Exploring the Role of Algorithms in Online Harmful Speech

August 2017

By David Talbot and Jeff Fossett

Originally Published by the Berkman Klein Center

Reflections from the workshop
Harmful Speech Online:
At the Intersection of Algorithms and Human Behavior
Held June 29-30, 2017

Sponsored by:

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

HARVARD Kennedy School
SHORENSTEIN CENTER
on Media, Politics and Public Policy

ISD
Powering new generations against extremism

Table of Contents

Introduction

The topic of online harmful speech—from harassment and cyber-bullying to terrorist recruitment and media manipulation—is a growing focus of academic research and government regulation. On June 29 and 30, 2017, the Berkman Klein Center, the Shorenstein Center on Media, Politics and Public Policy at the Harvard Kennedy School, and the Institute for Strategic Dialogue (ISD), a London-based think tank, co-hosted "Harmful Speech Online: At the Intersection of Algorithms and Human Behavior" to discuss how market dynamics, behavioral drivers, laws, and technology contribute to the spread of harmful speech online and inform measures to constrain it. This report provides a sample of the conversations that took place during the event.

The workshop, which brought together a diverse set of experts on the topic, grew out of prior work from each of the host institutions. The Berkman Klein Center has been actively studying harmful speech online for the past two years. A recent report summarizes some of this work and broadly describes research on the problem. The related problem of online false information, or "fake news", prompted the Shorenstein Center to host an event in February and later chart a research agenda. And on the matter of confronting online terrorist recruitment and extremist content, ISD has led significant research and worked extensively with industry and government to develop possible interventions and policy approaches to respond to this rising set of challenges.

Framing the Issues: "Uncharted Territory"

The event was held at Harvard Law School and convened over 60 stakeholders from academia, civil society organizations, and major technology companies—including Facebook, Twitter, Google, and Microsoft—and featured talks, group discussions and breakouts. The hosts included Rob Faris, Berkman Klein Center research director; Urs Gasser, the center's executive director; Sasha Havlicek, CEO of ISD; and Nicco Mele, director of the Shorenstein Center. The major goal was to foster collaboration and idea-sharing; not release new findings. The event was conducted under the Chatham House Rule; participants were recontacted to provide permission to use quotations or attributions appearing in this report.

Introducing the topic, Faris, Mele, and Havlicek pointed to the enormous gap—in terms of resourcing, activism, and even basic research—between the problems of harmful speech online and the available solutions. Harmful speech and extremism in online spaces can have an enormous impact on public opinion, inclusiveness, and political outcomes. And as Mele put it, we are in "uncharted territory" when it comes to addressing these problems, which underscores the importance of convening groups from academia, civil society, and industry to address the challenges.

Faris said the topic was one of the central challenges of the digital age and a serious impediment to greater and more inclusive online participation. But he cautioned that proposed remedies—which inevitably mean removing online content—hold the potential to be worse than the problem and veer into widespread censorship. His comments were prescient: weeks later, Russia enacted a law modeled after the German one. (For context, a recent Berkman Klein Center report documents trends in global internet censorship as measured by empirical research performed by Berkman Klein researchers.)

Havlicek described the adept migration of extremist groups online and the resultant hypercharging of their global reach and influence. According to Havlicek, there is currently a gap in professionalism, capacity, networks, resourcing, and technological skills which is enabling extremist groups to drown out more moderate voices online. Through a range of pilot studies, many in close partnership with the tech industry, ISD has developed a body of evidence about the types of interventions that can work for addressing online extremism— algorithms and tailored communication strategies with individuals at risk of recruitment, for example. However, in order to achieve scalable, long-term solutions, more resources and cross-sector collaboration is required, Havlicek said. She highlighted the importance of drawing from research and practice across different forms of online harm in order to develop more effective responses and avoid “reinventing the wheel”. Indeed, it was in part with this need in mind that the event was conceived among the partner organizations.

