Note to Office of the UN High Commissioner for Human Rights

By Sanjana Hattotuwa, Special Advisor, ICT4Peace Foundation
19 February 2016

Human Rights and Preventing and Countering Violent Extremism

The ICT4Peace Foundation thanks the Office of the High Commissioner for Human Rights for the invitation to write to it regarding best practices and lessons learned on how protecting and promoting human rights contribute to preventing and countering violent extremism. Our submission, in line with the Foundation's mission, will be anchored to technologies that can be adapted and adopted in countering violent extremism (CVE). Our submission will also highlight the roles and responsibilities of governments and civil society in this regard.

As noted in ICTs for the prevention of mass atrocity crimes[1] published by the Foundation in 2010:

> With growing access to new technologies and channels of communication, such as new media and mobile phones, an increasing number of hitherto marginalized, compelling accounts of violence are being recorded for posterity. These accounts can contribute to increased awareness on genocide and crimes against humanity. Crawford & Cole (2007) argue that ICTs can be used to build lasting peace through: providing information, helping people access information, improving decision making, reducing scarcity, supporting relationships and helping people understand each other.[2] ICTs can aid these tactics in many ways – high quality citizen journalism and low-cost technologies have helped in processes of transitional justice, accountability, truth seeking and reconciliation alongside other initiatives, including those by government.
>
> Civil society is becoming increasingly involved in the search and design of digital innovations for addressing the challenges of genocide. A recent example is Project

[1] http://ict4peace.org/icts-for-the-prevention-of-mass-atrocity-crimes/
[2] Crawford & Cole, 2007

> 10^100, a competition hosted by Google, where the idea of creating a genocide monitoring and alert system was one of the sixteen finalists. The ideas included reducing crimes against humanity by aggregating data, including pertinent statistics, the history and geography of specific conflicts, local cultures, geostrategic interests, by using e.g. updated dynamic web maps and hand-held GPS devices.[3] Another example is found in a recent report by Amnesty International entitled 'Geospatial Technologies', where technologies such as satellite images, GPS, virtual globes and infrared/multispectral sensing are assigned the purpose of assisting, monitoring and advocating the protection of populations at risk and advanced warning of crises.[4] Done well and over the long-term, initiatives like these can prevent recurrence of genocide and mass atrocity crimes.

As noted by the Foundation's Special Advisor, Sanjana Hattotuwa, in 2014, in a pivotal study of hate speech online in Sri Lanka (*Liking violence: A study of hate speech on Facebook in Sri Lanka*[5]):

> "Hate speech" on the Internet is a global concern and with no kill-switch solution. Depending on the location online, language and media used, context and sometimes even the nature of the actors concerned, dealing with hate speech is a vexed challenge from parent to policymaker. This hasn't stopped politicians, with little to no understanding of underlying technical challenges or repressive governments, who often seek a monopoly around the dissemination of defamatory propaganda seeking to control hate speech. Parochialism and expediency drive most efforts around hate speech related policy responses and legislation. In Sri Lanka, online social media and web based platforms, accessed increasingly over smartphones and tablets, provide an important, necessary vent for critical dissent, in a context where mainstream media does not and cannot afford the space for questioning or content that holds the government accountable for heinous crimes and outrageous corruption. The growth of content creation and consumption online, wider and deeper than any other media in the country and at an accelerated pace, has also resulted in low risk, low cost and high impact online spaces to spread hate, harm and hurt against specific communities, individuals or ideas. Conspiracy theorists, fringe lunatics and trolls have since the first days of the Internet inhabited online spaces and engaged with devoted followers, or sought to deny and decry those who question them. The growth of hate speech can be seen as a natural progression outward from these pockets of relative isolation, and is also pegged to the economics of broadband internet access and the double digit growth of smartphones – an underlying, coast to coast network infrastructure capable of rich media content production and interactive, real time engagement. This infrastructure has erased traditional geographies – hate and harm against a particular religion, identity group or community in one part of the world or country, can for example within seconds, translate into violent emulation or strident opposition in another part, communicated via online social media and mediated through platforms like Twitter, Facebook and also through instant

[3] Google's Project 10^100, Finalists
[4] Amnesty International, Geo-Spatial Toolkit
[5] http://www.cpalanka.org/liking-violence-a-study-of-hate-speech-on-facebook-in-sri-lanka/

> messaging apps for mobiles like iMessage and WhatsApp, in addition to the older SMS technology.
>
> A central challenge around addressing hate speech is that it is technically impossible – given the volume, variety and velocity of content production on the Internet today[6] – to robustly assess and curtail, in as close to real time as possible, inflammatory, dangerous or hateful content just in English, leave aside other languages like Sinhala or Tamil. Once content is produced for the web and originally for a single platform, given user interactions and responses, it often replicates and mutates into other content over dozens of other websites and platforms, making it impossible to complete erase a record of its existence even if the original was taken down, deleted or redacted. This makes it extremely hard to address the harm arising out of hate speech, since there is so much of it around in digital form over so many media.

