AMNESTY INTERNATIONAL

TIGO IOR/40/2017.066

20 December 2017

CONTENT REGULATION IN THE DIGITAL AGE

AMNESTY INTERNATIONAL SUBMISSION TO UNITED NATIONS SPECIAL RAPPORTEUR ON THE PROMOTION AND PROTECTION OF THE RIGHT TO FREEDOM OF OPINION AND EXPRESSION

Amnesty International is currently investigating the human rights implications of violence and abuse against women on social media platforms – in particular Twitter and Facebook. This submission sets out some of the findings in relation to individuals at risk, automation and content moderation and transparency in response to the call for contributions ahead of the June 2018 thematic report on *Content regulation in the digital age*.[1]

4. Individuals at risk: Do company standards adequately reflect the interests of users who face particular risks on the basis of religious, racial, ethnic, national, gender, sexual orientation or other forms discrimination?

Social media platforms are a critical space for individuals to exercise the right to freedom of expression. Online platforms have created an environment where individuals can debate, discuss, engage or collectively organize with other individuals and such spaces have helped create visibility and awareness around various gender and identity-based issues affecting women and other marginalized groups in society.

However, violence and abuse against women online is part of the same continuum of gender-based violence and abuse against women that affects women offline. Our research has shown that violence and abuse against women online negatively impacts women's right to freedom of expression, as well as their enjoyment of other human rights, such as the right to privacy.[2] Women often experience violence and abuse online that targets the various intersecting aspects of their identities, including those related to their gender, race, ethnicity, religion, sexual orientation, gender identity, gender expression, age or disability, which can compound their experience of violence and abuse online. Individuals who are gender non-conforming or non-binary can also be targeted with violence and abuse online for transgressing existing gender

[1] http://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulationInTheDigitalAge.aspx

[2] https://www.amnesty.org/en/latest/campaigns/2017/11/what-is-online-violence-and-abuse-against-women/

norms.

Community standards, in principle, recognize that users from marginalized communities face a particular risk for targeted abuse because of their identities.[3] However, transparency around how such standards are interpreted to ensure that these individuals are able to exercise their right to freedom of expression online equally, freely and without fear remains weak. Companies provide little information on how or whether these groups are meaningfully and regularly consulted with to address the specific risks of violence and abuse they may experience. Social media platforms should provide more information about which members of civil society they consult with on the development of policies and community standards and what the consultation process with such groups consists of. It is important that social media companies do not merely use such discussions as a tick-box exercise but instead take the specific concerns and risks users can face into consideration into all aspects of policy and product development for the platform.

8. Automation and content moderation: What role does automation or algorithmic filtering play in regulating content? How should technology as well as human and other resources be employed to standardize content regulation on platforms?

Automation and the use of machine learning to detect violence and abuse on social media platforms can be helpful in assessing trends and patterns of such content on these platforms, such as information about which users are being targeted with violence and abuse or how online violence and abuse online correlates with offline events.

In September 2017, Amnesty International used machine learning and developed an algorithm to detect online abuse against female parliamentarians in the UK. The algorithmic analysis helped demonstrate the intersectional nature of online abuse with the findings showing that Shadow Home Secretary and the UK's first black female MP Diane Abbott receiving 45.15% of all abuse against female MPs active on Twitter in the 6 weeks leading up to the Snap Election in June 2017 and 10X more abuse during this time than any other female MP included in the study. The findings also showed that Asian female MPs, despite only making up 8.8% of female MPs in the UK, received 35% more abuse than white female MPs.[4]

Tracking online violence and abuse is important for developing solutions to solve this problem – the more data that is readily available about how violence and abuse against women manifests on social media platforms, the more this can help companies, governments and civil society organizations to think of solutions. However, in the case of violence and abuse against women online, the context of such content is incredibly important. Automated systems that are used as the sole mechanism to take down content poses a serious risk to restricting legitimate expression online. Companies must be transparent about how they are using machine learning to curb online violence and abuse against women and how they balance the use of algorithmic systems to detect abuse with the right of users to express themselves freely online. There is also a risk that automated systems used to detect violence and abuse online will entrench existing discrimination and companies should take steps to mitigate against any bias in the design of machine learning systems.

9. Transparency: What information should companies disclose about how content regulation standards under their terms of service are interpreted and enforced? Is the transparency reporting they currently conduct sufficient?

---

[3] See: https://www.facebook.com/communitystandards#hate-speech and https://help.twitter.com/en/rules-and-policies/twitter-rules

[4] https://medium.com/@AmnestyInsights/unsocial-media-tracking-twitter-abuse-against-women-mps-fc28aeca498a

Tackling online violence and abuse on social media platforms and protecting women's rights online requires resources, transparency and coordinated action from social media companies and governments. Social media companies have a responsibility to respect human rights and this includes the right to freedom of expression. This means ensuring that the women who use their platforms are able to do so equally, freely and without fear. In some circumstances, online abuse - despite being offensive or disturbing - should not be censored or restricted in the form of take down measures. Instead, social media companies must enable and empower users to understand and utilize individual security and privacy measures such as blocking, muting and content filtering so women are easily able to curate a less toxic and harmful online experience.

The policies of social media platforms explicitly state that they do not tolerate targeted abuse on the basis of a person's gender or other forms of identity. In instances where online abuse or violence against women is in breach of the company's own 'hateful conduct' or 'harmful abuse' policies, social media companies should enforce their own policies and implement adequate and transparent reporting mechanisms that users have confidence in utilizing. Social media companies must also ensure that moderators are trained in identifying gender and other identity-related threats and abuse on their platforms, as well as in international human rights standards regarding freedom of expression and privacy.[5]

Social media platforms should also record and publicly share disaggregated data about the levels and types of abuse reported, as well as their response on a regular basis.[6] They should also record and share whether users are satisfied with the outcome of any reports made to the platform. Platforms should be transparent about the resources they are specifically investing into tackling online violence and abuse such as how many moderators they employ to tackle violence and abuse online - disaggregated by language and region.

In November 2017, Amnesty International commissioned IPSOS Mori to conduct an online poll with women aged 18-55 across 8 countries (UK, USA, New Zealand, Italy, Spain, Sweden, Poland and Denmark). Our findings show that many women feel the response of social media companies to tackle online abuse has been inadequate. Almost 1/3 (32%) of the women polled who use Facebook stated that the company's response to dealing with abuse or harassment online was inadequate. Twitter did not fare much better. Almost 30% of women polled who are Twitter users stated the company's response to abuse or harassment was inadequate, including 43% of women users in the UK and 41% in Sweden.[7] These figures demonstrate that social media platforms have much more work to do when it comes to increasing the confidence and trust of their women users.

[5] https://medium.com/@AmnestyInsights/unsocial-media-tracking-twitter-abuse-against-women-mps-fc28aeca498a

[6] http://www.refinery29.com/2017/12/185584/twitter-abuse-rules-amnesty-international-opinion

[7] https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4