Draft 'Effective Guidelines on Hate Speech, Social Media and Minorities'

An initiative by the Mandate of the United Nations Special Rapporteur on Minority Issues, Dr. Fernand de Varennes

Background

Throughout his mandate the UN Special Rapporteur on Minority Issues, Professor Fernand de Varennes, has adopted the unique approach of holding regional forums with the aim of making the annual UN Forum on Minority Issues more accessible for minorities, and to provide effective responses to problems affecting them, while taking into account regional contexts and realities.

The regional forums are convened by the UN Special Rapporteur on Minority Issues and are organised by the Tom Lantos Institute, in cooperation with local and regional partners. They inform the work of the UN Forum on Minority Issues with region-specific recommendations.

In 2020, the theme of the regional forums and the UN Forum in Geneva was "Hate Speech, Social Media, and Minorities" and had the following objectives:

- Promote the understanding of the various forms and the harmful impact of online hate speech against persons belonging to minorities, as well as the role of social media in the dissemination of hate speech;
- Analyse and discuss the legal, institutional and policy challenges concerning the regulation of hate speech against minorities on social media platforms;
- Develop effective strategies to restrict the dissemination of hate speech against minorities on social media platforms, referring to good practices, in accordance with international human rights norms;
- Strengthen the participation of minorities and their representatives in global discussions on online hate speech as well as in the development of relevant laws and policies;
- Strengthen partnerships among various stakeholders so they can address hate speech against minorities on social media platforms more effectively.

The European Regional Forum took place on 20-21 September 2020. The recommendations made at the Forum, and the video recording of the event can be accessed at Minority Forum Info. The Asia-Pacific Regional Forum took place on 19-20 October 2020.[1]

In total, both regional forums produced 127 specific recommendations aimed at a wide range of stakeholders to improve the protection of minorities from hate speech on social media platforms underpinned by international human rights law. The present draft 'Effective Guidelines on Hate Speech, Social Media and Minorities' seek to build on these in order to provide detailed guidance on how social media companies can improve their content standards and their enforcement to combat online hate speech, and in particular when targeting minorities.

The Special Rapporteur is pleased to share an initial draft of part of the Guidelines and the launch of a global consultation at RightCon 2022 to scrutinise and improve them. This is still very much a work in progress, and there remain some Guidelines that the Special Rapporteur is undecided on, which he has opted not to share in full.

The Special Rapporteur intends to subject the draft Guidelines to both an open public consultation and series of expert roundtables focusing on the application of international human rights law (and minority rights) to social media companies (SMCs). This has conventionally centred on balancing the right to freedom of expression with the prohibition on incitement to hatred. It is hoped that this

[1] For more information about previous and subsequent regional forums, please visit this page. The thematic focuses for 2019 and 2021 were "Education, Language and the Human Rights of Minorities" and "Conflict Prevention and the Protection of the Human Rights of Minorities", respectively.

framing can be advanced and enriched with a minority rights perspective grounded in ICCPR Article 27 and the UN Minorities Declaration[2].

[2] Declaration on the Rights of Persons belonging to National or Ethnic, Religious and Linguistic Minorities: resolution / adopted by the UN General Assembly (47th sess. : 1992-1993).

Introduction

The Holocaust was the single worst act of genocide in the twentieth century. It also sparked an international desire for peace and supplied the impetus for the work which would lead to our rules-based international order underpinned by the guarantee of fundamental human rights in 'recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family' as 'the foundation of freedom, justice and peace in the world'.[3]

The resulting commitment to 'liberate mankind from such odious scourge' and as such to 'prevent and punish'[4] the crime of genocide was not sufficient to avert genocides in Rwanda and Bosnia. What is often overlooked is that all three instances targeted and sought to eliminate minorities.[5] Despite reasserting 'never again' in the aftermath, we struggle even today to prevent acute violence against minorities in ongoing situations which may amount to ethnic cleansing and even genocide.

What is remarkable is that the above instances did not begin with killing, but with dehumanising words, usually targeting those belonging to ethnic, religious or linguistic minorities.[6] As I and others have pointed out in the past, the Holocaust did not begin at Auschwitz, it started with hate speech against a minority. While in the past, vile propaganda and identity-based hatred was carried in print, radio and broadcasts[7], today's hatred is conveyed online, instantaneously, without editorial controls, to an audience of millions, unrestricted by national boundaries, the commitment of any resources, or the need for the backing of an organisation. Any content standards are applied following content being posted publicly rather than before and associated systems are often slow and unable to cope with the sheer scale of online content being posted at any given time.

