A/74/48050

Advance Unedited Version

Distr.: General
9 October 2019

Original: English

Seventy-fourth session
Item 72 (b) of the provisional agenda*
Promotion and protection of human rights: human rights questions including alternative means of improving the effective enjoyment of human rights and fundamental freedoms

Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression**

Note by the Secretary General

The Secretary-General has the honour to transmit to the General Assembly the report prepared by the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, submitted in accordance with Human Rights Council resolution 34/18. In this report, the Special Rapporteur evaluates the human rights law that applies to the regulation of online 'hate speech'.

* A/74/50.

** The present report was submitted after the deadline in order to reflect recent developments.

Report of the Special Rapporteur on the promotion and protection of the freedom of opinion and expression

Contents

I. Introduction

1. ‘Hate speech’, a short-hand phrase that conventional international law does not define, has a double-edged ambiguity. Its vagueness, and the lack of consensus around its meaning, can be abused to enable infringements on a wide range of lawful expression. Many governments use ‘hate speech,’ like ‘fake news,’ to attack political enemies, non-believers, dissenters and critics. Yet the phrase’s weakness (‘it’s just speech’) also seems to inhibit governments and companies from addressing genuine harms such as the kind that incites violence or discrimination against the vulnerable or the silencing of the marginalized. The situation frustrates a public that often perceives rampant online abuse.

2. In a world of rising calls for limits on ‘hate speech,’ international human rights law provides standards to govern State and company approaches to online expression.[1] This report explains how those standards provide a framework for governments considering regulatory options and companies determining how to respect human rights online. It begins with an introduction to the international legal framework, focusing on UN treaties and the leading interpretations of provisions related to what is colloquially called ‘hate speech’. The report then highlights key State obligations and addresses how company content moderation may respect the human rights of users and the public. It concludes with recommendations for States and companies.

3. This is the sixth in a series of reports since 2015 that have addressed the human rights standards applicable to freedom of opinion and expression in the Information and Communications Technology (ICT) sector.[2] It should be read in light of the standards and recommendations proposed previously (and which are not necessarily repeated herein). Like the earlier ones, this report draws extensively from existing international standards and from considerable civil society input over the past several years.

II. Hate speech in international human rights law

4. Under international human rights law, the limitation of ‘hate speech’ seems to demand a reconciliation of two sets of values: democratic society’s requirements of open debate and individual autonomy and development with the also compelling obligation to prevent attacks on vulnerable communities and ensure equal and non-discriminatory participation of all individuals in public life.[3] Governments often exploit the resulting uncertainty to threaten legitimate expression such as political dissent and criticism or religious disagreement.[4] Yet freedom of expression and the rights to equality and life, and the obligation of non-discrimination, are mutually reinforcing; human rights law permits States and companies to focus on protecting and promoting the speech of all, especially those whose rights are often at risk, while also addressing the public and private discrimination that undermines enjoyment of all rights.

Freedom of expression

5. Article 19(1) of the International Covenant on Civil and Political Rights (ICCPR) protects the right to hold opinions without interference, and Article 19(2) guarantees the freedom of expression, that is, the right to seek, receive and impart information and ideas of all kinds, regardless of frontiers, through any media. Numerous other treaties, global and

[1] A/HRC/38/35, para. 45. This report uses ‘hate speech’ to refer to obligations and limitations in human rights law that do not use that particular term. See generally Susan Benesch, *Proposals for Improved Regulation of Harmful Online Content*, Paper for the Israel Democracy Institute (2019). Benesch coined a sibling term, ‘dangerous speech’, to identify a “capacity to catalyze violence by one group against another.” See Benesch, *Dangerous Speech: A Proposal to Prevent Group Violence* (2013).

[2] See A/HRC/29/32 (encryption and anonymity), A/HRC/32/38 (mapping the ICT sector’s impact on rights), A/HRC/35/22 (the digital access industry), A/HRC/38/35 (online content moderation), and A/73/348 (Artificial Intelligence and human rights).

[3] See especially Special Rapporteur Frank La Rue’s report on hate speech: A/67/357 (2012).

[4] Id., para. 51 – 54.

regional, expressly protect the freedom of expression.[5] The Human Rights Committee, the ICCPR's expert monitoring body, has emphasized that these freedoms are "indispensable conditions for the full development of the person. . . [and] constitute the foundation stone for every free and democratic society."[6] They "form a basis for the full enjoyment of a wide range of other human rights."[7]

6. Since the freedom of expression is fundamental to the enjoyment of all human rights, restrictions on it must be exceptional, subject to narrow conditions and strict oversight. The Human Rights Committee has underlined that restrictions, even when warranted, "may not put in jeopardy the right itself".[8] Article 19(3) of the ICCPR illuminates the exceptional nature of limitations, recognizing that States may restrict expression under Article 19(2) only where provided by law and necessary to protect "the rights or reputations of others, national security or public order, or public health or morals." These are narrowly drawn exceptions,[9] and the burden falls on the authority restricting speech to justify the restriction, not on the speaker to demonstrate she has the right.[10] Any limitations must meet three conditions:

(a) *Legality*: provided by law that is precise, public, and transparent, avoids providing authorities with unbounded discretion, and gives appropriate notice to those whose speech is regulated. Rules should be subject to public comment and regular legislative or administrative process. Procedural safeguards, especially those guaranteed by independent courts or tribunals, should protect rights.

(b) *Legitimacy:* justified to protect one or more of the interests specified in Article 19(3) – rights or reputations of others; national security; public order; public health or morals.