Who will lead the way? Mele discussed the example of Henry Luce, the publisher of *Time* magazine in the 1930 and 1940s and one of the most powerful men in the United States. Luce often promoted conservative candidates and causes. He also convened people to think about the power of the press to shape public opinion and policy—most notably by funding the Hutchins Commission, a body made up mostly of academics that was charged with evaluating how well the news media was serving the public good.

Mele pointed out that the institutions that influence and control public opinion today are very different and that we are consuming a “harmful information diet”—much of it served online, and against a backdrop of the disappearance of conventional news media organizations across the United States. (A recent report by Berkman Klein Center researchers showed, in part, how hyperpartisan and outright fake news spread on Twitter and Facebook.) Mele asked rhetorically whether anyone could imagine Mark Zuckerberg, the founder of Facebook, doing anything similar today to the Hutchins Commission’s inquiry on the role of the media in a democracy. Implicitly answering his own question, he said it was up to the people in the room to help figure out how to deal with harmful speech online.

Defining and Documenting Harassment

Although conference participants didn’t attempt to draft or circulate a definition of “harmful speech,” (the difficulties of this task were articulated in a 2016 Berkman Klein Center report), the term is generally understood to encompass many forms of online content that cause a variety of harms, from emotional distress to physical violence. Such content includes verbal harassment and stalking, threats or depictions of violence, nonconsensual image sharing (such as so-called revenge porn), violent or extremist content (such as terrorist recruitment), and fake news and misinformation.

Hard data on the rise of online harmful speech are also elusive, though survey data provides some insights. Amanda Lenhart, a senior research scientist at the Associated Press-

NORC Center for Public Affairs Research, said that the problem of harmful speech online is pervasive; in a survey she helped conduct while at another organization, Data & Society, 47 percent of Americans over 15 said they had experienced at least one instance of online harassment, and 72 percent said they had witnessed online harassment or abuse. (A more recent survey produced similar findings.)

Lenhart also found substantial differences by gender: while men and women were equally likely to have experienced some form of harassment, women generally faced a wider variety of online abuse, including more serious violations such as sexual harassment and long-term harassment, while men were more likely to experience abusive name-calling or physical threats. Men and women also experienced harassment differently, with women being almost three times as likely to say that an experience of harassment made them feel scared. These different experiences of harassment underscore the importance of considering diverse perspectives when defining and addressing harassment online.

The Challenges of Content Moderation at Scale

Today, pressure on major internet platforms to remove content is sharply rising. During the same week as the event, lawmakers in Germany approved a bill aimed at forcing major internet companies to banish “evidently illegal” content (Germany has tough laws imposing prison sentences for certain forms of prohibited speech, such as Holocaust denial) within 24 hours or face fines that can range up to $57 million. Such developments have already prompted some major online platforms to explore the deployment of algorithms to detect and speed up the removal of content that runs afoul of national laws or platform policies.

In theory, algorithms could be used to automatically identify and filter or flag potentially unsavory content; in the ideal situation, humans confirm that algorithmically flagged material (or material flagged by users) indeed violates platform policies or other regulations or laws. But as workshop participants pointed out, the methods used by Twitter and other companies are typically opaque: we often don’t know how their algorithms work, how they train their moderators, and so on. In a recent Shorenstein Center lecture, Jeffrey Rosen, president and CEO of the National Constitution Center and law professor at The George Washington University Law School, argued that these processes are sometimes inconsistent with First Amendment principles.

And sometimes these processes, despite being rooted in internal policy, seem to be in need of improvement. One insight into industry policies was provided by a ProPublica story released just a few days before the convening. The story, which became a topic of discussion, covered the confidential guidelines that Facebook provides to its team of “content reviewers,” whose job it is to decide which posts are allowed or deleted, and the sometimes strange outcomes that result.

Facebook’s guidelines draw clear (if sometimes complex) lines between what is or isn’t considered problematic content that should be immediately removed. For example, Facebook’s hate speech policies protect “subsets” of protected categories when both the group and the subset are protected but not when only one is. On this logic, “White Men” is protected because both race (“White”) and gender (“Male”) are mentioned. But “Black Children” is not, since race (“Black”) is a protected category, but age (“Children”) is not.