The phenomenon of social media in the last twenty years has both been a force for democratic mobilisation, but also of unrestrained proliferation of hateful narratives and stereotypes. Yet despite online hate speech mainly targeting minorities at greater risk of communal violence and even ethnic cleansing and genocide, there has been scant focus on their protection in the online space and in particular on social media platforms.

Cognisant of the above, it is intended that these Guidelines bring together previously disparate and unconnected areas of international human rights law in the specific context of online expressions on social media that facilitate the sharing of user-generated content. These are the right to freedom of expression, minority rights, the obligations to protect individuals from harm, advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, and the human rights responsibilities of private of multilateral companies, specifically those that are commonly referred to as social media companies or platforms (SMC).

---

[3] Universal Declaration of Human Rights, 1948 (UDHR), Preamble.

[4] Convention on the Prevention and Punishment of the Crime of Genocide, 1948 (Genocide Convention), Preamble & Art. I.

[5] See Genocide Convention, Art. II: "genocide means any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group".

[6] International Convention on Civil and Political Rights, 1966 (ICCPR), Art. 27: "In those States in which ethnic, religious or linguistic minorities exist, persons belonging to such minorities shall not be denied the right, in community with the other members of their group, to enjoy their own culture, to profess and practise their own religion, or to use their own language."

[7] Genocide Convention, Art. III(c) mandates the punishment of "Direct and public incitement to commit genocide".

The UN Guiding Principles on Business and Human Rights[8] have emerged as the golden standard by which private actors, including social media companies, are increasingly held and are holding themselves to account. It has at its core concerns related to severity of human rights harms, their causality to the companies' actions and most importantly the cross-cutting theme of protection of vulnerable and marginalised groups, including minorities. It is hoped and intended that these Guidelines can be utilised as an extension and elaboration of how such companies should achieve this in relation to the specific intersection of online hate speech targeting minorities on social media platforms.

[8] UN Guiding Principles on Business and Human Rights, UN OHCHR, 2011 (UNGP).

Underlying Principles & Proposed Definition

PRINICIPLE 1: Social media companies (SMCs) should not offer protection to minorities less than required under international human rights standards aimed at States in the area of incitement to hatred and hate speech.

PRINICIPLE 2: Regardless of the extent national laws incorporate international human rights standards, SMCs should adhere to international human rights obligations.

PRINICIPLE 3: SMCs should offer increased protections to community members, given the lesser interference with freedom of expression as is required from States. There is seldom a complete nullification of the right to freedom of expression or the granting of an absolute freedom of expression despite the harm it may cause. Often there is an interference or limiting of expression, which has to be justified with the purpose of that limitation. This can be seen as an exercise of 'proportionality'[9] or 'balancing' of competing rights[10].

Proposed definition of 'Online Hate Speech'

There has been no attempt to define online hate speech specifically. We do however have the standard of prohibition of incitement to hatred under ICCPR, Art. 20(2) and the Rabat Plan of Action.[11] Most recently, the UN Office of Genocide Prevention and Responsibility to Protect has elaborated a definition of hate speech more generally and broader in scope than ICCPR, Art. 20(2):

> "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."[12]

It has also become apparent that social media companies are having to define what 'hate speech' is on their platforms and that they are in need of assistance, support and guidance to ensure compliance with international human rights law, any minimum criteria and consistency of standards across industry. Drawing on and synthesising this pre-existing work, an authoritative working definition of 'online hate speech' and associated core responsibility to support and enhance existing approaches by social media companies can be proposed:

> *"Social media companies are responsible for effectively prohibiting and removing content in the shortest time possible that is discriminatory, hostile or violent towards those (community members) with protected characteristics on the basis of any identity factor and especially those belonging to national or ethnic, religious and linguistic minorities on the basis of their minority identity"*

---

[9] ICCPR, Art. 19(3) requires permissible limitations on the right to freedom of expression to be necessary, which is in turn establishment of proportionality.

[10] UNGP.

[11] Rabat Plan of Action, Appendix, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, 2013.

[12] UN Strategy and Plan of Action on Hate Speech, 2019 and UN Strategy and Plan of Action on Hate Speech, Detailed Guidance on Implementation for UN Field Presences, 2020.

Substantive Guidelines

1. Social media companies (SMC) should clearly and precisely define 'hate speech' in their content policies and expand their protected characteristics to include any identity factor. Hate speech targeting minorities should be a distinct category and include national or ethnic, religious and linguistic minorities.