(c) *Necessity and proportionality:* demonstrated by the State as necessary to protect the legitimate interest and the least restrictive means to achieve the purported aim. The Human Rights Committee has referred to these as "strict tests" according to which restrictions "must be applied only for those purposes for which they were prescribed and must be directly related to the specific need on which they are predicated."[11]

7. States regularly assert proper purposes for limitations on expression but fail to demonstrate that their limitations meet the tests of legality or necessity and proportionality.[12] For this reason, the rules are to be applied strictly and in good faith, with robust and transparent oversight available. Article 2 of the ICCPR obligates States to ensure that individuals seeking remedy for an ICCPR violation have that right "determined by competent judicial, administrative or legislative authorities, or by any other competent authority provided for by the legal system of the State".[13]

---

[5] See, e.g., Convention on the Elimination of All Forms of Racial Discrimination (CERD), Article 5; Convention on the Rights of the Child (CRC), Article 13; Convention on the Rights of Persons with Disabilities (CRPD), Article 21; International Convention on the Protection of Rights of All Migrant Workers (ICPMW), Article 13; American Convention on Human Rights, Article 13; African Charter on Human and People's Rights, Article 9; European Convention on Human Rights, Article 10.

[6] General Comment 34, CCPR/C/GC/34, para. 2 (hereinafter "General Comment 34").

[7] Id., para. 4. See also para. 5-6.

[8] Id., para. 21.The Human Rights Committee clarified that "restrictions must not impair the essence of the right." CCPR/C/21/Rev.1/Add.9, para. 13, adding that "The laws authorizing the application of restrictions should use precise criteria and may not confer unfettered discretion on those charged with their execution." Id.

[9] See especially A/67/357, para. 41; A/HRC/29/32, paras. 32 – 35.

[10] General Comment 34, para. 27.

[11] Id., para. 22

[12] See A/71/373 (2016).

[13] Article 2(3)(b), ICCPR. See also A/HRC/22/17/Add.4, para. 31.

‘Advocacy of hatred that constitutes incitement’

8. Article 20(2) of the ICCPR obligates States parties to prohibit by law “[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence”. It does not obligate States to *criminalize* Article 20(2) expression. The previous Special Rapporteur explained that Article 20(2) relates to (1) “advocacy of hatred,” (2) “advocacy which constitutes incitement,” and (3) incitement likely to result in discrimination, hostility or violence.[14]

9. UN human rights standards offer broader protection against discrimination than Article 20(2)’s attention to national, racial or religious hatred. Article 2 of the ICCPR guarantees rights “without distinction of any kind”, and Article 26 expressly provides that “the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground”. International standards ensure protections against adverse actions on grounds of race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, including indigenous origin or identity, disability, migrant or refugee status, sexual orientation, gender identity or intersex status.[15] The category expands with time, such that other categories, such as age or albinism, also constitute explicit protections today. Given the expansion of protection worldwide, the prohibition of incitement should be understood to apply to the broader categories now covered in international human rights law.

10. A critical definitional point: the individual whose expression is to be prohibited under Article 20(2) is *the advocate whose advocacy constitutes incitement*. A person who is not advocating hatred that constitutes incitement to discrimination, hostility or violence – for instance, a person advocating a minority or even offensive interpretation of a religious tenet or historical event, or a person sharing examples of hatred and incitement to report or raise awareness – is not to be silenced under Article 20 (or any other provision of human rights law). Such expression is to be *protected* by the State, even if the State disagrees with or is offended by the expression.[16] There is no ‘heckler’s veto’ in international human rights law.[17]

11. The International Convention on the Elimination of Racial Discrimination (ICERD), adopted the year before the ICCPR, calls on States to “eradicate all incitement to, and acts of” racial discrimination, “with due regard” to other rights protected by human rights law, including the freedom of expression.[18] Article 4 obligates States Parties, *inter alia*, to:

> “declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin”; and
>
> “declare illegal and prohibit organizations, and also organized and all other propaganda activities, which promote and incite racial discrimination, and shall recognize participation in such organizations or activities as an offence punishable by law”.

12. Article 20(2) of the ICCPR and Article 4 of the ICERD address specific categories of expression, often characterized as ‘hate speech’.[19] The language of these provisions has a

[14] A/67/357, para. 44

[15] Article 2(1), Article 26, ICCPR. Also see Article 19, ‘Hate speech’ explained: A toolkit (2015), at page 14. On online violence against women, see A/HRC/38/47 (2018).

[16] General Comment 34, para. 11

[17] See Evelyn M. Aswad, *To ban or not to ban blasphemous videos*, 4 Georgetown Journal of International Law 1313, 1320-22 (2013).

[18] ICERD, Articles 4 – 5.

[19] See Jeremy Waldron, *The Harm in Hate Speech* (2012).

kind of ambiguity compared to Article 19(2).[20] Whereas the freedom of expression in Article 19(2) of the ICCPR involves capacious rights embodied by active verbs (seek, receive, impart) and the broadest possible scope (ideas of *all* kinds, *regardless* of frontiers, through *any* media), the proscriptions under Article 20(2) and Article 4, while much narrower than generic 'hate speech' prohibitions, involve difficult-to-define language of emotion (hatred, hostility) and highly context-specific prohibition (advocacy of incitement). To be sure, the Human Rights Committee has concluded that Articles 19 and 20 "are compatible with and complement each other".[21] Even so, they demand interpretation.

