UNITED NATIONS
HUMAN RIGHTS
SPECIAL PROCEDURES
SPECIAL RAPPORTEURS, INDEPENDENT EXPERTS & WORKING GROUPS

Public Comment by UN Special Rapporteur on Freedom of Opinion and Expression Irene Khan on Facebook Oversight Board Case no. 2021-009

Uneven Content Moderation in the Middle East

There have been reports of a troubling pattern of uneven application of Facebook's Terms of Service and policies, resulting in arbitrary treatment of content relating to the recent conflict between Israel and Palestine. In this context, the Facebook Oversight Board has accepted to consider the appeal of a Facebook content moderation decision related to an Al Jazeera news story that was reposted in May 2021 by an ordinary user without comment, and whose repost was deleted while the original news article remained on the platform. The reposted Al Jazeera story showed members of the Qassam Brigades, the military arm of Hamas, with the text "He Who Warns is Excused.' Al-Qassam Brigades military spokesman threatens the occupation forces if they do not withdraw from Al-Aqsa Mosque." Facebook removed the repost for violating its Dangerous Individuals and Organisations Community Standard while the Al Jazeera news article remained on the platform.

Around that time, Facebook and its social networking website Instagram reportedly blocked numerous posts, hashtags and livestreams related to protests against planned evictions of Palestinian families from their homes in the Sheikh Jarrah neighbourhood, as well as content restrictions related to the Al Aqsa Mosque in Jeruselum and the term 'Zionist'. Where reasons were given, users were told that their posts violated the Community Standard on dangerous individuals and organisations, amongst others. There are concerns that at least some content restrictions may have been linked to requests from the Israeli Ministry of Justice's Cyber Unit.

Facebook has previously faced scrutiny for its content moderation errors in relation to Palestine, which were often ascribed to a shortage of content moderators and the inability of algorithmic content moderation to properly moderate content in languages other than English. Although Facebook has staffed a regional operations centre with native speakers of Arabic and Hebrew, the suppression of content by local activists, journalists and human rights and defenders has continued.

Since then, it appears that Facebook and Instagram have restored some content and accounts, as well as made changes to the algorithm, but have not fully disclosed the extent of content restrictions, or what changes were made to the treatment of content.

Conversely, there are reports of content posted by organized extreme right-wing Israeli groups on Facebook, some of which appears to amount to advocacy of hatred that constitutes incitement to violence against Palestinian communities living inside Israel and in Palestine.

Applicable international human rights standards

Facebook has a responsibility to respect human rights under international law. In its recently adopted Corporate Human Rights Policy, the company has pledged to respect human rights "as set out in the UN Guiding Principles on Business and Human Rights (Guiding Principles)".[1]

A rights-oriented approach to content moderation calls for the use of human rights impact assessments for product and policy development, which should inform operational and policy decisions and be periodically reassessed and subject to public and civil society consultation.[2] HRIAs can help Facebook to identify, prevent and mitigate adverse human rights outcomes linked to future revisions of its content moderation policies and processes.

Facebook's reliance on Community Standards that are based on a 'set of values' has led to unpredictable and potentially unsafe online spaces for users, as well as public criticism. In particular, when Facebook adopts ad-hoc measures in response to crises, such as the situation in Palestine and Israel, it is frequently criticised for arbitrariness and bias. In contrast, international human rights standards "enable companies to create an inclusive environment that accommodates the varied needs and interests of their users while establishing predictable and consistent baseline standards of behavior."[3]

Facebook should align its content moderation measures with international human rights standards, and transparently communicate to users and the public what measures are being adopted, their rationale, and how Facebook considers and applies international human rights standards. The Oversight Board should also exercise due care when drafting its decisions to ensure that its consideration of relevant principles of international human rights law are given appropriate prominence so that it is clear that they take precedence over Facebook's stated values.

Facebook's content moderation policy and practices should recognise that speech (on matters of public interest in particular) is entitled to strong protection under international human rights law. The Human Rights Committee, which is responsible for monitoring the ICCPR's implementation, has interpreted expression protected under Article 19(2) broadly to include "political discourse, commentary on one's own and on public affairs, canvassing, discussion of human rights, journalism, cultural and artistic expression, teaching, and religious discourse". The Committee has further stated that "expression that may be regarded as deeply offensive" is protected under international law.