The same report goes on to note:

> This brings us to a key challenge around hate speech – it always requires context to understand and address, and increasingly, the intermediaries in both supporting and curtailing the spread of it are corporate entities, not governments. Machine level and algorithmic frameworks to identify and block hateful and harmful content often fail, simply because they flag too many false positives (content erroneously flagged as hate speech) or allow so much of hate speech to pass through (in, as noted earlier, languages other than English) that their core purpose rendered irrelevant. This puts the burden of addressing this content on users themselves, who through reporting mechanisms baked into all the major only social media platforms, can choose to report hate speech with relevant context. Only as effective as the numbers who report hate speech, these reporting mechanisms also take some time to kick-in from the time of submission to the actual deletion or blocking of the original content, page, account or user. At a time of heightened violence, this time lag is unhelpful. There is also no guarantee the (corporate) owner of an app, service, platform or website agrees with the reporting of hateful content. Studies show, for example, significant variance in dealing with hate speech even within Facebook[7].

Precisely the same arguments and observations on hate speech in various web and mobile fora can be made of CVE online. Combatting CVE from a rights-centric framework requires a concert of measures by government, civil society and transnational institutions like the UN.

Some recommendations for the consideration of OHCR around CVE follow.

1. Counter-messaging is the production of content around CVE, and disseminated using the same platforms, apps, services and sites as more harmful content. This can include, on a case by case basis, direct engagement with accounts that promote

[6] http://www.ibmbigdatahub.com/infographic/four-vs-big-data
[7] http://ohpi.org.au/if-you-cant-recognize-hate-speech-the-sunlight-cant-penetrate/

violent extremism by debunking misinformation and disinformation campaigns, and calling their bluff on pseudo-science and myth-making.

2. As noted by Hattotuwa in 2013[8], is that "groups which attempt to portray a more inclusive and tolerant country, by critiquing the positions of the extremists, often come under attack, are subject to hate speech and fail to attract as many followers as the Facebook pages and groups with inflammatory content". This is turn calls for sufficient human, financial and institutional resources to support more sustained monitoring of these hate speech trends, so as to create early warning mechanisms that alert relevant authorities and civil society stakeholders around heightened tensions online that could explore into, or exacerbate, real world violence. Importantly, this monitoring should cover vernacular languages in addition to English content across key social media domains, apps, services and platforms.

3. Along with more sustained, deeper monitoring of violent extremism online, and digital media literacy campaigns around the critical appreciation of such content geared towards those between 18 – 30 in particular, recommendations to address hate speech online echo points made in the Bytes for All report on Pakistan's hate speech in cyberspace[9], pegged to "a multi-pronged approach to work, a plan of action that has multiple stakeholders involved would be necessary to maintain checks and balances, particularly to ensure that the issue of hate speech in cyberspace is not manipulated and used to further political agendas, increase censorship and/or target and discriminate against vulnerable individuals/groups". Bytes for All focuses on the role of government, mainstream media, online companies (e.g. Facebook) and organised civil society advocacy and activism as means through which online hate speech can be, to the extent possible, effectively contained and addressed.

4. However, especially in contexts where there is a democratic deficit, the challenge of engaging with government and informing progressive policymaking is even more acute if not downright impossible. Mainstream media in general is extremely risk averse and has neither the imagination nor independence to counter hate speech by extremist groups, especially when widely perceived to be protected by powerful sections of the government. Furthermore, little to no comment moderation guidelines across mainstream media website also result in trolls openly publishing comments full of hate, hurt and harm. In this light, comment moderation and content curation policies in line with what the World Editors Forum has published in 2013[10] can greatly contribute to the creation of official mainstream media websites and social media accounts that actively resist and combat hate speech and engage in CVE, complementing editorial policies that also, on principle, disallow defamatory and inflammatory content from the institution's articles, columns and broadcasts.

[8] https://sanjanah.wordpress.com/2013/02/01/anti-muslim-hate-online-in-post-war-sri-lanka/
[9] https://content.bytesforall.pk/sites/default/files/Pakistan_Hate_Speech_Report_2014.pdf
[10] http://www.wan-ifra.org/reports/2013/10/04/online-comment-moderation-emerging-best-practices

5. There are also other possibilities, arising from Dr. Tarlach McGonagle's work on addressing online hate speech in Europe[11]. Key amongst her ideas and fully worth embracing is to develop and effectively promote an 'Anti-Hate Speech Pledge' for politicians and political parties. As noted by Dr. McGonagle,

> … a certain minimum number of commitments [around combatting hate speech] would have to be entered into, in return for which, a party could display the logo for the Pledge on all of its official materials. In order to ensure seriousness of purpose and meaningful uptake, participating party leaders would be obliged to attend annual meetings to explain and evaluate their parties' actions to combat hate speech. A non-roll-back clause could be included in order to ensure that annual achievements would continuously be built on.