Commentary

SMC content policies[13] on hate speech should be legally certain imbuing foreseeability and accessibility. Broad, subjective or ambiguous terms such as 'direct' and 'attack' should be clearly defined, detailed and their thresholds elaborated to allow for transparent and objective interpretation and application. There would be increased legal certainty in referring to established and objective terms such as 'discriminatory' instead[14]. Examples of appeals and their decisions should be publicly shared to aid clarity and provide precedents. The included protected characteristics should not be arbitrarily chosen and instead brought in line with permitted heads of non-discrimination under international human rights law[15] and as per the understanding of 'hate speech' in the UN Strategy and Plan of Action on Hate Speech. As such, the included protected characteristics should not be exhaustive and establish principled criteria to ascertain inclusion such as 'any identity factor'.[16] Notably while, ethnic and religious minorities may be protected from hate speech to some extent on the basis of their ethnic, racial or religious characteristics, linguistic minorities are often excluded.[17] When there is intersectionality between two or more identity factors, the potential of harm and thus severity of hate speech is greater and should be recognised as such with appropriate and proportionate content moderation responses.[18]

While 'hate speech' defined as targeting those with protected characteristics may include minorities, it is nonetheless a distinct type of hate speech that is not only severer, but also poses the worst risk of widespread, systemic and group-based violence resulting in atrocity crimes such as ethnic cleansing and genocide.[19] Such hate speech targets an entire minority group on the basis of their national, ethnic, religious or linguistic identity and culture, leads to incitements to violence, violent hate crimes and ultimately violence en masse. The creation of such a distinct category is supported by several normative standards. These include protection of ethnic, religious or linguistic minorities under ICCPR Art. 27; prohibition of incitement to hatred, hostility and violence under ICCPR Art. 20(2), which singles out minority groups based on national, racial and religious basis; ICERD Art. 4, which notes the specificity of dissemination of ideas of racial superiority in relation to minorities[20]; the international crime of inciting genocide and States' obligation to prevent and punish genocide[21]; and the requirement in the UN Guiding Principles on Business and Human Rights[22] for private companies to take special note of

[13] These are referred to under various headings by social media companies such as 'Community Guidelines', 'Community Standards', 'Rules'.

[14] 'Incitement to discrimination' is the lowest threshold for 'advocacy of national, racial and religious hatred' under ICCPR, Art. 20(2).

[15] ICCPR, Art. 2 & 26. ICERD Art. 1. General Comments.

[16] UN Strategy and Plan of Action on Hate Speech 2019, and its Detailed Guidance, 2020.

[17] ICCPR, Art. 27 and the UN Declaration on Minorities, which also includes 'national' minorities.

[18] Detailed Guidance, 2020

[19] Genocide Convention, Rome Statute and other relevant instruments.

[20] CERD General Recommendation No. 35, Combating racist hate speech, 26 September 2013.

[21] Convention on the Prevention and Punishment of the Crime of Genocide, art. 1.

[22] UNGP, General Principles, p. 1: "These Guiding Principles should be implemented in a non-discriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by,

the potential harm to marginalised and vulnerable groups, who in most cases are minorities as understood under international law. As such, the types of minorities that should be included in this distinct category of hate speech can be 'ethnic', 'racial', 'national', 'religious', 'linguistic'. This should also be a non-exhaustive list and should have the principled criteria of any risk of group-based violence on the basis of group identity. Therefore, indigenous peoples should also be included.[23]

2. SMCs should not create overly broad and ill-defined 'public interest' exemptions which give them wide and opaque discretion to permit hate speech in violation of their content policies.

Commentary

Such a discretionary tool is sometimes referred to as the 'newsworthiness' or 'public interest' exemption and has been contested especially when applied to political actors and political leaders. It is justified on the basis of wider scope of freedom of expression for politicians given their elected and representative role or when it is necessary to make the public aware of negative statements given the status of the person and its perceived democratic importance.

Firstly, under the right to freedom of expression, certain categories of speakers have a broader scope of freedom of expression, notably academics, politicians, journalists, or judges.[24] At the same time public interest is the basis of permissible limitations to freedom of expression[25] and the prohibition of incitement to hatred, hostility and violence[26] with the aim of protection of minorities[27]. Therefore, both public interests must be balanced: the wider scope of freedom of expression against the harm caused or posed to a protected group or minority. Secondly, the wider freedom of expression for these specific groups is derived from their special role and status in the functioning of democratic societies. Hence, the extra leeway given must be related to the performance of this special function and cannot be contrary to a democratic society or aimed at the destruction of the rights of others.[28] While such political or other actors may be able to raise and debate controversial topics while fulfilling their role, they would not be able to on the 'street' or other public or private spaces.[29] Thirdly, the status, reach and influence that comes with these special categories of individuals can also lead to an increased likelihood of inciting hatred, hostility and violence.[30] As such, the application of a 'public interest' exemption should be made less likely.

