Inputs on Online Hate Speech to the

Office of the UN High Commissioner for Human Rights

IT for Change

June 2023

IT for Change

IT for Change's Inputs to OHCHR on Online Hate Speech

IT for Change[1]

June 2023

Introduction

'Hate speech', which can be described as offensive speech targeting a group or an individual based on inherent characteristics (race, ethnicity, gender, religion, nationality, sexual orientation, etc.), is a form of attack rooted in social prejudice. It reinforces widespread discrimination, ostracization, and delegitimization of members of such groups, thereby preventing their equal participation in public life. It violates the dignity of the members of the targeted group, imperils their right to life, restricts their freedom of expression, and inhibits their ability to lead autonomous and fulfilling personal lives.[2]

The internet has allowed for the dissemination of hate speech and hateful messages at an unprecedented scale, reach, and speed. Online hate speech takes various forms – outright hateful and vitriolic messages attacking an individual or group, covert forms of hate and abuse often laced in humor and wordplay (1, 2), and spreading deceptive or inaccurate information against certain communities and groups. The pervasiveness of online hate speech has upended the transformative potential of the internet for the oppressed and marginalized communities to meaningfully participate in political, economic, social, and cultural life. Big Tech platforms with network advantage mediate today's online communication space, deploying proprietary algorithms that amplify hateful and divisive content to profit from the ever-widening engagement with such content.

In India, despite constitutional guarantees of equality and non-discrimination on grounds of religion, race, caste, sex, descent, and place of birth,[3] there is widespread discrimination and prejudice against individuals and groups based on these identity characteristics. Often such discrimination takes the form of hate speech, both offline and online. For example, Northeastern Indians, targeted for their East-Asian looks, often face racist slurs and are referred to by pejorative terms (1,2,3,4); Muslims in India are frequently subjected to hateful and inflammatory speech and disinformation in a climate of Hindu majoritarianism in the country, often leading to offline violence (1,2); and individuals are ridiculed and discriminated against on the basis of their skin tone (1,2). Another form of discrimination which is deeply entrenched and pervasive in India is casteism. A recent study on Online Caste Hate Speech in India highlighted that casteism manifests in online spaces, "with social media platforms becoming sites of caste discrimination and humiliation." As a result of increasing international focus on caste-based discrimination and hate speech in India, General Recommendation 29 of the Committee

[1] Inputs by Anita Gurumurthy, Malavika Rajkumar, and Merrin Muhammed Ashraf
[2] Herz, M., & Molnár, P. (Eds.). (2012). The content and context of hate speech: Rethinking regulation and responses. Cambridge University Press.
[3] The Constitution of India, 1950. Articles 14, 15, & 16.

on Elimination of Racial Discrimination in 2002 stated that "discrimination against members of communities based on forms of social stratification such as caste and analogous systems of inherited status" is a form of racial discrimination.

In the Indian context, it is also pertinent to note the rising instances of gender-based hate speech. While this call for inputs by the Office of the United Nations High Commissioner for Human Rights (OHCHR) aims to feed into the report on forms of racism on online hate speech, through our submission we would also like to highlight how gender intersects with, compounds, and reinforces discrimination based on race as well as other social markers such as caste and religion to make women and gender minorities particularly vulnerable targets of hate speech and harassment online.

Points of Concern and Recommendations

1. Importance of recognizing gender-based hate speech and intersection of gender with other marginalized identities

Hate speech on the basis of gender is a form of gender-based violence that women and gender minorities face frequently, both online and offline. Gender-based hate speech includes expressions that spread, incite, promote or justify hatred based on one's gender identity. It takes various forms like reducing women to their basic biological and reproductive functions, attacking one's choice of gender identity, promoting hatred or discrimination against certain gender groups, attacking a person's credibility on the basis of their gender, and reinforcing harmful gender stereotypes (1,2). Gender-based hate speech perpetuates and exacerbates gender inequality and adversely impacts equal participation of women, including LBT women[4], and other gender minorities. The prevalence of misogynistic trolling and online hate, when coupled with the networked dynamics of platform sociality, routinizes censure and abuse against "erring" or "transgressing" women, resulting in a gendered restructuring of the digital space.

