## COMMENTS ON CRIMINAL ALLEGATIONS BASED ON NATIONALITY

Social media platforms, including Meta, promote public discourse and activism, particularly during political upheaval and conflict.[1] Platforms's hate speech policies should carefully balance preventing hate speech while allowing legitimate criticism of state actions and policies. The ability to criticize state actions is a fundamental aspect of free expression and is vital for accountability and transparency. Blanket bans on generalizations about nationality are overly broad and risk suppressing legitimate political discourse. For instance, censoring content that criticizes government actions can discourage free speech, as individuals might self-censor to avoid having their posts deleted or accounts suspended. This issue is especially troubling during conflicts, where social media platforms could otherwise play a vital role in documenting human rights violations and facilitating public discussion.[2]

The impact of content alleging criminality based on a person's nationality can contribute to the stigmatization, dehumanization, and discrimination of these groups, exacerbating their vulnerability and marginalization in already volatile environments.[3] In the context of crisis and conflict situations, where social tensions and polarization heighten, this can make it easier for perpetrators of violence and discrimination to justify their actions and can contribute to the normalization of hate and intolerance in public discourse. For example, during the COVID-19 pandemic, there was a significant increase in hate speech and discrimination directed toward Asian and Asian-American communities, fueled in part by the spread of misinformation and conspiracy theories on social media platforms.[4]

[1] Zeynep Tufekci and Christopher Wilson, 'Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square' (2012) 62(2) *Journal of Communication* 363 <http://onlinelibrary.wiley.com.ezp01.library.qut.edu.au/doi/10.1111/j.1460-2466.2012.01629.x/abstract> ('Social Media and the Decision to Participate in Political Protest').

[2] Jonathon W Penney, 'Internet Surveillance, Regulation, and Chilling Effects Online: A Comparative Case Study' (2017) 6(2) *Internet Policy Review* <https://policyreview.info/articles/analysis/internet-surveillance-regulation-and-chilling-effects-online-comparative-case> ('Internet Surveillance, Regulation, and Chilling Effects Online').

[3] Victoria M Esses, Stelian Medianu and Andrea S Lawson, 'Uncertainty, Threat, and the Role of the Media in Promoting the Dehumanization of Immigrants and Refugees' (2013) 69(3) *Journal of Social Issues* 518.

[4] Bing He et al, 'Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis' (No arXiv:2005.12423, arXiv, 10 November 2021) <http://arxiv.org/abs/2005.12423> ('Racism Is a Virus').

As a global platform with immense reach and influence, Meta is responsible for respecting and upholding human rights, including freedom of expression and non-discrimination.[5] Meta is responsible for ensuring that its policies and practices do not contribute to the spread of hate speech, discrimination, or incitement to violence against people based on their nationality or other protected characteristics.[6] Ambiguity in the policies can lead to inconsistent enforcement, where similar content is treated differently. To fully meet its human rights responsibilities, Meta should develop more detailed guidelines for assessing the impact of content on marginalized groups and invest in resources and training for content moderators to help them identify and address subtle forms of hate speech and discrimination.[7]

I propose the following criteria for establishing whether a user is targeting a concept/institution or a group of people based on their nationality:

1. **Intention of the user and past behavior:** Analyze whether the user has a history of sharing content that targets a specific concept or institution or if their stated intent is to promote hate or discrimination.[8]
2. **Specificity:** Analyze the language used in the content. The more specific the reference to a concept or institution, the more likely the user is targeting that concept or institution rather than a group of people.[9]
3. **Context:** Content shared in an ongoing conflict or historical tension may require a more nuanced interpretation.[10]
4. **Impact:** Assess the potential for the content to contribute to harm, discrimination, or violence against members of the targeted nationality or ethnic group.[11]

[5] Rikke Frank Jørgensen (ed), *Human Rights in the Age of Platforms* (The MIT Press, 2019).

[6] David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (Columbia Global Reports, 2019) (‘*Speech Police*’).

[7] ‘Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda - Nicolas Suzor, Tess Van Geelen, Sarah Myers West, 2018’ <https://journals.sagepub.com/doi/10.1177/1748048518757142>.

[8] Alexandra Olteanu, Kartik Talamadupula and Kush R Varshney, ‘The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection’ in *Proceedings of the 2017 ACM on Web Science Conference* (Association for Computing Machinery, 2017) 405 <https://dl.acm.org/doi/10.1145/3091478.3098871> (‘The Limits of Abstract Evaluation Metrics’).

[9] ‘Countering Online Hate Speech - UNESCO Digital Library’ <https://unesdoc.unesco.org/ark:/48223/pf0000233231>.

[10] Antoine Buyse, ‘Words of Violence: “Fear Speech,” or How Violent Conflict Escalation Relates to the Freedom of Expression’ (2014) 36 *Human Rights Quarterly* 779 (‘Words of Violence’).

[11] Karsten Müller and Carlo Schwarz, ‘Fanning the Flames of Hate: Social Media and Hate Crime’ (2020) 19 *Journal of the European Economic Association* (‘Fanning the Flames of Hate’).