At present, the legal certainty of these discretionary exemptions is dire and worse than the definitions of hate speech. To have resort to a discretion so ill-defined and broad, SMCs risk rendering their content policies arbitrary and devoid of consistency and predictability. Having such a veto even for violating hate speech content would also open allegations of commercial consideration interfering in determining and preventing harm in particular relating to hate speech. Transparent operation and rationales

individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men"

[23] UN Declaration on Indigenous Peoples.

[24] ICCPR, Art. 19, General comment No. 34, Article 19: Freedoms of opinion and expression

[25] ICCPR, Art. 19(3).

[26] ICCPR, Art. 20.

[27] ICCPR Art. 27 and UN Declaration on Minorities.

[28] ICCPR, Art. 19(3) and ICCPR, Art. 17.

[29] Jersdilk v. Denmark, Application no. 15890/89, ECHR, 23 September 1994.

[30] Rabat Plan of Action, 6 part test.

employed for the 'public interest' exemption, which could combat such a perception, remains absent due to opaque details of its application. This should be remedied by applying the above three principles.

3. SMC's should balance the potential harm to protected groups and minorities on the basis of necessity and proportionality, including when deciding the appropriate content moderation response.

Commentary

Beyond clearly defined content policies on hate speech compliant with international human rights laws, there is a need to adopt a focused application of existing standards to the online context to determine where the precise substantive line should be drawn between free speech and hate speech. The right to freedom of expression[31] extends to online expressions[32] and falls within the scope of "freedom to seek, receive and impart information and ideas of all kinds".[33] It is fundamental and essential to other human rights and the proper functioning of democracy[34] as well as its associated institutions. It even extends to controversial and offensive views[35]. It may be permissibly limited to protect the rights of others or on pressing public policy grounds[36] as long as it is legally defined, pursues a legitimate aim, and is necessary and proportionate in a democratic society.[37] This remains the most appropriate and correct framework to apply to permissibly limiting online hate speech provided that some aspects are adopted for the online context.

First, the application of 'legality' to SMCs content policies on hate speech requires that its definition and enforcement satisfies the standard in Guideline 1 above.

Second, the only 'legitimate aim' that SMC's can pursue is that of "protecting the rights of others". The remainder are intended for States and inapplicable to SMCs. As such, States may deem it necessary and proportionate to require limitations on others' expressions by SMCs in very narrow circumstances based on other enumerated legitimate aims such as 'national security'. This also correlates with the requirements of the UNGP to balance competing rights of their community members, which in this case would be the right to freedom of expression and the right to protection from hate speech.[38]

Third, 'necessity' and 'proportionality' should be determined on the basis of there being no alternative means to ensure the same result or the protection of the concerned protected group or minority.[39] In other words, it should be the least restrictive option available. 'Proportionality' also entails that a range of enforcement responses should be availed on a spectrum of necessity and severity. These can include warning labels, limits on virility (through disabling engagements such as 'likes', 'shares' and 'replies' as well as algorithmic downranking – sometimes referred to as 'friction'), removal of content, temporary suspension of community members and ultimately their permanent exclusion.

---

[31] ICCPR Art. 19.

[32] Report on Content Regulation, Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (David Kaye), 2018 (A/HRC/38/35).

[33] ICCPR, Art. 19(2).

[34] General comment No. 34, Article 19: Freedoms of opinion and expression.

[35] Ibid.

[36] ICCPR. Art 19(3): "a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals."

[37] ICCPR Art. 19(3).

[38] UNGP.

[39] General Comment 34, par. 20.

'Proportionality', in light of the worst possible response being the inability to make public statements on a given online platform, requires that it is not applied in an identical manner as would be the case of State practices, which mostly concern criminal laws that either prevent expressions vital to democratic governance or incarcerate those human rights defenders and journalists, among others, who express them. The range of responses available are limited to individual SMCs and where they do not extend to criminality and illegality (which should engage States' international obligations), may necessitate a higher level of protection prioritising the protection from hate, violence, hostility and discrimination and where their proportionality of response is assessed on the basis of inaccessibility to a single online platform rather than illegality or criminality.