However, it is pertinent to note that the experience of gender-based hate speech is not the same for all women and gender minorities. Gender identity intersects with other identity markers, such as race, caste, religion, ethnicity, etc., mutually reinforcing and constructing marginalities. As such, women and gender minorities falling at the intersections of multiple marginalized locations are at heightened risk of being targeted online. For example, women in Northeastern parts of India are subject to several stereotypes, based on both their race and gender, which inhibit their full and free participation in public life. With respect to intersection of gender with religion and caste, IT for Change's study on misogynist trolling and abuse against Indian women in public-political life indicated that Muslim women and Dalit women received an overwhelming majority of the abusive messages that were

[4] Lesbian, Bisexual, and Trans women

mapped for the study. The rise of the 'Sulli deals' and 'Bulli Bai' applications hosted on Github that had listed hundreds of Muslim women for "auction" with their photographs doctored and sourced without their permission is yet another example of magnified hate against Muslim women. Women and queer people from Dalit community face significant casteist abuses and slurs, including rape threats and gendered and queerphobic hate speech.

These examples show that discrimination based on race, caste, religion, etc. has a gender dimension, which is also reflected in forms of online hate speech against members of marginalized groups and communities.

Recommendations

- There should be international recognition at the UN level of the term gender-based hate speech and its impact on the human rights of women and gender minorities, both offline and online. A first step in this regard should be an explicit recognition by the Special Rapporteur on Freedom of Opinion and Expression on the responsibility of the State in preventing and redressing such hate speech. Further, the report should highlight the failure of digital platforms to combat gender-based hate speech on their platforms and their role in creating a conducive environment for the proliferation of hate speech, as done by the 2019 report of the UN Special Rapporteur on Online Hate Speech. This should lead to a call to institute an effective accountability framework to hold platforms responsible for the deliberate and systematic failure to take steps to prevent and mitigate harm from gender-based hate speech.
- Apart from a specific recognition of the issue of gender-based hate speech, there is also a need to address the gender dimension of discrimination and hate speech on the grounds of race, religion, ethnicity, etc. While the 2019 report of the UN Special Rapporteur on Online Hate Speech recognized gender as one of the grounds on which hate speech is perpetrated, it stopped short of calling out how gender intersects with other identity markers in amplifying the harm from hate speech. In this regard, a leaf can be taken out of the Durban Declaration on Racism, Racial Discrimination, Xenophobia, and Related Intolerance, which recognizes the need to integrate gender perspective into relevant policies, strategies, and programs of action against racism, racial discrimination, xenophobia, and related intolerance, in order to address multiple forms of discrimination.

2. Need for recognition and strong frameworks of platform accountability in addressing online hate speech, including gender-based hate speech

While technology mirrors society and its patriarchal culture, society is also co-constituted by the technological paradigm. Evidence of growing occurrences of gender-based hate speech has been continuously discussed, presented, and debated over the years. So much so that sexism, misogyny,

and gender-based violence on digital platforms are now trivialized and dismissed as a normal part of the online experience.

Online hate speech often involves many speakers acting in concert, like mob attacks, with coordinated bot threats, disinformation, and so-called deep fakes. The harm from online hate speech is exacerbated by the scale, virality, and velocity of content posted on social media platforms. Large platform companies do not just facilitate online violence but are also complicit in amplifying harmful, hateful, and violent content.[5] Algorithms for content curation tend to prioritize sensationalist content over verified information, reflecting the profit motive that drives the mainstream technological paradigm. The artificial intelligence (AI) systems of dominant digital platforms have also been reported to have failed to detect abusive posts that promote hate speech in a number of languages used on these platforms. Further, grievance redressal mechanisms of digital platforms are slow and inadequate to respond to online abuse and hence ineffective.

Despite this active complicity of digital platforms in perpetrating online hate speech, the architecture of corporate impunity, primarily in the form of laws revolving around the dumb-conduit argument, has allowed social media corporations to avoid accountability for the human rights violations resulting from their operations. This also derails attempts to construct a critical discourse for regulating their conduct.