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

In discussions on the topic of hate speech, Monika Bickert, the head of global policy management at Facebook, defended the platform’s policies, though she admitted that “the policies are not perfect.” Bickert emphasized the challenges that platforms face in moderating content at scale and highlighted “the need to draw bright lines to allow for consistent global practices.” She noted that Facebook continues to revisit and improve its moderation policies, including those around hate speech. In public statements on this topic, Facebook has also highlighted the risks of over-moderating online, noting that “it can feel like censorship” if Facebook removes a piece of content that users believe to be a reasonable political view.

Challenging the Platforms

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

Tufekci recounted that platforms have been slow to respond to horrific content in certain contexts. ISIS, during its takeover of Mosul in 2014, used YouTube and Twitter to post videos of its beheadings and other atrocities against local populations to effectively terrorize the city and help the organization take control, Tufekci said, but the platforms only really took action to remove such content in later stages of the ISIS campaign, once the beheading victims were westerners.

The Role of Algorithms in Producing and Addressing Harmful Speech

Other conversations more broadly explored the role of machine learning, algorithms, and artificial intelligence in both producing and addressing harmful speech online. For example, Camille François, a principal researcher at Jigsaw—a think tank and technology incubator within Google—discussed her group’s recent partnership with the *New York Times* to develop a machine learning tool to help the Times moderate its comment sections online.

The new tool, called Moderator, can automatically prioritize comments that are likely to be in need of review or removal, easing the job of content moderation. The tool was trained on more than 16 million moderated Times comments, and has allowed the Times to substantially increase the volume of commenting it allows. François emphasized the importance of transparency and collaboration in developing automated tools of this sort, and she also highlighted the value of data and experimentation.

Although clever machine learning models may provide some solutions for moderating harmful speech online, algorithms can also be part of the problem. As Tarleton Gillespie, a principal researcher at Microsoft asked in one session, “What role have our algorithms played in *calling forth* harmful speech online?” Gillespie explained that the sorting, filtering, and recommendation algorithms that structure many online spaces—such as Twitter’s “Trending” or Facebook’s News Feed—often form the “terrain” on which harmful speech occurs. Even as these algorithms may have been designed to promote the “best” content, they can also empower, surface, and aggregate harmful content in ways that platform designers may not have anticipated. For example, recent research has studied how partisan actors have used automated Twitter accounts, or “bots,” to attempt to manipulate political conversations (and trending algorithms) online.

YouTube faces similar challenges. Zeynep Tufekci noted that if you watch a Donald Trump campaign rally on YouTube, the site will suggest that the next thing you may want to watch is a collection of white nationalist videos. The algorithm may “push you down a radicalization rabbit hole because this looks like ‘engagement’ to the algorithm,” she said. In her view, some “recommendation” algorithms bring people to the fringes and drive polarization online, whereas a functioning democracy requires people to convene. Several weeks after the conference, YouTube released a blog post promising to implement tougher standards on extremist content, including placing flagged videos in a “limited state,” in which they cannot be monetized through advertising and won’t be recommended to users.

Gillespie also raised questions about the second-order effects of using algorithms to moderate online content. For example, he raised concerns about the impact on constituents’ sense of autonomy if algorithmic decision-making goes unexplained. Will users still feel in control if Twitter, say, automatically removes a post or comment without explanation?

Broader use of such algorithms in media—including social media—is one reason why the Berkman Klein Center is exploring “media and information quality” as one of three major focus areas of its AI Ethics and Governance Initiative, a joint effort with the MIT Media Lab that will explore how algorithms and AI can best serve the public interest.

Oversight on Industry Algorithms

These sorts of questions are likely to take on increasing significance. Susan Benesch, a faculty associate at the Berkman Klein Center and founder of the Dangerous Speech Project, explained that the internet industry—under increasing pressure to remove and regulate content—was entering a phase in which algorithms and artificial intelligence will be deployed at enormous scale to take down content and that these processes need to become more transparent. There should be some oversight or auditing of takedown, she said, by people who do not work for the platforms, such as researchers.

