Working Meeting on Youth-Oriented Online Hate Speech

Summary

Spring 2015
Harvard University

I. Introduction

This working meeting, co-hosted by the Berkman Center for Internet & Society at Harvard University and the Institute for Strategic Dialogue, brought together participants from multiple fields and perspectives, including academics, practitioners, and company representatives, from the Berkman and Institute for Strategic Dialogue communities who care deeply about youth-oriented hate speech online. The meeting was envisioned as one in a series of conversations, including both real-space and virtual engagements, which will deepen our understanding of the online hate speech phenomenon as it relates to youth and of the development, implementation, and evaluation of strategies to counter such speech. Certain conversations may focus on narrower topics within a broader set of hate speech phenomena that concern youth; the March 31 working meeting paid particular attention to violent extremism and its potential countermeasures, and sought to locate them within the broader set of relevant phenomena.

Two broad thematic areas structured discussion during the meeting: problems and solutions, respectively, related to youth-oriented online hate speech and violent extremism. Within this structure, discussion ranged widely across numerous themes (within and adjacent to the two broad areas noted above), reflecting the participants’ diverse backgrounds and perspectives. This report seeks to concisely summarize the conversation’s main themes and highlight certain suggestions for future action the meeting generated. In the following section, the main themes and observations listed include issues that participants dealt with explicitly and at length, as well as those that more quietly (and sometimes implicitly) surfaced at multiple points during the afternoon. Section III, which concerns suggested approaches for future action, does not comprehensively describe all ideas, large and small, that came up. Rather, it highlights a few suggestions that elicited a significant amount of interest from meeting participants, or which crystallized through the course of extended discussion and exploration.

II. Main Themes and Observations
A. Rhetorical Tactics for Spreading Online Hate

From the beginning of the meeting, participants wrestled with the question of how to conceptualize the relationships between and among disparate manifestations of hate speech and hateful online behavior, especially those that disproportionately concern or affect youth. Some drew attention to the proliferation of various hateful online behaviors that do not fit neatly within the bounds of the traditional conception of hate speech, which is concerned with different forms of discrimination and bias—e.g., racial, ethnic, religious, sexual, or gender-related in nature—expressed through language. Slurs are a paradigmatic example of such expression. Looking at a wide range of online behaviors that constitute, approximate, or evoke hate speech—such as revenge porn, doxxing, misogynist abuse, and cyberbullying, to name a handful—shows that the rhetorical forms used to convey hate on the Internet are more varied. Critical race theory, which emerged from the legal academy, offers a conceptual framework for understanding the negative impact of such varied rhetorical forms on targets. For instance, in the case of revenge porn the hateful expression involves, at minimum, the posting of content intended to harm a victim; in the case of doxxing, it is the dissemination of personally identifiable information. Abuse that many women suffer online involves being the target of misogynist speech, consonant with the traditional conception of hate speech. But the abuse can also entail rape and death threats—expressions surely fueled or informed by misogyny, yet which state things about an individual rather than a group. Similarly, online bullying can overlap with identity-based discrimination, but can also involve harmful expressive behaviors that do not express bias against a group or demographic.

The ways in which propaganda is used online as a rhetorical tactic by extremists merits scrutiny. Multiple participants were keen to explore the ways in which traditional hate speech and propaganda (along with other forms of rhetoric, such as incitement to violence) are distinct yet intertwined. Propaganda often functions to advance extremists’ work of cultivating hatred and distrust of target groups, but may not express hatred in a direct or heavy-handed manner (as do slurs). Tactics used by extremist groups in their creation and dissemination of propaganda online include: presenting a vision of “the good life” that prospective followers could enjoy (upon embracing the extremists’ ideology); depicting the extremist group as the victim of wrongdoing by the target group; disseminating false information about the traits and actions of the target group (but in a manner that appears reliable and objective); distorting, or taking out of context, certain truths or facts; appealing to positive emotions such as pride and a sense of belonging to a social group; and, channeling feelings of fear and frustration into hatred and blame of the target group.

B. Main Types of Solutions (current and potential)

Participants discussed a variety of approaches to addressing the broad problem of youth-oriented hate speech online, including approaches that are (or have been) used as well as those that remain untested. The approaches fell roughly into three overlapping categories: prevention, technological regulation, and counter-narrative.

