SCHOLARS
STRATEGY NETWORK

Countering Online Toxicity and Hate Speech

Ben Miller, University of Canterbury

People communicate online to connect to family and friends, manage and collaborate with coworkers, obtain news, and participate in their communities. An increasing part of people's social time, professional life, play, and training takes place in the digital world, and new research suggests that people transfer behaviors they learn in online settings to their everyday social and professional activities. Online chats in games, online comments about news articles, and bulletin boards for online communities could, in principle, further civil discussions. Unfortunately, digital culture, despite forging new connections among many kinds of people, turns out to be rife with hate speech and harassment.

Understanding Online Toxicity

According to research on computer-mediated communication by Lincoln Dahlberg and others, trolls were originally understood as mischievous tricksters trying to be annoying or disruptive. In early Internet culture trolls created false personas to integrate into an online community and ultimately derail group conversations. Since then, however, the term has become a catchall to describe any sort of antisocial or disruptive online behavior. A term now heavily used by reporters, "trolling" is frequently used interchangeably to refer to bullying and hate speech, muddying the waters around the word's definition and descriptive power. As a catchall media label, "trolling" invokes a kind of nebulous Internet folk devil rather than an actual person or persons behind the computer screen. It obscures the underlying hate speech. If observers were to shift away from such uses of "troll" and "trolling," they could actually name specific toxic behaviors the sexism, racism, homophobia, transphobia, that they actually represent.

Toxic behavior is pervasive in every online environment. Maeve Duggan's "Online Harassment," a study released by Pew Research in 2014, leads with the finding that 40% of internet users have faced harassment and 73% of users have seen others get harassed. Although physical threats were only witnessed by a quarter of respondents in this study and only 8% said they were physically threatened, these numbers misrepresent experiences online. Seven in ten of Internet users aged 18 to 24 have been harassed while online; and 26% of women in that age group report being stalked online. Such statistics provide a first glimpse at the scale of the problem of the toxic online environments, and they show that common practices of community self-selection fail to address harassing online behaviors.

Recent research shows that toxicity also exists across online gaming groups, and is not isolated to a particular game or specific player community. Alexis Pulos' research finds that player posts to online forums like the World of Warcraft player community often create a culture of hostility toward gay, lesbian, bisexual, and transgender people. Similarly, Kishonna Gray's ethnography of the Xbox Live gaming community reveals a constant barrage of gendered and racially motivated harassment directed at women of color who opt to communicate with teammates via voice chat. Problems are worsened by gaming community leaders who

claim that gender-based harassment is a “non-issue” and dismiss their responsibility for fostering rape cultures. As these evasions show, the industry will likely be resistant to change unless external pressure is applied. Yet unless hostile online behaviors are reduced, vulnerable people, marginalized groups, and the public generally will all be further harmed.

Recent efforts to understand and respond to such pervasive toxicity include a 2015 panel on online harassment convened by Caroline Sinders at South by Southwest; the 2017 workshop on Abusive Language Online held at the annual conference of the Association for Computational Linguistics; and the “Notoriously Toxic” project funded by the National Endowment for the Humanities that brought together a working group of game developers, legal experts, social scientists, computational linguistics, and humanists. Also relevant are experiments like Google Jigsaw’s Perspective, which attempts to use a machine learning classifier to classify text strings on a scale from “very toxic” to “very healthy.”

From Online Toxicity to Offline Hate Speech

Online hostility goes hand in hand with offline hate speech directed at targets defined by gender, race, class, ethnicity, nationality, and various individual vulnerabilities. In one case, described by Lisa Nakamura, Chinese players of the online game World of Warcraft were verbally attacked and harassed in-game and via social media by attackers who used using stereotyped images identical to those deployed against exploited Chinese railroad laborers in the mid-1800s. Online communities can be disrupted by anti-social behavior in games like League of Legends, Call of Duty, and Minecraft, along with harassment on social media platforms like Facebook and Twitter. However, such pathologies could be reduced with social codes and regulatory systems.

Ways to Address Online Pathologies

Efforts to reduce online toxicity can target players, groups of players, or the servers on which games take place. Each relies on different type of approach. Work with players focuses on the psychology of individual gamers. A focus on player groups requires an understanding of social structures. And a focus on the servers that host games requires an understanding of how regulations can influence population-level behaviors. Techniques aiming to moderate conduct can operate at each of the three levels and strike some balance among commercial, regulatory, technical, and ethical concerns. Some techniques do natural language processing to filter individual comments; others encourage community self-moderation. A blended approach might apply a language filter, a censoring mechanism, and a behavior warning system, bolstered in extreme instances by litigation against offenders or bans for spans ranging from minutes to lifetimes.

Although no one could reasonably argue that the average gaming company encourages use of its platforms to radicalize young people or intentionally disseminate hate speech, game players nevertheless face such pathologies every time they go online. If nothing is done, online social and professional interactions will likely be plagued by ever more hostile modes of communication. To protect the next generation from the flood of hate speech encountered by today’s youth when they search the Internet and chat friends or play games online, policymakers and other advocates must address the systems-level features that foster and even encourage toxic online speech. A mix of approaches must be tried, until effective, sustainable solutions are devised.

Read more in Ben Miller, "Notoriously Toxic: Understanding the Language and Costs of Hate and Harassment in Online Games" (working paper, 2018).