She also suggested that “regulation of online content is a form of law.” Facebook and other platforms all have documents governing use of their services, often called community guidelines, or standards. These are analogous to constitutions in that they offer general rules that are comparatively vague. The rules take on clearer meaning when they are applied, by company staff (or algorithms) deciding which content to censor. These decisions are analogous to case law. But unlike decisions by courts, platform decisions are generally secret. “All of this de facto law is operating in the dark,” Benesch said.

She challenged workshop participants to think about the ways that platforms might allow democratic oversight of algorithmic decisions over content. “There is absolutely no doubt that the platforms will make increasing use of algorithms for ‘takedowns,’ which is the term they use. It’s also legitimate to use the term ‘censorship,’” she said. “And now platforms and governments are going to regulate speech on a much larger extent and scale. This is super scary; people are going to get used to this; most people are not even going to be aware of it. Content will simply evaporate.” She also encouraged the group to look beyond takedowns, which have been described as a losing game of Whack-a-Mole, as the lone solution to harmful speech online. Researchers and internet platforms should work together on experiments to discourage the posting of harmful speech in the first place, she said.

A Proposal From Europe

Dan Shefet, a Paris attorney specializing in internet, intellectual, and competition law, offered one approach to creating a common accountability platform: his proposed creation of an “Internet Ombudsman,” whose office would establish a process to decide whether online content was lawful or unlawful according to national laws.

At a recent Council of Europe meeting, he made a motion that described how this would work. Through the ombudsman, platforms could address good-faith questions about content moderation and receive recommendations about what to do. Platforms would have a choice about whether to follow the recommendations but would have a strong incentive to follow them in that by doing so they could receive immunity from fines.

Shefet has already had an impact. In 2014 he forced Google to remove links to defamatory information about him; the ruling meant that anyone in an E.U. nation affected by Google’s links to, say, a libelous story, could obtain an injunction against the local Google subsidiary rather than having to get a judgment against Google in the United States. Since then he has sought ways to make sure that search engines delist harmful content globally and established the Association for Accountability and Internet Democracy. The debate over removal requests and freedom of expression remains robust.

Ideas From Participants

During an interactive session, participants were invited to describe other ideas for how to handle the challenges. Many of the ideas shared themes: obtaining data from the major platforms to enable basic research and collaboration; setting up an oversight or auditing mechanism for current and future algorithms used by platforms; and systematically generating, testing, and scaling algorithms that could improve methods involved with online recommendations and takedowns.

Users also proposed specific steps, including adding small forms of “friction”—delays, confirmatory steps, or warnings—to allow users to reflect before impulsively sharing content, and creating alerts when the entity responding to a tweet is a bot. Participants additionally cited the need to aggregate, source, and target existing counternarrative material to present to people who are at risk of becoming radicalized, as well as the role of education in building digital literacy and resilience against online hate. In general, participants cited a need to use open source code and open data to replicate, repeat and test interventions.

Major questions posed by the group included: Who would pay for the necessary work at scale? How can companies share data without violating privacy? How can industry collaborate with researchers to review how changes in the user interfaces can be audited?

Conclusion

On the closing day of the event, which was supported by the MacArthur Foundation and the Siezen Foundation, Gasser said that questions about technology inevitably lead back to humans. Engineers who build these systems need to have an ethical mindset, and business leaders and business models need to be moral and reflect broader societal interests. In addition, a core challenge is how we reconcile the needs of an increasingly heterogeneous society with the centralization of dominant online platforms. In the future, will we see a broader and more diverse array of platforms to reflect the diversity of society? And finally, though many issues are local and contextual, governance also plays a role in drawing bright lines when doing so is possible and morally necessary.

Many participants said they learned new things about what others are working on in the space (this reading list was also distributed) and developed an appreciation for the fact that the scale of algorithmic takedown of speech defined as or called “harmful” by governments or courts would accelerate, increasing the need for oversight. Findings from the sessions will help shape written reports, mechanisms for network building, continued knowledge sharing across sectors, and avenues for future collaboration.