   It is clearly in the interests of all political parties to explore ways through which their party leadership, officer bearers and supporters can counter hate speech in general and online hate speech in particular, by signing up to a pledge – with public and visible punitive measures taken against anyone who goes on to produce or disseminate hate and harm. A political culture of zero tolerance over hate speech can deeply influence the production and appraisal of hate speech online.

6. Demographics are important: Youth (those between 18-24 in particular) stand the risk of radicalisation upon entering and engaging with online and mobile chat based fora. To appeal to this segment, iconic figures from youth (singers, actors, sportspersons, YouTube producers, hackers, IT industry leaders, young entrepreneurs) are more important to leverage in counter-speech initiatives than say expressions from or iconography based around the dhamma. It is also the case that combined with geo-targetting, those who are held in high regard by this age group in local communities (ranging from monks in a community temple where this segment has gone for tuition or Sunday school to local business owners) can be leveraged, the emphasis being on the identification of influencers within that demographic, and furthermore, by geography.

7. Geo-targeting/geo-fencing: Easily done on Facebook, counter-speech content (ranging from pages to specific posts on Facebook) can be targetted to specific regions, at specific times, for specific communities. Wide-scatter promotions simply don't work, either displaying on the screens of those who are already partial to the counter-speech content, or only sporadically appearing on the screens of those for whom it is most relevant. The larger the terrain of an audience, the greater the emphasis should be on geo-fencing counter-speech content. For example, during an election, constituencies that have witnessed heightened communal or partisan violence can be targetted well before the day of the election with counter-speech messaging to prevent the spread of rumours and other inflammatory content.

---

11 https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016800c170f

8. Language: In a multi-lingual country, CVE is largely ineffective if it isn't conducted in the language that dangerous and hate speech fora use in their interactions. Hate and dangerous speech on Facebook often exist only in the vernacular, and counter-speech initiatives in English alone have no relevance or traction. Iconic counter-speech examples like Panzagar in Myanmar can be very effective, since they transcend the barriers of language. Short-form video can also be a powerful vector for counter-speech to reach target audiences, without necessarily being anchored to a single language.

9. Translation: Good translations of CVE content that communicate ideas and meaning are hard to come by, and good translators are generally over-worked. Idioms, nuances, aphorisms and adages in languages differ, and native speakers of the language counter-speech content was originally produced in or for, and the language into which it will be translated into are very hard to come by.

10. Time: CVE is a long-term process, and timing is important in so far as what is expected as a result. Counter-speech to address and reduce electoral violence requires a different timeline to content that seeks to address deep-rooted communal or religious tension. Project oriented counter-speech campaigns, which are often driven by relatively short-term funding opportunities, are often too short for any meaningful impact.

11. Reasons for (social media) engagement: CVE proponents need to do far more, and better research around why, and at what times, hate and dangerous speech content is produced and received with high levels of engagement. What drives the production cycles? Are there links to key political or cultural events? Is there a connection between the utterances of key individuals and the production of hate speech in online fora? Is there a connection between the speeches of political groups, politicians, religious leaders or other individuals and the engagement online using dangerous speech? Does hate speech increase in the lead up to an election, and if so, at what key points?

12. Law of diminishing returns: CVE proponents need to create content that doesn't just go viral once. They should also keep in mind that content addressed to the same demographic will, unless very inventive, generate progressively less interest and interaction over time. The higher the frequency of content production sometimes risks the perception of counter-speech as spam, whereas too infrequent production also risks ineffective audience engagement. Context is critical to content.

In addition to these recommendation, the following broad observations may hold some relevance to OHCHR's process.

- Study the generation and spread of hate and dangerous speech by spoilers and other groups who are the lead architects of discord
- Demographics – carefully target those who haven't yet been radicalized by their entry and participation in known FB groups that incite hate

- Geo targeting – Locate and address cities, provinces and locations that have historically had a prevalence to (violent) act on content that is digitally produced and disseminated
- Language – Use effective means through which to craft and communicate counter-speech
- Make sure the counter-speech is localized and appeals to the target audience(s) in terms of optics
- On Facebook, counter-speech pages, groups and accounts must focus on engagement more than likes
- Constantly examine reasons for engagement and try to strengthen known drivers to enhance reach
- Encourage leading social media companies like Facebook to invest more in the algorithmic or machine examination of content posted.

****