13. The Human Rights Committee found in General Comment 34, in 2011, that whenever a State limits expression, including Article 20(2) expression, it must still "justify the prohibitions and their provisions in strict conformity with article 19."[22] A year later, a high-level group of human rights experts, convened under the auspices of the UN High Commissioner for Human Rights, adopted an interpretation of Article 20(2).[23] The Rabat Plan of Action defines key terms as follows:

> "'**hatred'** and '**hostility'** refer to intense and irrational emotions of opprobrium, enmity and detestation towards the target group; the term '**advocacy'** is to be understood as requiring an intention to promote hatred publicly towards the target group; and the term '**incitement'** refers to statements about national, racial or religious groups which create an *imminent risk of discrimination, hostility or violence* against persons belonging to those groups."[24]

14. The Rabat Plan of Action also identifies six factors to identify the severity necessary to criminalize incitement:

(a) The "social and political **context** prevalent at the time the speech was made and disseminated";

(b) The status of the **speaker**, "specifically the individual's or organization's standing in the context of the audience to whom the speech is directed";

(c) **Intent**, such that "[n]egligence and recklessness are not sufficient for an offense under article 20," which provides that "mere distribution or circulation" does not amount to advocacy or incitement.

(d) **Content and form** of the speech, in particular "the degree to which the speech was provocative and direct, as well as the form, style, nature of arguments deployed";

(e) **Extent or reach of the speech act**, such as the "magnitude and size of its audience", including whether it was "a single leaflet or broadcast in the mainstream media or via the internet, the frequency, the quantity and the extent of the communication, whether the audience had the means to act on the incitement . . ."; and

(f) Its **likelihood, including imminence**, such that "some degree of risk of harm must be identified," including through determination (by courts, it suggests) of a "reasonable probability that the speech would succeed in inciting actual action against the target group".[25]

---

[20] The ambiguity is not surprising, considering the negotiating history. See Jacob Mchangama, *The Sordid Origin of Hate-Speech Laws,* Policy Review (2011).

[21] General Comment 34, para. 50.

[22] Id., para, 52. In the context of Article 20(2) in particular, see id., para. 50.

[23] See, e.g., CERD Committee, General Recommendation 35, CERD/C/GC/35 (2013).

[24] Rabat Plan of Action, A/HRC/22/17/Add.4, footnote 5 (emphasis added), *citing* ARTICLE 19, *The Camden Principles on Freedom of Expression and Equality* (April 2009). Special Rapporteur Frank La Rue defined as a key factor in the assessment of incitement whether there was "real and imminent danger of violence resulting from the expression." A/67/357, para. 46. See also ARTICLE 19, *Prohibiting incitement to discrimination, hostility or violence* (2012), p. 24-25.

[25] Rabat Plan of Action, para. 29.

15. In 2013, the Committee on the Elimination of Racial Discrimination ("CERD Committee"), the expert monitoring body for the ICERD, followed the lead of the Human Rights Committee and the Rabat Plan of Action. It clarified the "due regard" language of ICERD Article 4 to require strict compliance with freedom of expression guarantees.[26] In a sign of converging interpretations, the CERD Committee emphasized that criminalisation under Article 4:

> "should be reserved for serious cases, to be proven beyond reasonable doubt, while less serious cases should be addressed by means other than criminal law, taking into account, inter alia, the nature and extent of the impact on targeted persons and groups. The application of criminal sanctions should be governed by principles of legality, proportionality and necessity."[27]

16. The CERD Committee explained that ICCPR Article 19's conditions also apply to restrictions under ICERD Article 4.[28] Dissemination and incitement, the CERD Committee found, require that States take into account a range of factors in determining whether particular expression falls into a prohibited category, including the speech's "content and form", the "economic, social and political climate" at issue during the time of the expression, the "position or status of the speaker," the "reach of the speech," and its objectives.[29] The Committee recommended consideration of "the imminent risk of likelihood that the conduct desired or intended by the speaker will result from the speech in questions."

17. The CERD Committee also found that the ICERD requires the prohibition of "insults, ridicule or slander of persons or groups or justification of hatred, contempt or discrimination", emphasizing that such expression may only be prohibited where it "clearly amounts to incitement to hatred or discrimination."[30] The language of ridicule and justification are extremely broad and generally precluded from restriction under international human rights law, which protects the right to offend and mock. Thus, the tie to incitement, and to the Article 19(3) framework, helps constrain such a prohibition to the most serious category.

18. The Rabat Plan of Action also clarifies that criminalization should be left for the most serious sorts of incitement under Article 20(2), and generally other approaches deserve consideration first.[31] These approaches include public statements by leaders in society that counter hate speech and foster tolerance and inter-community respect; education and inter-cultural dialogue; expanding access to information and ideas of a range of kinds that counter hateful messages; promotion and training in human rights principles and standards. The recognition of steps other than legal prohibitions highlights that prohibition will often not be the least restrictive measure available to States confronting hate speech problems.

Hateful expression that may not constitute advocacy or incitement

19. Other kinds of speech may not meet the Article 20(2) or Article 4 definitions or thresholds but involve, for example, advocacy of hatred. The question arises: May States restrict 'advocacy of hatred' that does not constitute incitement to discrimination, hostility, or violence? May they restrict 'hate speech' when defined, as a UN panel did recently, as speech "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,

[26] General Recommendation 35, para. 35. The CERD Committee understands the due-regard clause as having particular importance with regard to freedom of expression, which it states is 'the most pertinent principle when calibrating the legitimacy of speech restrictions.' Id.

[27] Id., para. 12.

[28] Id., para. 4, 19-20.

[29] Id., paras. 15 – 16.

[30] Id., para. 13.

[31] Id., para. 35

ethnicity, nationality, race, colour, descent, gender or other identity factor."[32] Clearly this is language short of Article 20(2) and Article 4's incitement, and while States and companies should combat such attitudes with education, condemnation and other tools, legal restrictions will need to meet the strict standards of international human rights law.