The notion that the broad classification of expressions as 'hate speech' goes beyond the incitement to discrimination, hostility and violence that requires the specific response of prohibition by law[40] and need only satisfy the tests of legitimate limitations to the right to freedom of expression is reinforced by the Rabat Plan of Action. It sets out three categories of hate speech: incitements to discrimination, hostility or violence that must be criminalised following consideration of the 6-part threshold test, incitements that must be limited by law and those that can be legitimately limited under ICCPR Art. 19(3). This has been further asserted by the special procedures and the UN Strategy and Plan of Action on Hate Speech. This means that 'hate speech' can be categorised broadly and should not be limited to proven examples of incitements of hatred required to be prohibited by States under ICCPR Art. 20(2), but just the expression of hate speech as necessitated by the framework set out in ICCPR, Art. 19(3) and as stipulated in the UNGP.[41] The 'prohibition' standard for incitements to discrimination, hostility and violence should not be applied by SMCs, States or other stakeholders in defining SMC's responsibilities with respect to protect their users from hate speech.

4. SMCs should clarify their definition of 'incitement' and at least meet and ideally exceed the protection offered by the prohibition on incitement to discrimination, hostility and violence applicable to States.

Commentary

SMCs have distinct policies on 'incitement' distinct from 'hate speech'. However, such policies focus too narrowly on the incitement of violence only. International human rights law requires the prohibition by law of such incitement along with incitements to discrimination and hostility. It may be argued that these two forms of lesser incitement are subsumed within the general definition of 'hate speech'. However, it remains possible that an 'incitement to discrimination and hostility' may not be captured by 'hate speech' policies for the reason that incitement is often aimed at prospective perpetrators of hate speech rather than aimed at protected groups or minorities. It may on the other hand be included within the meaning 'hate speech' but not be categorised as a far more severe type of hate speech and thus requiring not only speedier detection but a more robust enforcement response.

States must also bring their laws in compliance with the requirements of ICCPR, Art. 20(2), hold perpetrators to account and partner with SMCs as well as stipulate their precise responsibilities to identifying and reacting to such illegal content. Where States do not meet the requirement for protection from incitement to hatred, SMCs should still at a minimum apply the general standard laid out in ICCPR Art. 20(2) and as obliged by the UNGPs.

[40] ICCPR, Art. 20(2).

[41] UNGP.

The ideal standard adopted by SMCs and required by States should ideally go higher as stated above in line with the aspect of online speech to proliferate at speed, the extent of the potential response as removal of community members from one platform. This would in practice mean that the threshold of criminal law laid out in RPoA should not need to be met for incitement to be applicable in the context of SMC content moderation. This has a number of facets. Firstly, all 6 parts of the test need not be satisfied for incitement to be found. As such if focus is placed on protecting protected groups and especially minorities from hate speech or incitement to hatred, emphasis should be placed on the outcome. Intent may be replaced with 'effect' or 'likelihood' and 'imminence' with 'some causality and risk' (UNGP). It should also be noted that fulfilling some of the 6 parts in the online context need not be as difficult to satisfy as in other ones. These include the idea of 'reach' and 'status' in the online setting need not attach to a prominent personality, figure or politician with a vast dedicated following. Such expressions will now doubt be even further amplified, but it does not discount the possibility of a publicly unknown individual gaining an audience of millions to acutely hateful online rhetoric. The social media context also necessitates the introduction of a further aspect not mentioned in RPoA and relevant to the UNGP of exposure which is a product of virility multiplied by time. This underscores the vast reach and instantaneous speed at which hate and incitement to hatred can proliferate.

A separate dimension to the types of expression which may be included in incitement policy is the necessity to expand the definition to include incitement to discrimination and hostility, not just violence. Such a distinction is arbitrary and becomes engaged too late in the process of escalation envisaged by ICCPR Art. 20(2) in that what culminates in violence begins with incitements to discriminate and then to hostility. Therefore, it remains imperative for an approach focused on the protection of protected groups and minorities from incitement to hatred that results in eventual violence that protection should also be offered beginning that the level of incitement to discrimination. In aspiring towards this higher level of protection of human rights that SMCs can seek to implement it could be argued that any statements which are discriminatory in nature or may normalised discriminatory attitudes towards protected groups and minorities should be limited.

It must be emphasised that any limitations on online expressions based on an 'incitement' policy must satisfy the three-part test for it to be permissible in compliance with the right to freedom of expression legality, legitimacy, and necessity and proportionality (elaborated and contextualised in detail above). Ultimately, any such 'incitement' content policy must ensure that whatever criteria are employed, the protection offered does not become illusory to such an extent that it never practically succeeds in preventing violence, but rather only becomes engaged as an enforcement response once the violence has occurred. This is the ultimate test of an effective 'incitement' content policy.