Recommendations

We recommend explicit recognition of the role and responsibility of digital platforms in preventing online hate speech, particularly gender-based hate speech. In this regard, guidance can be taken from the global norms given by the UN Guiding Principles on Business and Human Rights to address the systemic issue of corporate impunity and pin down accountability of digital platforms for perpetuating hate speech online. In pursuance of this, we recommend certain key measures:

- The Special Rapporteur's report should highlight the role of states in instituting an accountability framework to hold platforms liable for their actions that threaten user rights or facilitate various forms of hate speech, including gender-based hate speech.
- Platforms should be required by law to conduct periodic human rights due diligence of their policies and operations and take techno-design measures to stem the virality of sexist, misogynist, and other illegal and harmful content. In addition to internal expertise, they should also draw on independent human rights expertise through "meaningful consultation with potentially affected groups and other relevant stakeholders", in order to regularly evaluate the effectiveness of their approaches to human rights harms.

[5] Also see, Gurumurthy, A., & Dasarathy, A. (2022, July). Profitable Provocations. A Study of Abuse and Misogynistic Trolling on Twitter Directed at Indian Women in Public-political Life. IT for Change. https://itforchange.net/sites/default/files/2132/ITfC-Twitter-Report-Profitable-Provocations.pdf

- Further, platforms should take measures such as removing, downranking, and labeling hate speech, as the case may be, to prevent and mitigate harm from such content, balancing considerations of freedom of expression and information. Factors or criteria used by platforms to determine whether a content is potentially harmful and the action that will be taken with respect to it must be made transparent and public.

3. Legislative, policy and regulatory efforts needed to address hate speech

Currently, there is a gap in both knowledge and policy action at the global level to combat hate speech, particularly gender-based hate speech. The first aspect that must be recognized within any legislative, policy or regulatory tool is the definition of hate speech and the aspects covered under it. In the European Union for instance, 'hate speech' and 'hate crime' do not have universal definitions and thus are interpreted differently by the member states, with different areas of law providing protection based on the type of crime. For effective adjudication, enforcement, and legislative practice, there must be clarity and comprehensiveness in law while defining hate speech.

In terms of a legal liability mechanism, criminal liability for hate speech has mostly taken the form of defamation laws. However, criminal liability is not a foolproof solution, and it often creates problems in maintaining a balance between fighting hate speech on the one hand, and safeguarding freedom of speech on the other. If criminal laws to deal with hate speech are not drafted carefully, it could give way to overbroad interpretations and be misused to silence minorities and suppress criticism of official policies and political opposition.

Recommendations

- Only with a common definition of hate speech that is attentive to social markers of identity can legislative and regulatory tools be effective in addressing the issue. One of the most widely accepted attempts to define hate speech was made by the Council of Europe (CoE), whose Recommendation No. R (97) 20 defines hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance". A step forward has been made by the CoE's Gender Equality Strategy 2014-2017, which explicitly includes tackling sexism as a form of hate speech under its strategic objective. The report made by the Special Rapporteur must inform and clarify such a definition for international consensus, reflecting the approach taken by CoE. This definition of hate speech should become a baseline for states to enact, enforce or amend legislation or regulatory efforts to combat hate speech.
- We also recommend clear directions for national level responses to address hate speech, including greater focus on civil law-based remedies which hold digital platforms and other entities liable for promoting hate speech. Further, recognizing the role of digital platforms in

facilitating sexist hate speech, the CoE also recommends instituting measures for effective moderation of social media, including by setting clear standards for the industry and putting in place mechanisms to monitor progress. While instituting these measures to address corporate impunity, it is also important to ensure that the liability provisions do not cause these platforms to engage in overzealous censorship of content, thereby endangering the freedom of expression of users. It is useful here to refer to the recommendations of the CoE which proposes taking a broader approach to the issue that goes beyond criminalization against users to tackle hate speech. This includes taking measures such as eliminating discriminatory laws, promoting gender equality and media literacy training, having in place clear policy frameworks, and legal remedies for women and girls who are subjected to sexist and harmful content, etc.