Two types of prevention were brought up: efforts targeting children and young people early in life that reduce the risk of becoming radicalized, or hateful speakers; and, efforts targeting those going through the process of radicalization from endangering others. In the context of the latter, participants wrestled with the risks and challenges of employing network analysis of social media.
data. Overall, however, discussion focused more on the former kind of prevention—specifically, the ways in which education, broadly defined, can serve this purpose. Education of children and young adults can cultivate morals and strengthen social norms that reduce the attraction of hateful behavior later in life. In addition to schools—which, at least in the U.S., remain the primary venue for reaching youth, especially disaffected ones—teaching and learning take place in real-space settings such as libraries and homes, as well as through media content (e.g., cartoons) and on social media platforms. Digital literacies and social justice-oriented curricula were two particular areas suggested as worthwhile for purposes of prevention. Additionally, several participants spoke about their direct experience creating or working with media content designed to impart moral lessons, such as appreciation of diversity and the importance of tolerance.

Technological regulation strategies that private companies could (or do) use to block, flag, monitor, and remove content (and users) were a subject of much discussion and some debate. Content removal was the focal point of much of this thread of conversation. Some participants mentioned examples of how takedown of content effectively supported counter-extremism work (e.g., against ISIS); others questioned the efficacy of takedowns for this purpose, and expressed concerns about freedom of expression. The limitations of social media companies’ systems for reviewing content were also discussed. It is difficult for human reviewers to take context and the speaker’s intent into account (given the volume of content to be dealt with); it’s as of yet unclear what the implications of using algorithms for this purpose would be. Without the capacity to attend to account for context and intent, companies’ human reviewers risk removing speech that is politically or socially valuable (and protectable), while leaving up content that harms people.

The third main approach highlighted was the use of counter-narratives. Counter-narratives seek to diminish the attraction and impact of hateful narratives by engaging and directly combating them. Strategies that employ counter-narratives are best suited for those who have already begun down the path of radicalization, and can be carried out by peers, authority figures in young peoples’ lives, governments, and social media platforms. Specific strategies suggested included correcting misinformation and connecting youth at risk of radicalization with adults whom they can learn from, such as former extremists. Counter-narrative work clearly overlaps with efforts to cultivate morals and strengthen norms discussed in the context of prevention. Tapping into emotions (as does propaganda, as discussed above) is important for counter-narrative work.

Suggestions that do not fit neatly in any of these categories, such as improving legal tools, modifying the architecture of social spaces online, and modifying reporting/flagging process online, were also briefly touched upon. Some participants proposed that when content is flagged or reported online, both the flagger and the reporter should have the opportunity to connect with relevant educational content or other humans—authority figures—from whom they could learn. Others contemplated the ways in which social media sites could offer users simple ways to express more nuanced responses (e.g., critique and disagreement, rather than mere dislike) to others’ speech.

C. Value of Collaboration

Participants considered a variety of collaborations that could leverage the diverse backgrounds represented at the meeting to improve existing projects create new ones. Suggestions included partnerships among teachers, parents, media makers, and other educators; outreach efforts or campaigns led by former extremists together with non-profit organizations or governments (former
extremists’ narratives’ have credibility, non-profits and governments have reach); and, research (data collection and analysis) and design (informed by that analysis) enabled by collaboration between social media companies and academics.

Some participants mentioned that potential collaborations could productively involve some combination of the following specialties and interests: target groups (e.g., Muslims, women, people of color); different types of hate speech (e.g., cyberbullying, racism, extremism); different stages of radicalization (e.g., after first exposure, partially radicalized, fully radicalized); spaces for education (schools, libraries, the Internet); different disciplines’ traditions and methodologies for data analysis (e.g., critical race theory, public health, network analysis). In addition to project-based work and ongoing exchange, participants noted that one-time feedback sessions, and even the practice of drawing attention to one another’s work, would be valuable.

D. Limitations of Generalized Approach

Discussion of the limits of generalization and abstraction came up repeatedly. Grouping together, or juxtaposing, disparate manifestations of online hate speech and hateful behaviors together has its merits, but some expressed concern about the risks of adopting a “one size fits all” approach to conceptualizing, studying, monitoring, and combatting hate speech. Participants talked repeatedly about the importance of maintaining a nuanced understanding of cultural, social, and historical context, related to which are concerns about self-determination and cross-cultural respect. Moreover, some participants noted the difficulty of extracting broadly applicable insights from individual interventions, especially in the context of counter-narratives. Skepticism of the use of the term “hate speech” to encompass diverse phenomena was also mentioned (though participants did not devote much time to brainstorming—or agreeing upon—alternatives).

E. Interaction Between Online and Offline

The question of the permeability of the online/offline divide was regularly alluded to throughout the proceedings, especially in discussions about issues related to education. Participants wondered how much online experiences, such as reporting content, having one’s content removed, or encountering counter-narratives, affect life offline. Would a young social media user told that using a certain slur on a social media site is unacceptable then consider the implications of using the word in the classroom? If not, how could the young person be taught to do so? Much of the discussion involved exploring the effects of online experiences (namely, exposure to extremist content, victimization, and radicalization) on offline behavior. On the other hand, participants spoke with greater confidence about effects in the reverse direction, offline to online, as indicated by the emphasis on offline education as a cornerstone of proposed prevention and intervention strategies.

