, and journalists who have come under threats online is account verification. A verified badge appears as a checkmark next to an account's name in search and on the profile. It means that social media platforms have confirmed that the account is authentic and represents the actual account holder. In practice, verified profiles enjoy more protection against false reporting and politically driven flagging of content. They appear to have more leverage on removing fake accounts or misinformation about the account holder. In brief, the small blue badge has proved to be protective of freedom of expression for its recipients.

Interviews I conducted with 20 prominent Iranian human rights activists, artists, and journalists confirm their limited success in drawing social media platforms' attention to their cases or at a minimum getting the platforms to verify their accounts. Only journalists affiliated with "reputable employers," e.g., international media, have had the privilege of receiving the blue badge with minimum difficulty.

However, the verification success of these journalists is the exception rather than the norm. Among my interviewees, all Iranian women's rights activists and LGBTQ public figures had failed to gain verification for their profiles even after they sent in the required documentation. The lack of a badge is particularly challenging for those who work in an individual capacity, unless they manage to establish a connection with social media platforms through third parties. While Twitter has detailed steps on how to request a verified badge for an individual account, Instagram and Facebook refer to verification as possible for "only some public figures, celebrities and brands."[4,5,6]

In addition to the unclear process, there are other complications.

First, the guides explaining how to become verified are not available in many local languages, including Farsi, and this language gap is not limited to the verification rules.[7] For instance, there is no information available in Farsi to guide individuals on reporting and documenting harassment.[8] This becomes particularly worrisome when we consider the many threats of violence that Iranian human rights activists and dissidents often receive as direct messages on Facebook, Instagram, and Twitter.[9]

Second, the requests for verification by most influential activists are repeatedly rejected on the grounds that they are not "famous enough," according to multiple interviewees. Lack of context and obscure understanding of the significance of their work in "human rights-unfriendly" environments like Iran seem to impede the protection that these activists desperately need. In addition, it creates a climate of mistrust between the targets and intermediaries of harmful speech online.

In the past few years, social media platforms have taken noteworthy measures and demonstrated more accountability against harmful speech online. Yet there are still gaps to address, particularly concerning vulnerable communities whose work is deeply influential and not Western-centric. In addition, their audience, and their attackers, largely reside far from where major intermediaries are headquartered. These individuals are often deprived of protection from law enforcement in their respective countries. In some cases, there is even evidence that the government likely supports the perpetrators of this harassment. Therefore, the ultimate hope for these targets is to be granted more leeway by the only other stakeholder involved — the intermediaries.[10]

More transparency about the functionality and processing of verification requests and reports of abuse can go a long way toward maintaining trust with the end users worldwide. For effective engagement in

addressing the concerns of affected stakeholders, the intermediaries also need to take language and other barriers into account. Making relevant information available to local communities reflects care and respect for the rights of regular targets of harmful speech online.

As the U.N. Guiding Principles on Business and Human Rights outline,[11] one of the initial human rights due-diligence measures is to mitigate the specific impact in a particular context of business operations. To this end, local vulnerable groups require dedicated attention from these intermediaries. Making verification more accessible to human rights defenders, at-risk groups, and marginalized communities is only a partial remedy for the adverse impacts that these individuals endure while advocating for the rights of many others, but it represents a concrete and useful step deserving of better attention and support.

References

1 - Claudio Guarnieri and Collin Anderson, "Iran and the Soft War for Internet Dominance," paper presented at Black Hat 2016, Las Vegas, August 2016, https://iranthreats.github.io/us-16-Guarnieri-Anderson-Iran-And-The-Soft-War-For-Internet-Dominance-paper.pdf.

2 - Thomas Erdbrink, "Rapper Faces Death Threats in Iran Over Song," *The New York Times*, May 15, 2012, http://www.nytimes.com/2012/05/15/world/middleeast/shanin-najafi-iranian-born-rapper-faces-death-threats-over-song.html.

3 - http://online-shia.ir

4 - Twitter, "Request to Verify an Account," accessed May 22, 2017, https://support.twitter.com/articles/20174631.

5 - Instagram, "Verified Badges," accessed May 22, 2017, https://help.instagram.com/854227311295302.

6 - Facebook, "What is a Verified Badge or Profile?," accessed May 22, 2017, https://www.facebook.com/help/196050490547892.