5. SMC's should publicly make available how they identify content that falls short of hate speech content policies and is algorithmically demoted, and likewise steps in place to ensure hate speech is not promoted or becomes virulent prior to removal.

Commentary

SMCs are responsible not just for the presence of user-generated content on the platforms but how that content is curated, packaged, presented, promoted and amplified. In this sense, the harm emanating from 'hate speech' and 'incitement to discrimination' targeting minorities does not just depend on its severity but also its reach, which in turn is determined by a number of automated systems employing artificial systems reliant on complex algorithms. Historically, the greatest criticism of SMCs has been that their design was intended to cause controversy, adversarial interactions and create self-reinforcing and polarised echo chambers. This had positive ramifications through the stimulation of democratic public debate and for community members to form groups of likeminded members with similar

interests. However, it also created an environment where provocative hate speech often elicited higher levels of engagement and allowed a digital climate where extreme, intolerant or violent elements gravitated towards each other and became further radicalised. This, the argument continued, maximised community member engagement and hence monetised of extreme views and spaces.

Two interrelated issues need to be addressed in compliance with international human rights law. The first, the criteria for any online expressions or content being identified as borderline or not quite violating hate speech or incitement policies must be publicly stated and clearly defined. To have a confidential policy for demoting content that is close to violating content policies but does not cross the threshold cannot and should not be treated differently without the rationale being elaborated and opened to scrutiny and appeal. Without this process, such confidential policies do not satisfy the lowest standards of legal certainty and raise serious concerns about interfering with controversial expressions explicitly protected under ICCPR Art. 19 rather than hate speech. These intermediate content policies focused on limiting the proliferation of content rather than its removal should be clearly stated, intended and necessary for the prevention of harm to protected groups and minorities from hate speech. The responses taken must then be proportionate to the perceived harm whether algorithmic non-amplification, the appending of warning labels, click-through shield, disabling/limiting engagement or sharing. Such policies and their enforcement measures must not be disproportionate in exceeding what is necessary to prevent harm to minorities, but also must not apply measures which do not achieve the desired aim. An example of this would be the attaching of warning labels or creating click-through shields that do little to mitigate the spread of hateful content, especially when targeted at devoted and ardent following rather than the general public. This is especially true for extremist views and sentiments.

The second is that hate speech that does violate SMCs' publicised content policies should not benefit from any amplification through algorithms or user engagement designed to accelerate virility. Put differently, automated content moderation systems must be able to distinguish virility from violating content and virility from permissible content. When assessing the performance of SMC's under UNGP and by regulators and Governments as well as academics and journalists in preventing and combating hate speech, it should not only be whether hate policies were implemented against violating content or whether they were proactively detected but how much exposure through distribution and broadcast it garnered. As such time before removal of content should not be the only factor in this but also the number of people it reached. This should be referred to as the 'rate of exposure to harmful content'. A 'low rate of exposure' over several days may cause less harm than a 'high rate of exposure' over a few hours. Any indication of what is the level of exposure that should be established as an ideal would be to some extent arbitrary. Instead, SMCs should publicly share such data/metrics in their transparency reports and strive to constantly improve their systems and these measurements.

6. Both automated and human moderation must be available in the languages most prevalent on the platform but also the languages of minority groups or others at risk of human rights abuses in particular violence.

Commentary

SMCs have a responsibility to anticipate and prevent harm against their community members in particular those that are vulnerable and marginalised in particular minorities.[42] There is no reasonable or objective basis to offer a higher level of protection in some regions of the world and not others. While there may be commercial and reputational reasons for content moderation resources to be allocated to

[42] UNGP.

Western or European States, international human rights law necessitates no such discrimination and in fact require the allocation of moderation resources on the basis of risk of harm. This inadvertently happens to be in less developed states from the global South. Indirect discrimination may also occur in the absence of several languages in which Facebook allows use of its platforms, more so in languages that it does not accommodate on its platform or even the moderation of others.

When an SMC opts to operate in a certain country, it undertakes to be responsible for the potential harm it could cause to prospective users of its services. SMC's merely due to resource implication or arbitrary criteria facilitate automated and human content moderation for some linguistic groups and not others. Furthermore, if only the majority languages are subjected moderation systems and enforcement of content policies, then it will be mostly minorities who will be excluded which being some of the most at risk of discriminatory treatment and violence.

This is exacerbated by even a smaller number of languages in which content policies are available. For those languages in which content policies are available, it is not clear that community members in those States or regions are aware of their rights to protection and complaint under content policies. For this to be effective SMCs must run periodic public awareness and education campaigns, in particular on their own platforms, and in all spoken languages, and if not that the number that covers the most people and the most marginalised and vulnerable people so that their rights and protections offered by SMCs are not illusory.

