GLOBAL CENTRE
FOR THE RESPONSIBILITY
TO PROTECT

16 February 2020

Hate Speech and Social Media: Preventing Atrocities and Protecting Human Rights Online

Speech delivered by Dr. Simon Adams on 16 February 2020 in Doha, Qatar, at the International Conference on "Social Media, Challenges and Ways to Promote Freedoms and Protect Activists," hosted by National Human Rights Committee of Qatar, UN Office of the High Commissioner for Human Rights, European Parliament, Global Alliance of Human Rights Institutions, and the International Federation of Journalists.

Social media allows us to connect across borders, to communicate more easily than at any other time in human history, and even to expose human rights abuses in faraway places. But in this new digital era, Facebook, Instagram, YouTube, Twitter and other platforms are not just places for information sharing and social networking, they are also places where vilification, targeting and incitement take place. Hate speech is not only proliferating in the dark corners of the internet, it is increasingly common on all major social media platforms.

Hate speech is speech that marginalizes and targets people on the basis of their religion, ethnicity, gender, sexual orientation or race. In other words, it is language that demonizes and threatens people not for anything that they have done, but for who they are. And while there are legitimate debates about the need to defend freedom of speech online (including speech that some might consider "offensive"), there should be no freedom of speech for genocide deniers, for extremists who deny the fundamental humanity of others, and for those who want to use words to try to build a path to the concentration camp or the mass grave.

For those of us who work in the world of human rights, the question of hate speech and incitement is crucial to our understanding of how and why mass atrocities are committed. If you examine the two exhaustive paragraphs of the 2005 UN World Summit Outcome Document where R2P was adopted, there is a very deliberate use of three words, enclosed within two discrete commas:

138. Each individual State has the responsibility to protect its populations from genocide, war crimes, ethnic cleansing and crimes against humanity. This responsibility entails the prevention of such crimes, including their incitement, through appropriate and necessary means…

The inclusion of *incitement* was a conscious political decision. Inflammatory rhetoric or hate speech targeting people on the basis of their identity creates an environment for the potential commission of atrocity crimes, especially when political leaders drag these poisonous ideas into public discourse and the media.

The propagation of hate speech has taken many forms throughout history as media platforms have evolved. The Nazis used virulently anti-semitic newspapers like Der Stürmer to help incite the German people into active persecution of Jews. Five decades later, during the 1994 genocide in Rwanda, the RTLM radio station played a key role promoting hate speech regarding the Tutsi, inciting the genocide, and even directing the killers with the names of potential victims who had not yet been murdered.

Most recently, social media has been weaponized to target "the other," and incite violence. Hate speech is used, in particular, by those who want to reshape public discourse and discriminate against ethnic and religious minorities, migrants and refugees.

Ralph Bunche Institute
for International Studies
CUNY Graduate Center
365 5th Ave., Suite 5203
New York, NY 10016 USA

info@globalr2p.org
T: +1 212-817-2104
F: +1 212-817-1565

www.globalr2p.org

A very disturbing example is Myanmar where in 2017 the military systematically utilized Facebook as a tool in their persecution of the Rohingya Muslim minority. The military harnessed Facebook over a period of years to disseminate propaganda, false news and inflammatory anti-Rohingya posts. In 2018, the UN Fact Finding Mission for Myanmar (FFM) described the role of social media as having had a "determining role" in inciting atrocities against the Rohingya. It reported that "Facebook has been a useful instrument for those seeking to spread hate in a context where, for most users, Facebook is the Internet."

Before 2015, Facebook only had English-speaking staff reviewing content posted in Myanmar – although it did later hire two Burmese speaking staff members. Under pressure from the FFM and human rights activists, Facebook eventually removed 18 personal accounts, one Instagram account, and 52 Facebook pages that it deemed were inciting hatred, disinformation and violence against the Rohingya. These pages were followed by tens of thousands of users. However, these pages were only removed in 2018, more than a year after the genocide took place – by which time more than 300 Rohingya villages had been burned down, tens of thousands of Rohingya were killed, and more than 750,000 people had been forced to flee across the border to Bangladesh. It was too little too late, and the FFM righty found that Facebook's overall response had been "slow and ineffective."