7 - The error on top of the text reads "Sorry but this text is not available in your language." (https://support.twitter.com/articles/20174631). While the drop-down list includes 32 languages, a few generate the same message stating that content is not available in the language selected. English, French and other major languages are all available.

8 - Facebook, "Reporting Conversations," accessed May 22, 2017, https://www.facebook.com/help/messenger-app/iphone/1165699260192280/?helpref=hc_fnav. The text in yellow box states that the guide is not available in the selected language i.e. Farsi but that users can choose from "supported" languages.

9 - These are the most popular social networking platforms among Iranians, hence the emphasis.

10 - Rob Faris, Amar Ashar, Urs Gasser and Daisy Joo, "Understanding Harmful Speech Online," *Berkman Klein Research Center*, December 10, 2016, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882824.

11 - U.N. Guiding Principles on Business and Human Rights, 2011, accessed May 22, 2017, http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. These principles were proposed by the UN Special Representative on Business and Human Rights, John Ruggie, and endorsed by the U.N. Human Rights Council in June 2011.

Ensuring Beneficial Outcomes of Platform Governance by Massively Scaling Research and Accountability

J. Nathan Matias

Because online platforms observe and intervene in the lives of billions of people, many have come to expect that they should address enduring social problems, including terrorism,[1] hate speech,[2] suicide prevention,[3] police brutality,[4] eating disorders,[5] and human trafficking.[6] Interventions often involve complex assumptions about combinations of individual psychology, collective behavior, choice of architectures in platform design, and the behavior of algorithms that learn and adapt to human activity. If these factors can be orchestrated wisely, they offer powerful opportunities for social change.

Platform-based social change will remain a risky, dangerous endeavor without progress on two fundamental challenges: evaluation and governance. First, without systematic evidence about the outcomes of platform interventions, policymakers risk increasing harms rather than reducing social ills. At the scale and breadth of global human society where platforms are expected to intervene, estimating policy outcomes would require a thousandfold increase in research over what is currently available. Second, as platforms discover powerful interventions to direct our moral and political behavior, that power must itself be governed.

Upon evaluation, many of the most lauded platform interventions have been shown to have damaging side effects or even increase the behavior they were intended to reduce. In 2014, 17 years after the invention of the “downvote” in comment discussions, researchers showed that each vote of disapproval in political discussions causes people to behave more badly over time.[7] Elsewhere, researchers discovered four years after Instagram's algorithmic efforts to obfuscate self-harm that their policy may have increased participation in harmful communities.[8] After Wikipedia introduced powerful vandalism detec-

tion systems, researchers discovered that these systems had caused a dramatic decline in participation over the next six years. In all three cases, the systemic harms from these harm-reduction strategies were not observed until after they had affected tens to hundreds of millions of people for years.[9]

Causal research methods would allow platform policymakers to predict the likely outcomes of an intervention on average.[10] Because platforms are designed to intervene, monitor, and process data about billions of people, they already possess the potential to make these evaluations.[11] Policymakers can choose from a rich palette of research techniques for piloting interventions, including randomized trials and post hoc quasi-experiments.[12] When policy goals are difficult to measure, qualitative field experiments allow powerful inferences on the outcomes of an intervention.[13]

The breadth and scale of platform power places new demands on the scale of research. Because the nature of problems such as hate speech and suicide varies across regions and cultures, it is likely that the most effective interventions will vary as well. While platforms do possess the ability to tailor interventions to context, this tailoring would require hundreds of new studies per policy. Platforms have developed infrastructures to scale research in sales and marketing, conducting up to hundreds of randomized trials per day — which adds up to tens of thousands of studies every year per platform.[14] Yet mass evaluation of policy on a similar scale has never been attempted. By my rough estimation, less than a dozen field experiments have ever been published on public interest uses of platform policies.