7. Human and automated moderation should be unbiased and involve consultation with minorities. Ideally, human moderators should be from diverse backgrounds and representative of protected groups and minorities.

Commentary

SMCs should ensure that automated artificially intelligent detection and removal systems are free from bias of any kind against protected groups and especially minorities. This can be in the form of the types of content they have been designed to identify. Pertinent questions are whether the lexicon of hate speech terms is up to date and includes all protected groups and minorities. Such terms are constantly changing and developing, and keeping AI systems effective and current, mechanisms must be created for continuous consultation with those local linguistic and cultural knowledge, in particular members belonging to minorities or minority civil society organisations.

Similarly, human moderation is the most effective means of content moderation but is the most resource intensive and time consuming. However, there are situations where high levels of human moderation can be dedicated to a particular language and country. It can however still miss important hate speech if they are not culturally aware of the local context. A greater issue arises when the moderators are linguistically competent and culturally aware, but themselves are consciously or unconsciously biased and belong to the majority ethnic, religious or linguistic group or even themselves believe in hateful tropes and stereotypes against minorities and protected groups. This should be combatted through training that identifies and rectifies such biases but the most effective way to implement this guideline is to strive to employ human moderators belonging to protected groups and minorities.

8. SMCs have a responsibility to provide content policies in the various languages used by their community members, in particular those languages that SMCs function in. Content policies on hate speech must be especially accessible to linguistic minorities at risk of violence or incitement to hatred or hostility.

Commentary

When SMCs make their services available in a language and thus to speakers of those languages, they have an increased responsibility to make content policies accessible in the same languages. There is often a disparity between the number of languages that SMCs accommodate so as to be commercially advantageous and the translation of content policies and moderation which takes place in far fewer languages which is resource intensive in developing AI systems and assigning increased human moderators which is thus commercially disadvantageous.

Choosing which languages, the content policies are translated into is normally based on scale of use and commercial considerations. This means that it will often be minorities who are excluded from accessing content policies as they will, by and large, also constitute numerical minorities and may even privilege majority populations. The responsibility might be lesser, but still SMCs could have the additional responsibility to remove spoken languages that are expressed through the script of other languages.

9. Transparency reports should provide data on all content moderation relating to hate speech and minorities. This should be disaggregated in a manner such that those protected groups or minorities most at risk or under threat should be discernible and States and regions in question. It should not be limited to just content removals but should be across the range of responses taken.

Commentary

SMCs are increasingly issuing periodic transparency reports carrying data on removal of violating content across categories of harmful content. While these show the number of removals for violating policies such as hate speech, incitement and dangerous organisations and individuals, they do not show which protected groups or minorities were targeted the most or the least and which States this data relates or the division of languages the content was posted in. Being transparent with regards to those who are most at risk and where they are located can play a vital role publicly demonstrating where and why resources need to be allocated to mitigated escalations and severity of harms.

It can also be relied on as a vital advocacy tool by civil society organisations and spur concerned States to take concrete policies to address such societal issues. Simultaneously it can also instil and inspire public and governmental confidence and trust in the relevant SMC in genuinely being concerned about community members who used the service and the potential harm to society that can occur. In the worst of cases, it can allow for an allocation of increased human moderators and the limiting or cessation of services to prevent the incitement and organisation of mass violence, up to the level of genocide.

Disaggregating moderation data along the lines of perpetrators, terms used, prevalent languages, type of hate speech and that which targets minorities as well as severity of hate speech can all considerably improve in mapping where, against who, by whom and severity of hate speech to encourage collaboration thinking action to address the issue. The range of responses taken should also be listed

such as which moderation response was taken beyond removal such as warning labels, limiting engagement, sharing, geo-blocking, account limitations.

10. Independent, academic and civil society actors should be allowed meaningful access to data to collaborate with SMCs on verifying and gauging effectiveness of content moderation of hate speech particularly against minorities with a focus on how much hateful content is not removed and why, as well as community members' experiences and perceptions.

Commentary

The current models for gauging effectiveness of hate speech content moderation focus on content removals. These do not cater for hateful content that is not detected and receives no complaints. They also do not factor in wrongful removal by automated systems that were not appealed. Both of these occurrences are worsened if community members are not aware of the hate speech policy, or do not know how it is applied and do not know how to flag such content.