20. For content that involves the kind of speech as defined by the UN panel, hateful but not constituting incitement, Article 19(3) of the ICCPR provides appropriate guidance. Its conditions must be applied strictly, such that any restriction – and any action taken against speech – meets the conditions of legality, necessity and proportionality, and legitimacy. Language like the UN Panel's above, if meant to guide prohibitions under law, would be problematic on legality grounds, given its vagueness – though it may serve as a basis for political and social action to counter discrimination and hatred. Any State adopting such a definition would also need to situate a restriction among the legitimate grounds for limitation. It most instances, the rights of others may provide the appropriate basis, focused on rights related to discrimination or interference with privacy, or protecting public order. But in each case, it would remain essential for the State to demonstrate the necessity and proportionality of taking action, and the harsher the penalty, the greater the need for demonstrating strict necessity.[33]

21. Some restrictions are specifically disfavoured under international human rights standards. For instance, the Human Rights Committee noted that "[p]rohibitions of displays of lack of respect for a religion or other belief system, including blasphemy laws, are incompatible with the Covenant, except" in the context where blasphemy also may be defined as advocacy of religious hatred that constitutes incitement of one of the required sorts.[34] To be clear, anti-blasphemy laws fail to meet the legitimacy condition of Article 19(3), given that Article 19 protects *individuals* and their rights to freedom of expression and opinion; neither it nor Article 18 of the ICCPR protect ideas or beliefs from ridicule, abuse, criticism or other 'attacks' seen as offensive. Several human rights mechanisms have affirmed the call to repeal blasphemy laws because of the risk they pose to debate over religious ideas and the role that such laws play in enabling government to preference one religion's ideas over other religions, beliefs, or non-belief systems.[35]

22. Second, laws that "penalize[] the expression of opinions about historical facts are incompatible" with Article 19, calling into question the criminalization of Holocaust and other atrocity denial and similar laws, which are often justified by reference to 'hate speech'. Opinions that are "erroneous" and "incorrect interpretations of past events" may not be subject to general prohibition, the Human Rights Committee noted, and any restrictions on the expression of such opinion "should not go beyond what is permitted" in Article 19(3) or "required under Article 20".[36] In light of these and other interpretations, denial of the historical accuracy of atrocities should not be subject to criminal penalty or other restriction without further evaluation under the definitions and context noted above. The application of any such restriction under international human rights law should involve evaluation of the six factors noted in the Rabat Plan of Action.

23. A third kind of non-incitement speech may involve a situation where a speaker is "individually targeting an identifiable victim" but not seeking to "incite others to take an

---

[32] *United Nations Strategy and Plan of Action on Hate Speech*, May 2019. The Rabat Plan of Action alludes to speech that is below Article 20(2) thresholds but either "may justify a civil suit or administrative sanctions" or, giving rise to no sanctions, "still raises concern in terms of tolerance, civility and respect for the rights of others". Rabat Plan of Action, para. 20

[33] Article 19(3)'s public morals exception would be an unlikely basis, but it bears noting that the Human Rights Committee has clarified that "'the purpose of protecting morals must be based on principles not deriving exclusively from a single tradition'". General Comment 34, para. 32

[34] Id., para. 48. In this case, the blasphemy would be beside the point; only the advocacy constituting incitement would be relevant.

[35] See especially A/HRC/31/18 (2015), paras. 59 – 61.

[36] General Comment 34, para. 49. See Sarah Cleveland, *Hate Speech at Home and Abroad*, in *Free Speech Century* 226 (Bollinger & Stone, eds.) (2019). See also A/67/357, para. 55.

action against persons on the basis of a protected characteristic".[37] Again, by reference to Article 19(3), such speech may be subject to restriction in order to protect the rights of others or public order. Often States restrict such expression under the general rubric of 'hate crimes' – whereby the penalty for a physical attack on a person or property is exacerbated by the hateful motivation behind it.

24. Fourth, it is important to emphasize that expression that may be offensive or characterized by prejudice, and raise serious concerns of intolerance, may often not meet a threshold of severity to merit any kind of restriction. There is a range of expression of hatred, ugly as it is, that does not involve incitement or direct threat, such as declarations of prejudice against protected groups. Such sentiments would not be subject to prohibition under the ICCPR or ICERD, and other restrictions or adverse actions would require analysis of Article 19(3)'s conditions. The six factors identified by the Rabat Plan of Action for criminalizing incitement will also provide a valuable rubric for considering how to evaluate public authorities' reactions to such speech. Indeed, the absence of restriction does not mean the absence of action; States may (and should, consistent with Human Rights Council Resolution 16/18) take robust steps – government condemnation of prejudice, education, training, public service announcements, community projects, etc. – to counter such intolerance and ensure that public authorities protect individuals against discrimination rooted in these kinds of assertions of hate.

25. Finally, the Genocide Convention requires States to criminalize incitement to genocide. In some situations, such as Myanmar, State inaction against incitement to genocide may contribute to very serious consequences for vulnerable communities. Such inaction itself is condemnable, just as the incitement itself must be opposed and punished.[38]

Human rights norms at the regional level

26. European, Inter-American and African human rights systems also articulate standards related to 'hate speech'. The European Court of Human Rights (ECtHR) has emphasized that freedom of expression protects the kinds of speech that may "offend, shock or disturb".[39] Yet the Court has adopted relatively deferential attitudes toward States that continue to ban blasphemy by law on the grounds of prohibiting 'hate speech' or criminalize genocide denial, in contrast to trends observed at the global level.[40] Often the Court avoids the 'hate speech' question altogether, relying not on freedom of expression but 'abuse of rights' grounds to find claims of violation inadmissible.[41] European norms may be in flux when it comes to imposing liability on intermediaries for hate speech on their platforms.[42] By contrast, standards in the Inter-American Commission for Human Rights have tended to run closer to the international standards elucidated above, while standards in the African system are at a comparatively early stage.[43] Regional human rights norms

[37] ARTICLE 19, *Hate Speech Explained: A Toolkit* (2015), page 22.