If platforms can genuinely provide effective mechanisms for shaping the behavior of billions of people, we need methods to govern the use of that power. Policy research findings could play an important role in holding platform policymakers accountable. In *The Open Society and Its Enemies*, Karl Popper imagined the role of social experiments in democratic and authoritarian societies. In closed societies, argued Popper, unaccountable policy evaluators could shape human life in secret and toward ends that the public could not influence. Popper proposed the "open society" as an alternative — a society in which the goals and means of social policy are subject to democratic processes, and public knowledge of research guides the public to reject harmful or ineffective uses of power.[15]

Local community policymaking — a model with 40 years of history on the internet — may offer a powerful opportunity to achieve needed research scales along with the public accountability of an open society. Across the social web, community moderators, sysops, group admins, and other local policymakers already do substantial work to support and govern millions of people online.[16] In a series of pilot studies I led recently, the CivilServant project offered communities the ability to conduct evaluations on the effects of their local policies, share findings openly, and replicate one another's research. In just a few months, these early studies and associated community deliberations have shaped decisions affecting tens of millions of people and spread policy ideas to over a hundred other communities (civilservant.io). If communities continue to embrace this model, they have the potential to generate thousands of transparent, accountable policy experiments in platform governance each year.

Whether platform operators or local communities create and enact policies governing the lives of billions of connected people, responsible use of their power will require new models of dramatically scaled research and democratic participation. Platforms already possess the potential for both, but neither is very common. Until reliable methods to evaluate and govern platform power become commonplace, our efforts to reduce social ills through platforms present large-scale, unestimated risks to society, even as we search for the benefits.

References

1 - Natasha Lomas, "Twitter Nixed 635k+ Terrorism Accounts Between Mid-2015 and End of 2016," *TechCrunch*, March 21, 2017, https://techcrunch.com/2017/03/21/twitter-nixed-635k-terrorism-accounts-between-mid-2015-and-end-of-2016/.

2 - Danielle Keats Citron and Helen L. Norton, "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age," *Boston University Law Review*, 91 (2011): 1435.

3 - Rachel Metz, "Facebook Live's New Suicide-Prevention Tools Come with Good Intentions but Many Questions," *MIT Technology Review*, March 1, 2017, https://www.technologyreview.com/s/603772/big-questions-around-facebooks-suicide-prevention-tools/.

4 - Caroline O'Donovan, "Nextdoor Rolls Out Product Fix It Hopes Will Stem Racial Profiling," *BuzzFeed,* August 24, 2016, https://www.buzzfeed.com/carolineodonovan/nextdoor-rolls-out-product-fix-it-hopes-will-stem-racial-pro.

5 - Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert and Munmun De Choudhury, "#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities," (paper presented at 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, USA, February 27-March 2, 2016), http://dl.acm.org/citation.cfm?id=2819963.

6 - Mitali Thakor and danah boyd, "Networked Trafficking: Reflections on Technology and the Anti-Trafficking Movement," *Dialectical Anthropology*, 37 no. 2 (2013): 277–290.

7 - Justin Cheng, Cristian Danescu-Niculescu-Mizil and Jure Leskovec, "How Community Feedback Shapes User Behavior," (paper presented at ICWSM 2014, Ann Arbor, USA, June 2-4, 2014), http://arxiv.org/abs/1405.19.

8 - Chancellor et al., "#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities."

9 - Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan and John Riedl, "The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline," *American Behavioral Scientist,* 57, no. 5 (2013), http://journals.sagepub.com/doi/abs/10.1177/0002764212469365.

10 - Donald T. Campbell, "Reforms as Experiments," *American Psychologist*, 24, no. 4 (1969): 409.

11 - David Lazer, Alex (Sandy) Pentland, Lada Adamic et al., "Life in the Network: The Coming Age of Computational Social Science." *Science*, 323, no. 5915 (2009): 721–723, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745217/.

12 - Joshua Angrist and Jörn-Steffen Pischke, "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *National Bureau of Economic Research*, Working Paper No. 15794 (March 2010), https://doi.org/10.3386/w15794.

13 - Elizabeth Levy Paluck, "The Promising Integration of Qualitative Methods and Field Experiments," *The Annals of the American Academy of Political and Social Science*, 628, no.1 (2010): 59-71.

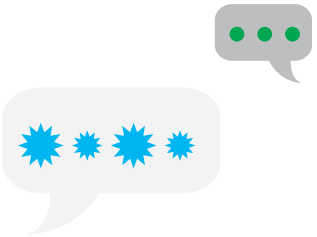
14 - Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu and Nils Pohlmann. "Online Controlled Experiments at Large Scale," (paper presented at the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, USA, August 11-14, 2013), http://dl.acm.org/citation.cfm?id=2488217.