Participants also debated how much of the online hate speech problem is simply a manifestation of a pre-existing condition or offline problems (such as quality of life, material circumstances, fear, etc.), and how much of it is specific to the online space. To what extent do anonymity, mob mentality, and disinhibition, which are often associated with the Internet, cause, accelerate, or otherwise shape hateful behavior? This question is especially important in the case of extremists and others who perpetrate acts of violence offline.

F. Need for More Data
Throughout the discussion, participants expressed a keen interest in new opportunities for quantitative data collection (especially from social media sites) and analysis, and new methodologies for such analysis. One particularly appealing idea suggested by a participant would be to design a study to evaluate the effects of counter-speech/counter-narrative techniques in multiple countries so as to generate comparable cross-country data. Two current data-driven efforts, which could inform or inspire future efforts, were mentioned. One was a network analysis of approximately 40,000 Twitter accounts, which sought to identify users with significant influence in their social networks who were disproportionately responsible for catalyzing or contributing to others’ radicalization. In addition to identifying such influential users, the study indicated the value of cross-platform analysis to understand social networks. The other collected data on views of and engagement with counter-narrative YouTube videos. Data collected included views, comments, and shares, and other metrics provided by Google Adwords.

However, the limitations of such data sets were also highlighted. Participants spoke to a need for more robust metrics and analyses that focus on longer-term outcomes for those who are targeted by intervention and prevention efforts. These metrics could help inform future work by changing the way companies design flagging and reporting processes, informing decisions to attempt to scale or generalize interventions, and direct future research more effectively.

III. Suggested Approaches for Future Action

A. Engagement with Emotions

The primacy of emotions in hate speech and extremism was addressed multiple times while discussing both the problems of and solutions to online hate. Participants noted that emotions play a large role in recruitment, as hateful propaganda frequently manipulates existing emotions, such as fear or frustration, and seeks to cultivate new ones, such as loyalty (to a particular group). In order to reach radicalized individuals and those that are in the process of being radicalized, appeals to these same emotional forces may be necessary to penetrate their newfound beliefs and dismantle the propaganda produced by the other side.

Several participants affirmed that appeals to logic and reason are more effective when they are built on a foundation of emotional engagement. Such a foundation could be established with messaging that recognizes and validates the existence of anger or pride, without letting them fester and grow into hate. Finally, it is vital to direct recipients of the messaging another toward opportunities for action that are outlets for emotions and alternatives to hateful behavior. This approach—which has emotional, logical, and practical elements—should be incorporated into prevention and intervention efforts of various stripes, including one-on-one counseling or conversations with former radicals, reporting and take-down processes, or the design of educational media.

B. Collaboration with the Private Sector

The promise of greater collaboration between the private sector and others (academics, educators, other practitioners, etc.) was remarked upon multiple times. Such collaboration could pursue
various ends. For example, more exchange about companies’ processes for reviewing flagged and reported content could benefit both sides. Other issues that further exchange among these groups could discuss further together include the design of platform architecture (opportunities for more nuanced expression of dislike, disagreement, etc., as mentioned above); ways to balance neutrality and “being on the right side of the problem,” and the related question of the pros and cons of content removal; etc.

C. Important Elements of Counter-Speech/Counter-Narratives

Participants stated that practitioners engaged in counter-speech/counter-narrative work should learn from the branding and advertising techniques employed by private sector actors to identify, reach, and engage their target audiences online. Other important factors to keep in mind when crafting counter-speech/counter-narratives are the importance of: professional, polished content; appearing credible to target audiences (note that credibility is distinct from reliability and accuracy, and that different audiences may conceive of credibility differently); appealing to emotions, in addition to reason and pragmatism (as discussed above); and using audience engagement tools such as Google AdWords can to target ads and help get content to the people who most need to see it.

D. Working Groups

As noted previously, the multifaceted nature of youth-oriented online hate speech makes collaboration all the more vital to increase our understanding of the problem and our ability to combat it. One specific proposal put forward was for meeting participants to form three working groups focused on the state of knowledge on online hate speech (i.e., mapping the landscape), data collection and methodology, and the youth angle (i.e., a youth and media lab), respectively. The mapping group might seek, for instance, to aggregate resources and build on conversation initiated at the meeting to identify knowledge gaps between different disciplines and fields. The data group could adopt a similar approach, but focus specifically on opportunities to design, replicate, or advance quantitative studies. The youth and media lab could think deeply about how to engage young people in the process of exploring youth-oriented online hate speech and incorporate them into the process of directing future activities. Participants were invited to take part in any of the groups that interest them; the teams from the Berkman Center and Institute for Strategic Dialogue offered to help the working groups get off the ground.