[38] See especially A/HRC/39/64 (2018), at para. 73. The Genocide Convention calls for the criminalization of "direct and public incitement to commit genocide". See Convention on the Prevention and Punishment of the Crime of Genocide, Article III (c).

[39] ECtHR, *Handyside vs United Kingdom*, 7 December 1976, para. 49. See Sejal Parmer, *The Legal Framework for Addressing 'Hate Speech' in Europe*, 6-7 November 2018 (Council of Europe Conference: Addressing Hate Speech in the Media, Zagreb, Croatia).

[40] See Council of Europe, Fact Sheet on Hate Speech (March 2019); Evelyn M. Aswad, *The Future of Freedom of Expression Online*, 17 Duke Law and Technology Review 26, 44 (2018).

[41] For an overview of practice, see Council of Europe, Guide on Article 17 of the European Convention on Human Rights (updated 31 August 2019).

[42] Compare Delfi AS v Estonia, application no. 64569/09, ECtHR, [GC] judgment 16 June 2015 with Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, application no. 22947/13, ECtHR, judgment 2 February 20116. See also ARTICLE 19, *Responding to 'hate speech': Comparative overview of six EU countries* (2018).

[43] See Inter-American Commission on Human Rights, *Hate Speech and Incitement to Violence Against Lesbian, Gay, Bisexual, Trans and Intersex Person in the Americas* (12 November 2015).

cannot, in any event, be invoked to justify departure from international human rights protections.

27. The Human Rights Committee has specifically rejected the European Court's margin of appreciation doctrine, noting that "a State party, in any given case, must demonstrate in specific fashion the precise nature of the threat to any of the enumerated grounds listed in paragraph 3 that has caused it to restrict freedom of expression."[44] The Human Rights Committee does not grant discretion to the State simply because the domestic authorities assert that they generally are better placed to understand their local context.

Summary of UN instruments on hate speech

28. The international human rights framework has evolved in recent years to rationalize what appear, on the surface, to be competing norms. In short, the freedom of expression is a legal right of paramount value for democratic societies, inter-dependent with and supportive of other rights throughout the corpus of human rights law. At the same time, anti-discrimination, equality, and equal and effective public participation underpin the entire corpus of human rights law. The kind of expression captured by Article 20 of the ICCPR and Article 4 of the CERD presents challenges to both sets of norms, something that all participants in public life must acknowledge. Thus, restrictions on the right to freedom of expression must be exceptional, and the State bears the burden of demonstrating their consistency with international law; prohibitions under ICCPR Article 20 and ICERD Article 4 must be subject to strict and narrow conditions of Article 19(3); and States should generally deploy tools at their disposal other than criminalization and prohibition – such as education, counter-speech, promotion of pluralism, and so forth – to address all kinds of 'hate speech'.

III. Governing online hate speech

A. State obligations and the regulation of online hate speech

29. Strict adherence to international human rights law standards protects against governmental excesses. As a first principle, States should not use internet companies as tools to limit expression that they themselves would be precluded from limiting under international human rights law. What they demand of companies, whether through regulation or threats of regulation, must be justified under and in compliance with international law. Certain kinds of action against content are clearly inconsistent with Article 19(3) – such as internet shutdowns and the criminalization of political dissent or government criticism online.[45] Penalties on individuals for engaging in unlawful hate speech should not be enhanced merely for being online.

30. It is useful to contemplate a hypothetical State that is considering legislation that would hold online intermediaries liable for failure to take specified action against 'hate speech'. Such an 'intermediary liability' law is typically aimed at restricting expression, whether of the users of a particular platform or of the platform itself, possibly aimed at fulfilling the obligation under Article 20(2). Any legal evaluation of such a proposal must address the cumulative conditions of Article 19(3) to be consistent with international free expression standards.[46]

Legality

31. Article 19(3) requires that the imposition of liability for the hosting of 'hate speech' define the phrase and the factors involved in identifying its instances. A proposal imposing liability for a failure to remove 'incitement' must define the content consistent with Article

[44] General Comment 34, para. 36.

[45] See A/HRC/35/22.

[46] For a statement of the principles that should apply in the context of intermediary liability, see Manila Principles on Intermediary Liability (2015).

20(2) of the ICCPR and the ICERD Article 4, including by defining the key terms noted above in the Rabat Plan of Action. If a State wishes to regulate hate speech on grounds other than Articles 20 and 4, it must define the content that is in fact unlawful[47]; precision and clarity mean that State laws should constrain excessive discretion in government actors to enforce the rules or private actors to use them to suppress lawful expression and must give individuals appropriate notice to regulate their affairs.[48] Without clarity and precision in the definitions, there is significant risk of abuse, restriction of legitimate content, and failure to address the problems at issue. States addressing 'hate speech' should tie their definitions closely to the standards of international human rights law, such as Article 20(2).

32. Several States have adopted or are considering adopting rules that require internet companies to remove "manifestly unlawful" speech within a particular period, typically twenty-four hours or even as brief as one hour, or otherwise unlawful content in a lengthier period. The most well-known of these laws, Germany's Network Enforcement Act (widely known as NetzDG), imposes requirements on companies to remove from their platforms speech that is unlawful under a number of specifically identified provisions of the German Criminal Code.[49] For instance, Section 130 of the Criminal Code provides, inter alia, for the sanction of a person who "in a manner capable of disturbing the public peace . . . incites hatred against a national, racial, religious group or a group defined by their ethnic origins, against segments of the population or individuals because of their belonging to one of the aforementioned groups or segments of the population or calls for violent or arbitrary measures against them."[50] The law evidently does not define its key terms (especially "incite" and "hatred")[51] and yet, through NetzDG, it imposes significant fines on companies that fail to adhere to them. The underlying law is problematically vague. While NetzDG should be understood as a good faith effort to deal with widespread concern over online hate and its offline consequences, the failure to define these key terms undermines the claim that its requirements are consistent with international human rights law.