[1] A/HRC/17/31, see especially Principle 11: companies should "avoid infringing on the human rights of others and . . . address adverse human rights impacts with which they are involved."
[2] *Ibid*
[3] A/HRC/38/35

Freedom of expression may only legitimately be restricted where it meets the three-part test under article 19(3) of the ICCPR, meaning that it should pursue a legitimate aim as prescribed under international human rights law, have a basis in law that is of sufficient quality, and be necessary and proportionate to achieve the legitimate aim. The Human Rights Committee has clarified that "law" must be "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly." Although Facebook does not issue laws, its rules and policies could usefully be guided by these general principles.[4]

Responding to 'hate speech'

Article 20(2) of the ICCPR requires States to ban "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence". Restrictions pursuant to article 20(2) of the ICCPR must also satisfy the conditions of legality, necessity and legitimate aim as set out in Article 19(3) of the ICCPR.[5]

In making a determination whether content should be restricted, Facebook should also consider the six factors outlined in the Rabat Plan of Action: the context of the speech, the status of the speaker, their intent, the content and form of the speech, its reach, and the likelihood and imminence of it causing harm. Prohibitions should focus on speech that is intended and likely to incite the audience of that speech to engage in acts of discrimination, hostility or violence against a protected group, rather than advocacy of hatred without regard to intent or likelihood of actually inciting a prohibited action against a protected group. Importantly, the ICCPR does not permit prohibition of advocacy of minority or even offensive views that do not amount to incitement.

International law requires use of the least restrictive measure available to confront the problems of 'hate speech'. In the context of content moderation, this means that Facebook should carefully craft proportionate responses within its content moderation policies, avoid suspending/banning users or deleting content unless the test above has been met, and ensure that due process standards are in place and followed so that users may understand the reasons for an enforcement decision and promptly appeal it where necessary. It may be appropriate to use measures such as downranking, demonetizing, friction, warnings, geoblocking and countermessaging, and the criteria for application of these measures should also be transparently disclosed to allow users to govern their behaviour accordingly and understand when these measures have been applied.

Clarifying and disclosing the list of Dangerous individuals and organisations

In addition, Facebook should continue to revise its policies concerning 'dangerous' individuals and organisations, and ensure that the terms it uses are guided by international human rights instruments on counterterrorism and are strictly limited by the principles of legality, necessity and proportionality.[6] In particular, Facebook should review its use of the terms "praising", "glorifying", or "justifying" terrorism, so that they are clearly defined to avoiding causing unnecessary or disproportionate interference with freedom of expression.[7] Bans using the terms 'praise and 'support' are 'excessively vague'.[8] Similarly, there is a risk that "poorly defined concepts [are used] to suppress political opposition or ideological dissent from mainstream values".[9]

---

[4] OL OTH 24/2019
[5] General Comment 34, para. 50. See also A/67/357
[6] A/HRC/16/51, Practice 7 and 8 and see also A/59/565 (2004), para. 164(d)
[7] General Comment 34, para. 46
[8] A/HRC/38/35, para. 26
[9] A/HRC/31/65

Facebook's Community Standard on dangerous individuals and organisations, updated on 23 June 2021, is an improvement on the previous definition in that it gives more detailed definitions and descriptions of designated individuals and organisations, as well as examples of prohibited conduct. However, the current definition continues to require users who "report on, condemn or neutrally discuss" designated individuals or organisations to "clearly indicate their intent", and notes that if the intent is unclear, Facebook "may remove content". Users are not told how to indicate their intent, giving rise to the possibility of arbitrary interpretation or application of the rules.