In working with third party independent researchers, SMCs can work towards independently verifying the effectiveness of their content moderation systems by allowing for a more detailed and nuanced picture where it may not be working. It will also lead to developing methodology and a multi-dimensional approach to measuring harm that goes beyond views, but also reflects virility and severity. It will allow for more authoritative and accurate research into the linkages of online hate and offline hate and violence and to understand how to disrupt such causation.

Annex I - Consultation Questions:

Underlying Principles and Proposed Definition

1. The 'Underlying Principles' seek to reconcile the tensions/conflicts between i) international law being aimed at States and not companies, ii) some States' national laws or Government actions falling foul of minimum international standards and iii) that ideally online freedom of expression will require a lower threshold for necessity and proportionality given the severest sanction being termination of an account. Do you agree with this framing and solutions offered by the principles? If so, could the wording be improved?

2. Definitions exist of 'incitement to hatred' under ICCPR Art. 20 and one is suggested for 'hate speech' in the UN Strategy and Plan of Action on Hate Speech. No similarly authoritative definition for 'online hate speech' has yet been offered nor has there been any consensus on the precise definition of 'minority' under ICCPR Art. 27 yet. The proposed definition of 'online hate speech' seeks to build on the above. Do you have any views on whether it works, will be of utility and how it could be improved?

Guideline 1

3. Is it feasible to set a common minimum standard for SMCs in relation to their 'hate speech' policies?

4. Do you agree that the list of protected characteristics should be non-exhaustive and inclusion of non-listed protected characteristics should be on the basis of them constituting an identity factor?

5. Do you think that reference should be made to hate speech specifically targeting minorities in light of its unique nature (i.e. targeting of cultural, religious or cultural identity) and the risk of worse consequences (i.e. communal / group-based violence)?

Guideline 2

6. Are there any ways to reduce the risks posed by opaque 'public interest' exemptions and risk that they may be applied in an unfair or discriminatory manner?

Guideline 3

7. Is the approach taken relating to the application of ICCPR, Articles 19 and 20, the correct one? Could it be improved in any way?

Guideline 4

8. As per Guideline 4, is there agreement that there may be a gap that has not been addressed by SMCs between prohibiting 'hate speech' and 'incitements and violence'?

Guideline 5

9. This Guideline is concerned with two issues related to algorithmic amplification and demotion or downranking. Conventionally, 'borderline' content not being removed and not amplified has not been considered an interference with the right to freedom of expression. This Guideline seeks to reframe it as such and thus requiring the application of the 3-part test for a permissible limitation. It is further not clear to what extent is hateful content amplified prior to removal. Any

feedback on these two notions and international law that would strengthen the current framing would be appreciated.

Guideline 6

10. To what extent, should moderation systems (automated and human) incorporate different languages and thus protection of a greater number of communities/societies around the world? Any suggestions how this should be determined in a proportionate way would be instructive to us.

Guideline 7

11. SMCs are asked to consult with those likely to be affected by online hate speech and in particular minorities. What would be the best way to make this expectation measurable rather than just aspirational?

Guideline 8

12. Can SMCs improve accessibility to and understandability of their policies on hate speech for non-English speaking societies and minorities around the world where they operate? In such markets, what should be expected of them?

Guidelines 9 and 10

13. These focus on 'transparency' reporting and increasing 'researcher access', specifically with the reduction of hate speech and the protection of minorities in mind. Apart from recommending the disaggregation of data by group/region and facilitating 3rd party independent research, are there any other ways to strengthen these two Guidelines?

Guidelines under consideration and suggestions for additional Guidelines

14. There are a number of relevant issues that the Special Rapporteur is still undecided on the necessity of dedicated Guidelines. These include:
    - The right to redress and whether and how liability should be legally provided for failure to comply with international human rights obligations,
    - The ability to appeal to a 3rd party and
    - Whether SMCs should be expected to proactively initiate public awareness and education campaigns to help community members better understand hate speech policies and for those likely to post such content to be dissuaded from doing so.

    The Special Rapporteur would welcome views on whether any of these merit inclusion or there are other significant omissions which he should consider adding to the Guidelines.

15. Connected to the above, the Special Rapporteur would welcome any views on whether SMCs should be asked to conduct independent human rights assessments and how best to ensure that they include considerations around hate speech generally and specifically about minorities.

16. The Special Rapporteur would also appreciate any views on whether the Guidelines would benefit from a section that pointed to how 'other actors' could use the Guidelines in their engagement with SMCs such as, but not limited to, advertisers, infrastructure providers, intermediaries or gatekeeper (e.g. app stores), governments, regulators, international organisations, civil society groups.