33. Few States have involved their courts in the process of evaluating platform hate speech that is inconsistent with local law, but they should allow for the imposition of liability only according to orders by independent courts and with the possibility of appeal at the request of the intermediary or other party affected by the action (such as the subject user).[52] Governments have been increasing the pressure on companies to serve as the adjudicators of hate speech. The process of adoption should also be subject to rigorous rule of law standards, with adequate opportunity for public input and hearings and evaluation of alternatives and impact on human rights.[53]

---

[47] States have largely distinguished terrorist and 'extremist' content from 'hate speech', but the same principles of legality must apply to those subjects as well. See, e.g., A/HRC/40/52, para. 75(e). The term 'extremism' often is deployed as a substitute for 'hate speech', albeit one that is not rooted in law. 'Violent extremism' does little to add clarity. Governments that use the phrase in good faith in an online context seem to focus on the problem of the virality of "terrorist and violent extremist ideologies" and to aim to counter "extremist" narratives and "prevent the abuse of the internet".

[48] This is not to preclude the possibility of civil claims that one individual may bring against another for traditional torts that take place online instead of offline. But defining the expression that may cause legally redressable harm is required under Article 19 of the ICCPR.

[49] Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), Section 1(3).

[50] German Criminal Code (last amended 24 September 2013), Section 130. Similar references are made in the proposed French law concerning online hate speech. See OL FRA 6/2109 (20 August 2019) and Response of the Government of France (23 August 2019).

[51] But see *BGH Urt. v. 3 April 2008* – 3 StR 394/07, BeckRS 2008, 06865.

[52] The previous Special Rapporteur noted that "any restriction must be applied by a body that is independent of political, commercial or other unwarranted influences in a manner that is neither arbitrary nor discriminatory, and with adequate safeguards against abuse". A/67/357 para. 42.

[53] See OL AUS 5/2019; and response from the Permanent Mission of Australia, 23 September 2019.

Necessity and proportionality

34. Legislative efforts to incentivize the removal of online hate, and impose liability on internet companies for a failure to do so, must meet the necessity and proportionality standards identified above. In recent years, States have pushed companies toward nearly immediate takedown of content, demanding that they develop filters that would disable the upload of content deemed harmful. The pressure is for automated tools that would serve as a form of pre-publication censorship. Problematically, an upload filter requirement "would enable the blocking of content without any form of due process even before it is published, reversing the well-established presumption that States, not individuals, bear the burden of justifying restrictions on freedom of expression".[54] Because such filters are notoriously unable to address the kind of natural language that typically constitutes hateful content, they can cause significant disproportionate outcomes.[55] What's more, there is research suggesting that such filters disproportionately harm historically under-represented communities.[56]

35. The push for upload filters for hate speech (and other kinds of content) is ill-advised, as it drives the platforms toward the regulation and removal of lawful content. They enhance the power of the companies with very little, if any, oversight or opportunity for redress. States should instead be pursuing laws and policies that push companies to protect free expression and counter lawfully restricted forms of hate speech through a combination of features: transparency requirements that allow public oversight; the enforcement of domestic law by independent judicial authorities; and other social and educational efforts along the lines proposed in the Rabat Plan of Action and Human Rights Council Resolution 16/18.

36. Some States have taken steps to address illegal hate speech through other creative and seemingly proportionate means. While India has problematically adopted internet shutdowns as a tool to deal with content issues in some instances, interfering disproportionately with the population's access to communications[57], some states in India adopted alternative approaches. One approach involved the creation of hotlines for individuals to report WhatsApp content to law enforcement authorities, while another involved the establishment of "social media labs" to monitor online hate speech. While these kinds of approaches require careful development to be consistent with human rights norms, they suggest a kind of "creative" and "out of the box" approach to address hate speech without outsourcing to distant companies the role of content police.[58]

37. In 2019, an official commission in France proposed an approach to the regulation of online content that would seem to protect expression while also giving room to address unlawful hate speech. While the status of the commission's work is unclear at the time of writing, its proposals involve judicial authorities addressing hate speech problems and multi-stakeholder initiatives to provide oversight of company policies. The commission concluded as follows:

> "Public intervention to force the biggest players to assume a more responsible and protective attitude to our social cohesion therefore appears legitimate. Given the civil liberty issues at stake, this intervention should be subject to particular precautions. It must (1) respect the wide range of social network models, which are

[54] OL OTH 71/2018. The European Commission recommendation on measures to effective tackle illegal content online, para. 36 (1 March 2018), calls for "proactive measures, including by using automated means, in order to detect, identify and expeditiously remove or disable access to terrorist content."

[55] See Center for Democracy and Technology, *Mixed Messages? The Limits of Automated Social Media Content Analysis* (2017).

[56] On the serious freedom of expression concerns raised by upload filters, see Daphne Keller, *Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Pieczek v. Facebook Opinion* (Stanford: 2019).