Facebook's current definition risks limiting legitimate forms of expression, such as reporting conducted by journalists and human rights organizations on the activities of terrorist groups and on counter-terrorism measures taken by authorities, in violation of the right to freedom of expression. International standards are clear that journalists and others reporting on hate speech should be protected against content restrictions or account actions.[10]

The updated Community Standard also continues to define Dangerous Individuals and Organisations based on Facebook's own definition without reference to any external standard, and without disclosing Facebook's internal list of designated individuals and organisations. This is problematic for a number of reasons. First, it is unclear how Facebook interprets and applies the term 'dangerous', and how it determines whether individuals belong to organisations labeled as 'dangerous'. Second, the use of a secret list prevents public comment on the inclusion of listed organisations and individuals. Third, the decision to designate organisations is privately made without consultation or a publicly disclosed procedure, and without an effective means for an individual or organisation to challenge a listing decision once made.[11] Fourth, Facebook has withheld the necessary information for users to understand the limits of permissible speech. Fifth, in many instances users are not provided with any granularity regarding the reason for their content's removal, which can result in downranking of content by the newsfeed algorithms and/or the automatic suspension or revocation of their user access and privileges. Finally, it should be noted that Facebook's policy is at risk of being applied in a way that disproportionately restricts political speech by minority groups and political dissidents.

Facebook should publish its entire list of designated individuals and organisations and put in place due process standards empowering users to understand how assessments are made and appeal decisions regarding the inclusion of individuals or organisations. Facebook should further consider adopting the model definition of incitement to terrorism advanced by the mandate of the United Nations Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism[12] and should be guided by the standards spelled out in the Rabat Plan of Action when addressing advocacy of national, racial or religious hatred that may constitute incitement to discrimination, hostility or violence. For clarity, content should not be removed solely on the ground that it mentions a prohibited individual or organisation.

Transparency

Under the Guiding Principles, "business enterprises whose operations or operating contexts pose risks of severe human rights impacts should report formally on how they address them," and "provide information that is sufficient to evaluate the adequacy of an enterprise's response to the particular

[10] A/74/486
[11] A/65/258, paras. 53-58
[12] A/HRC/16/51, Practice 8

human rights impact involved."[13] The Guiding Principles provide for transparency in "a variety of forms, including in-person meetings, online dialogues, consultation with affected stakeholders, and formal public reports."[14]

In light of these standards, Facebook should engage more closely with relevant local communities and their representatives, including in Palestine, as set out in the UN Guiding Principles 17−19, and provide clearer guidance on how external input is solicited and integrated into the company's decision-making process. The platform should also disclose more information about how it applies the processes it has developed for flagging content. As well, greater transparency is needed on governmental requests to either remove or otherwise restrict speech or conversely give more prominence to government messaging.[15]

Recommendations

By promising to adhere to international human rights law, Facebook has raised the public's expectation that it will realign its Community Standards to human rights standards, and the Oversight Board should hold the company to that commitment. The Oversight Board should call upon Facebook to incorporate a human rights approach into its guidelines, standards, considerations, and practices through the following measures:

1. **Disclosure of designated individuals and organisations:** The list of designated Dangerous Individuals and Organisations should be published, and Facebook should put in place due process standards empowering users to understand how assessments are made and appeal decisions regarding the inclusion of individuals or organisations.

    a. Facebook should consider adopting the model definition of incitement to terrorism advanced by the mandate of the United Nations Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism.
    b. Facebook should be guided by the six factors set out in the Rabat Plan of Action when addressing advocacy of national, racial or religious hatred that may constitute incitement to discrimination, hostility or violence.

2. **Transparency:** All requests to remove content or suspend/delete accounts by governments, including the Israeli Ministry of Justice's Cyber Unit, should be disclosed to the public. The data provided should include the basis for the request (whether there is a violation of national law or a request for 'voluntary' removal), the number of requests, and action taken in response (whether to restrict the visibility of content or conversely increase the reach of pro-government messaging).

3. **Algorithmic transparency:** Facebook should ensure transparency of its automation and machine learning algorithms so that users and the public can understand how the platform moderates content, in particular relating to the Palestinian conflict. The transparency should include error rates as well as classifiers used.

[13] UN Guiding Principles, Principle 21
[14] *Ibid*
[15] A/HRC/47/25 at para. 82

4. **Consultation:** Facebook should undertake meaningful consultation with potentially affected groups and other stakeholders, including in Palestine, and appropriate follow-up action that mitigates or prevents these impacts. Facebook should also conduct ongoing review of its efforts to respect rights, including through regular human rights impact assessments, consultation with stakeholders, and frequent, accessible and effective communication with affected groups and the public, in line with Guiding Principles 20−21.[16]

[16] A/HRC/38/35