15 - Karl Popper, *The Open Society and Its Enemies*, (Abingdon, Oxon: Routledge, 1945).

16 - J. Nathan Matias, "The Civic Labor of Online Moderators," (paper presented at the Internet Politics and Policy conference, Oxford, United Kingdom, 2016), http://ipp.oii.ox.ac.uk/sites/ipp/files/documents/JNM-The_Civic_Labor_of_Online_Moderators__Internet_Politics_Policy_.pdf.

Looking Ahead

A reflection by the Harmful Speech Online team



The pieces in this collection offer many entry points for reflection, critical thought, and future research. As highlighted in the essays, governance models take different forms around the world, no two platforms share the same moderation policies, and interventions and solutions are often context-specific. Best practices are yet to be distilled and universally applicable standards are likely to remain elusive but, as J. Nathan Matias suggests in this collection, experimentation and data sharing are promising ways to come closer to consensus.

Indeed, there are several key questions to which better answers could help guide more effective interventions with fewer collateral costs: Does regulating harmful speech chill freedom of expression? Or does regulating harmful speech enable more speech? Are we better served by leaving it up or taking it down, so to speak?

As a global society, we are still far from consensus over the question of whether harmful speech should be taken down when and where it emerges. Protecting harmful speech when it lies within the bounds of legality has long been seen as a cornerstone of democratic discourse, but this notion is challenged where removals do more to support broader participation in the public sphere than they do to suppress it. This is fundamentally an empirical question worthy of further exploration, and it leads us to ask about effective interventions: When does counterspeech work? Can platforms nudge users into being more considerate citizens?

Addressing harmful speech online with *more* speech is deeply rooted in a central tenet of freedom of expression theory. If constructive speech counters destructive speech, then more aggressive legal approaches, such as takedowns and filters, are less appropriate responses. This is particularly salient where the very act of openly engaging with harmful speech might help to address underlying issues that would otherwise be swept under the carpet, ultimately fostering a more open, tolerant, and civic-minded community. Understanding the conditions and context in which counterspeech is a potent antidote to harmful speech will require much greater attention from researchers as well as support and assistance from companies and governments to enable such initiatives.

At a grassroots level, many of us put stock in the notion that better education and awareness-building focused on the impact and ramifications of harmful speech will help to deter and prevent harmful speech at its source. This prompts the question: How far does public education on these issues take us? We need to establish whether this is in fact valuable and worthy of attention and resources that might otherwise be invested in other interventions.

Looking ahead to the future of harmful speech online, we anticipate the growing role of algorithms and artificial intelligence used by platforms to influence how different actors engage in this ecosystem. Through machine learning and algorithmic approaches, and with robust and reliable data, communities and platforms may be able to more aggressively and thoroughly remove negative speech while minimizing (although not completely preventing) collateral damage. We have yet to know how far automated tools can really go toward identifying and mitigating problematic speech while protecting freedom of expression. Algorithms may be key to new approaches, interventions, and solutions, but they will also generate new quandaries as well as introduce new dynamics to problems that already exist.

Collaboration and information sharing across these many endeavors and sectors will be crucial to understanding and addressing problems, implementing novel approaches to countering harmful speech online, and, perhaps most importantly, enabling cooperation between the many actors in the space who seek progress.

Contributor Bios

Kendra Albert works at Zeitgeist Law, a boutique technology law firm in San Francisco. They are also an affiliate at the Berkman Klein Center for Internet & Society, and a writer and speaker on a diverse set of internet issues, including computer security and online harassment. Kendra holds a JD from Harvard Law School and a BHA from Carnegie Mellon University.

Susan Benesch is Faculty Associate of the Berkman Klein Center for Internet & Society at Harvard University. She founded and directs the Dangerous Speech Project, to study speech that can inspire violence - and to find ways to prevent this, without infringing on freedom of expression. To that end, she conducts research on methods to diminish harmful speech online, or the harm itself. Trained as a human rights lawyer at Yale, Susan also teaches at American University.

Adnan Ahmad Chaudhri is an Associate Researcher with Digital Rights Foundation, focussing on surveillance and the right to privacy, and tackling online harassment. He has a background in political science, history and archaeology, and has worked with the media, development sector and academia.