Responses from tech companies and social media platforms to online hate speech has been uneven and ad hoc. But monitoring social media and protecting against incitement is almost impossible to do on an ad hoc basis. The very nature of social media, with numerous platforms and groups and online spaces, makes the task more complicated than ever.

Regrettably, many social media companies don't invest resources in filtering or blocking hate speech. Instead, social media platforms rely on a combination of artificial intelligence, user reporting, and "content moderators" to monitor whether content is appropriate or not. Content moderators often lack knowledge of the specific political and historical context of situations where hate speech is prevalent. In addition, those tasked with monitoring harmful content usually don't have specific training or expertise in human rights.

Exacerbating these constraints is the fact that content on social media is usually posted in local languages. But most major tech companies are based in the United States and English-speaking content moderators often don't understand the danger and harmfulness of certain content, since it is in another language and/or is culturally-specific.

But some things are crudely obvious and don't require specialist training. For example, when the genocide against the Rohingya was underway in late 2017 I would often get trolled by people from Myanmar who would send me grotesque cartoons of stereotypical Muslim characters gang-raping Buddhist women. There would often be an adjoining message saying something like, "how would you feel if this was your sister?", followed by some comment about how the Army was protecting people from "Bengalis" (ie: the Rohingya) who were all terrorists, rapists and land thieves. Although I always reported these individual users, the images proliferated and were widely available across Twitter, Facebook and elsewhere.

So what steps are needed to prevent hate speech and incitement on social media and in turn prevent atrocities? Firstly, technology and social media companies need to employ experts to tackle the problem on their platforms. Algorithms and the complexity of social media cannot be an excuse for inaction. A first step would be for social media companies to establish designated staff for monitoring human rights and hate speech.

Governments and multilateral bodies must also assume their responsibility to address warning signs and prevent hate speech and incitement online. Technology is just a tool. The Nazis used newspapers, the Hutu genocidaire used radio, and the Army in Myanmar used Facebook. Technology is neutral, but those who use it are not. And not all content is equal. Governments should increase pressure on Facebook, Twitter, and other platforms to actively stop hate speech from being propagated on networks that operate in their country.

Numerous governments have also adopted different approaches to confronting and criminalizing hate speech and online incitement. Importantly, hate speech legislation should never be used as a guise for arbitrarily restricting freedom of expression. People have a right to be offensive or disagreeable. But they should not be allowed to utilize social media to incite violence, to foster persecution, or to promote genocide.

Promisingly, since May 2016 the European Commission has implemented a "code of conduct on countering illegal hate speech online" together with Facebook, Instagram, Twitter and YouTube. Since the code was adopted, companies have provided a swift response to racist and xenophobic hate speech when notified. According to reports, 89% of content flagged to these social networks now gets analyzed within 24 hours, and companies have removed 72% of content reported as illegal hate speech if requested to do so by a respected authority.

Overall, addressing hate speech and incitement on social media requires a coordinated response. Lessons from history are worth reiterating, not just because anti-semitic tropes and the far-right are making a disturbing online comeback in Europe, but because hate speech has been a precursor for mass atrocities in almost every major example from history, including in Rwanda, in Bosnia, in South Sudan and Myanmar.

Hate speech has always been about forging identities amongst both the targeted and the targeters. Identities are redefined and weaponized. But this means that they can also be contested and disrupted. We need better legislation to restrict hate speech, and we need more responsible and responsive social media companies. But all of us who use social media can also help deny online space to those who want to denigrate and divide. We need to strengthen national, regional and global identities that promote diversity and are only intolerant of intolerance. Because the ultimate antidote to online hate speech is speech that promotes and protects everyone's human rights.

Thank you.