[57] See OL IND 7/2017; OL IND 5/2016; See also the press release *UN rights experts urge India to end communications shutdown in Kashmir, 22 August 2019.*

[58] Chinmayi Arun & Nakul Nayak, *Preliminary Findings on Online Hate Speech and the Law in India Defining Hate Speech*, Berkman-Klein Center Publication 2016-19, at 11.

particularly diverse, (2) impose a principle of transparency and systematic inclusion of civil society, (3) aim for a minimum level of intervention in accordance with the principles of necessity and proportionality and (4) refer to the courts for the characterisation of the lawfulness of individual content."[59]

38. This approach deserves further development and consideration, addressing issues of freedom of expression and "social cohesion" in ways that appear to enable respect for international human rights law.

Legitimacy

39. Government regulation of online intermediaries should be subject to the same guidelines for legitimacy as human rights law applied to all government restriction of speech. As noted above, certain kinds of speech that States may characterize as 'hate speech' should not be subject to prohibition under Article 20(2) or Article 19. In addition, legal terms that restrict incitement that, for instance, instigates 'hatred against the regime' or 'subversion of State power', are unlawful bases for restriction under Article 19(3).[60] Overly broad definitions of hate speech – for instance proscribing incitement of 'religious discord,' or speech that might subject a country to violent acts[61] – typically enable speech restrictions for illegitimate purposes, or in this case, demands on intermediaries that are inconsistent with human rights law.

B. Company content moderation and hate speech

40. It is on the platforms of internet companies where hateful content spreads online, seemingly spurred on by a business model that values attention and virality.[62] The largest deploy "classifiers" such that AI software can identify proscribed content, with perhaps only intermittent success, based on specific words and analysis. They operate across jurisdictions, and the same content in one location may cause a different impact elsewhere. Online hate speech often involves unknown speakers, with coordinated bot threats, disinformation and deep fakes, and mob attacks[63].

41. Internet companies shape their platforms' rules and public presentation (or brand).[64] They have massive impact on human rights, particularly but not only in places where they are the predominant form of public and private expression, where a limitation of speech can amount to public silencing or a failure to deal with incitement can facilitate offline violence and discrimination[65]. The consequences of ungoverned online hate can be tragic, as illuminated by Facebook's failure to address incitement against the Rohingya Muslim community in Myanmar. Companies do not have the obligations of governments, but their impact is of a sort that requires them to assess the same kind of questions about protecting their users' rights to freedom of expression.[66]

---

[59] *Creating a French framework to make social media platforms more accountable: Acting in France with a European vision*, Interim Mission Report Submitted to the French Secretary of State for Digital Affairs, May 2019.

[60] See A/67/357, paras. 51 – 55.

[61] See OL JOR 3/2018.

[62] See Tim Wu, *The Attention Merchants* (2016).

[63] See Gayathry Venkiteswaran, *'Let the mob do the job': How proponents of hatred are threatening freedom of expression and religion online in Asia*, Association for Progressive Communications (2017).

[64] See Kate Klonick, *The New Governors: The People, Rules and Processes Governing Online Speech*, 131 Harvard Law Review 1598 (2018); David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (2019).

[65] See A/HRC/42/50, paras. 70-75

[66] A/HRC/32/38 paras. 87 – 88. See also Business for Social Responsibility and World Economic Forum, *White Paper: Responsible Use of Technology* (August 2019).

42. Previous reports have argued that all companies in the ICT sector should be applying the UN Guiding Principles on Business and Human Rights and establish human rights by design and by default. Yet companies manage 'hate speech' on their platforms almost entirely without reference to the human rights implications of their products.[67] This is a mistake, depriving the companies of a framework for making rights-compliant decisions and articulating their enforcement to governments and individuals, while hobbling the public's capacity to make claims using a globally understood vocabulary. This report reiterates the call for company human rights policies that involve mechanisms to:

(a) conduct periodic review of their impact on human rights,

(b) avoid adverse human rights impacts and prevent or mitigate those that do arise, and

(c) implement due diligence processes to "identify, prevent, mitigate and account for how they address human rights impacts" and have a process for remediating harm.[68]

43. There will always be difficult questions about how to apply UN human rights standards to a wide range of content (just as there are difficult questions about national laws and regional human rights law).[69] However, the guidance above can help shape company protection of rights at each stage of the moderation of content: product development, definition, identification, action, and remedy. Global norms provide a firm basis for companies with global users communicating across borders, and they are called for by the UN Guiding Principles (Principle 12).[70]

Human Rights Due Diligence and Review

44. Dealing with hate speech should have started with due diligence at the product development stage. Unfortunately, it seems likely that few if any major internet companies have conducted rights-oriented product review related to hate speech; if so, it has not been public. However, products in the ICT sector are under constant updating and revising, and thus it is critical for companies to conduct regular impact assessments and reassessments in order to determine how their products infringe upon the enjoyment of human rights. Under the UN Guiding Principles, businesses should (among other things) have an ongoing process to determine how hate speech impacts human rights on their platforms (Principle 17), including through a platform's own algorithms[71].They should draw on internal and independent human rights expertise, including "meaningful consultation with potentially affected groups and other relevant stakeholders" (Principle 18). They should regularly evaluate the effectiveness of their approaches to human rights harms (Principle 20).

45. The lack of transparency is a major flaw in all the companies' content moderation processes. There is a significant barrier to external review (academic, legal, other) of hate speech policies as required under Principle 21: while the rules are public, the details of their implementation, at aggregate and granular levels, are nearly non-existent. Finally, the companies must also train their content policy teams, general counsel, and especially content moderators in the field, those conducting the actual work of restriction. (Principle 16, commentary) The training should identify the norms of human rights law that their content moderation aims to protect and promote. In particular, companies should assess whether their hate speech rules infringe on freedom of expression by assessing the legality, necessity and legitimacy principles identified above.