Kate Coyer is a Fellow at Berkman Klein Center for Internet & Society at Harvard University and directs the Civil Society and Technology Project at the Center for Media, Data and Society. Her research examines the use of new and old technologies for social change and the impacts on human rights and freedom of expression, as well as the role of social media company policy and practice. A longtime radio producer and organizer, her work supports digital rights advocacy, community media, and communication access for refugees, which has been featured on NPR, BBC, Washington Post, New Scientist, Wired and others. She holds a PhD from Goldsmiths College and previously held a postdoctoral fellowship from the Annenberg School for Communication, University of Pennsylvania.

Nighat Dad is the founder and Executive Director of Digital Rights Foundation, a Pakistan-based privacy and surveillance non-governmental organisation that tackles the rise of gendered digital violence and advocates for civil liberties and privacy protection. TIME magazine listed her as one of their 2015 Next Generation Leaders for her work aiding women to fight online violence and harassment. In June 2016 she was a recipient of the Atlantic Council Freedom Award, and was the 2016 recipient of the Government of the Netherlands' Human Rights Tulip Award.

Nani Jansen Reventlow is an Associate Tenant at Doughty Street Chambers and a 2016-2017 Fellow at the Berkman Klein Center for Internet & Society at Harvard University. She is a recognised international lawyer and expert in human rights litigation responsible for groundbreaking freedom of expression cases across several national and international jurisdictions. Between 2011 and 2016, Nani has overseen the litigation practice of the Media Legal Defence Initiative (MLDI) globally, leading or advising on cases before various national and international courts. At the Berkman Klein Center, Nani's work focuses on cross-disciplinary collaboration in litigation that challenges barriers to free speech online. She also acts as an Advisor to the Cyberlaw Clinic.

Amy Johnson is a linguistic anthropologist and Science, Technology and Society scholar who studies interactions of language and technology across English, Japanese, and Arabic. Recent research focuses on parody, platform governance, public scholarship, harassment online, social media management by US governmental agencies, manual bots in Japanese-language Twitter, and standup comedy and the law in the UAE. Johnson holds a PhD from MIT and is a Fellow at Amherst College's Center for Humanistic Inquiry and an Affiliate at the Berkman Klein Center for Internet & Society at Harvard University.

Rey Junco is an associate professor of education and human computer interaction at Iowa State University and a faculty associate at the Berkman Klein Center for Internet & Society. He is the director of the Junco Applied Media (JAM) Research Lab. Rey applies quantitative methods to analyze the effects of social technologies on youth psychosocial development, engagement, and learning.

Simin Kargar is a human rights lawyer with specific focus on gender, media and communication laws and policies in Iran. Presently, Simin is a fellow at Harvard's Berkman Klein Center for Internet & Society where she focuses on harmful speech online, the interplays of media, power and propaganda and how online platforms are facilitating its dissemination. In addition, Simin studies how discourses of geopolitical issues are created and disseminated through social media. Simin has published research analyzing the formation of online movements that led to changes in policies, or to creating tangible social impact. Her current research addresses cyber abuse and online harassment against communities of journalists, artists and activists who - due to their profession, activities or beliefs- are more likely to become targets of coordinated online mobs.

Andres Lombana-Bermudez is a researcher and designer working at the intersection of digital technology, youth, innovation, and learning. His approach is transdisciplinary and collaborative, combining ethnographic and quantitative research methods, design-based research, and media technology design. He is a post-doctoral fellow at Harvard University's Berkman Klein Center for Internet & Society and a Research Associate with the Connected Learning Research Network.

J. Nathan Matias recently completed a PhD at the MIT Media Lab Center for Civic Media and is an affiliate at the Berkman-Klein Center. Nathan conducts independent, public interest research on flourishing, fair, and safe participation online. With CivilServant, Nathan has supported millions of people to test ideas for improving social life online. Nathan has extensive experience in tech startups, nonprofits, and corporate research, including SwiftKey, Microsoft Research, and the Ministry of Stories. His work has been covered extensively by international press, and he has published data journalism and intellectual history in the Atlantic, Guardian, PBS, and Boston Magazine.

Holger Melas was born in Schäßburg in 1972, grew up in the Rhineland and studied sports at the German Sport University Cologne. Additionally, he wrote articles for the "Kölnische Rundschau" and various sports magazines, composed radio contributions for WDR II, shot video clips for creatv and cast candidates for the ARD show "Geld oder Liebe". In 2000 he moved to Berlin for his diploma thesis which was on the history of the German participation in the Olympic Games 1896. After an intermezzo in the PR sector, he changed over to the editorial department of meinberlin.de and tagesspiegel.de in 2003. Since January 2008 he is the Chief Editor for the areas of blogs, community and reader service at welt.de.

Grace Mutung'u is an associate at the Kenya ICT Action Network (KICTANet) . She is currently an OTF Senior Fellow in Information Controls at the Berkman Klein Center for Information and Society at Harvard University researching on information controls during elections in East Africa. She is an advocate of the High Court of Kenya and her interests include culture and contemporary arts.

Helmi Noman is a Research Affiliate of the Berkman Klein Center. His research focuses on internet censorship in the Middle East and North Africa; exploring the impact of information and communication technologies on the Arab information societies; how the use of the internet defies the social and political structures; and the potential systemic changes cyberspace can bring to real space in the Arab region.

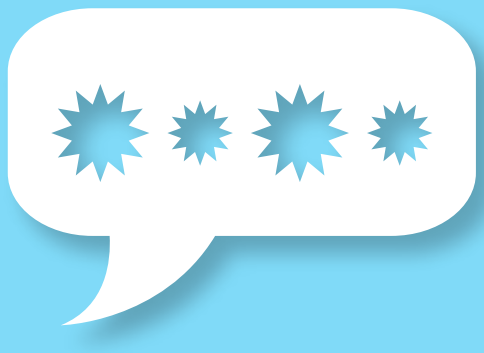
Felix Oberholzer-Gee is the Andreas Andresen Professor of Business Administration in the Strategy Unit at Harvard Business School. A member of the faculty since 2003, Professor Oberholzer-Gee received his Masters degree, summa cum laude, and his Ph.D. in Economics from the University of Zurich. His first faculty position was at the Wharton School, University of Pennsylvania. He currently teaches competitive strategy in executive education programs such as the Program for Leadership Development, the Senior Executive Program for China, and in a program for media executives titled Effective Strategies for Media Companies. Professor Oberholzer-Gee has won numerous awards for excellence in teaching, including the Harvard Business School Class of 2006 Faculty Teaching Award for best teacher in the core curriculum, and the 2002 Helen Kardon Moss Anvil Award for best teacher in the Wharton MBA program. Prior to his academic career, Professor Oberholzer-Gee served as managing director of Symo Electronics, a Swiss-based process control company.

Jonathon W. Penney is a legal academic, Research Fellow at the Citizen Lab, Munk School of Global Affairs, University of Toronto, and a doctoral candidate in information/communication sciences at the Oxford Internet Institute, University of Oxford (Balliol College). A recent Berkman Fellow and Research Affiliate at Harvard's Berkman Klein Center for Internet & Society, as well as a Google Policy Fellow at the University of Toronto, Jon's interdisciplinary research focuses, among other things, on human rights, privacy, censorship, and security, especially as they intersect with information law and policy. He also teaches law at Dalhousie University in Canada.

Anke Sterzing is a postdoctoral researcher at the University of Magdeburg and an affiliate at the Berkman Klein Center for Internet & Society at Harvard University. She holds a Ph.D. in economics from the University of Kaiserslautern, as well as a Diploma (M.Sc. equivalent) in business economics from the University of Magdeburg. She has been a visiting scholar at University of Chicago, Yale School of Management, University of Zurich, and University of California, Santa Barbara. Her research work is dedicated to the study of behavioral economics and media economics. Anke is currently working on whether online hate speech influences cooperative behavior.

Casey Tilton is a project coordinator for the Internet Monitor project at the Berkman Klein Center for Internet & Society. His recent work includes researching the means and extent of state-sponsored Internet censorship around the world.

Lumi Zuleta holds a Master's degree in cultural studies. Since 2009, she has been a project manager at the Danish Institute for Human Rights where she works with equality and non-discrimination in Denmark, with a particular focus on gender equality. Her latest publication is a report on hate speech in the public online debate.



HARMFUL SPEECH ONLINE