---

[67] At the time of writing, Facebook had just released a revised statement of values indicating that it would "look to international human rights standards" to make certain judgments concerning community standards. See Facebook, *Updating the Values that Inform Our Community Standards*, 12 September 2019.

[68] UN Guiding Principles on Business and Human Rights (2011) no 12 with commentary, 13 and 15.

[69] See Susan Benesch, *Proposals for Improved Regulation of Harmful Online Content*, Paper for the Israel Democracy Institute (2019).

[70] See generally Business for Social Responsibility, *Human Rights Impact Assessment: Facebook in Myanmar* (2018).

[71] A/73/348.

The legality standard

46. Company definitions of hate speech are generally difficult to understand, though companies vary on this score. Some are non-existent, and others are vague. For instance, Russia's VK prohibits content that "propagandizes and/or contributes to racial, religious, ethnic hatred or hostility, propagandizes fascism or racial superiority" or "contains extremist materials". China's WeChat prohibits "content . . . which in fact or in our reasonable opinion . . . is hateful, harassing, abusive, racially or ethnically offensive, defamatory, humiliating to other people (publicly or otherwise), threatening, profane or otherwise objectionable." Others are dense and detailed, with serious efforts to spell out exactly the kind of content that constitutes hate speech subject to restriction, and yet the density paradoxically can create confusion and a lack of clarity. The policies of the three dominant American companies – YouTube, Facebook and Twitter – have evolved and improved over many years, each layering their policies in ways that have converged to a recognizably similar set of rules. However, while they use different terms to signal restriction of content that "promotes" violence or hatred against specific protected groups, they do not clarify how they define promotion, incitement, targeting groups, and so forth. Among other issues, subjects such as intent and result are difficult to identify in the policies.[72]

47. The companies should review their policies, or adopt new ones, with the legality test in mind. A human rights compliant framework on online hate speech would draw from the definitional guidance above and provide answers to the following:

(a) *What are protected persons/groups?* Human rights law has identified specific groups requiring express protection. Companies in the ICT sector should aim to apply the broadest possible protection in keeping with evolving laws and normative understandings. The companies should be clear that they would not restrict "promotion…of a positive sense of group identity" particularly in the context of historically disadvantaged groups (while acknowledging that some expressions of group identity, such as white supremacy, may in fact constitute hateful content).[73]

(b) *What kind of hate speech constitutes a violation of company rules?* Companies should develop 'hate speech' policies by considering what kinds of interference users may face on the platform. Human rights law provides guidance, particularly by noting the legitimacy of restrictions to protect the rights of others. For instance, companies could consider how hateful online expression can incite violence that threatens life, infringes on others' freedom of expression and access to information, interferes with privacy or the right to vote, and so forth. The companies are not in the position of governments to assess threats to national security and public order, and hate speech restrictions on those grounds should be based not on company assessment but legal orders from a State, themselves subject to Article 19(3)'s strict conditions.

(c) *Is there specific 'hate speech' content that the companies restrict?* Companies should indicate how they prohibit, if they do, the kind of expression covered by ICCPR Article 20(2) and ICERD Article 4. If so, they should draw from the instruments identified above in defining it. But incitement is only one part of the problematic content that may constitute 'hate speech'. Beyond it, companies should identify what that category includes, much as the evolving policies of some companies have done. They should do more than simply identify; they should also show, through the development of a kind of case law, exactly how their categories play out in the actual enforcement of the rules.[74]

(d) *Are there categories of users for whom the hate speech rules do not apply?* International standards are clear that journalists and others reporting on hate speech should be protected against content restrictions or account actions. Moreover, an application of the context standards of Rabat would lead to the protection of such content. Politicians, government and military officials, and other public figures are another matter. Given their

[72] A/HRC/38/35 para. 26.

[73] ARTICLE 19, Camden Principle 12.

[74] A/38/35, para. 71.

prominence and potential leadership role in inciting behaviour, they should be bound by the same hate speech rules that apply under international standards. In the context of hate speech policies, by default public figures should abide by the same rules as all users. Evaluation of context may lead to a decision of exception in some instances, where the content must be protected as, for instance, political speech. But incitement is almost certainly more harmful when uttered by leaders than by other users, and that factor should be part of the evaluation of platform content.

48. Where company rules vary from international standards, the companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation. For instance, were a company to decide to prohibit the use of a derogatory term to refer to a national or racial or religious group – which, on its own, would not be subject to restriction under human rights law – it should clarify its decision in accordance with human rights law. Moreover, companies should be especially alert to the abuse of their platforms through disinformation that constitutes hate speech; particularly in environments of rising tensions, the companies should clearly state their policies, develop comprehensive understanding through community and expert engagement, and firmly counter such incitement. International human rights standards can guide such policies, while the virality of hateful content in such contexts may require rapid reaction and early warning protect fundamental rights.

49. The companies should define how they identify when users violate the hate speech rules. Presently, it is difficult to know the circumstances under which the rules may be violated. There seems to be very significant inconsistency in the enforcement of rules. The opacity of enforcement is part of the problem. The Rabat Plan of Action identified a set of factors applicable to the criminalization of incitement under Article 20(2), but those factors should have weight in the context of company actions against speech as well. They need not be applied in the same way that they would in a criminal context. However, they offer a valuable framework for examining when the specifically defined content – the posts, the words or images that comprise the post – merits a restriction.

50. Companies may find the kind of detailed contextual analysis to be difficult and resource-intensive. The largest rely heavily on automation in order to do at least the first-layer work of identifying hate speech, and as a result, they require having rules that fall neatly into one box (ignore) or another (delete). They use the power of AI to drive these systems, but they are notoriously bad at evaluating context.[75] However, if the companies are serious about protecting human rights on their platforms, they must ensure that they define the rules clearly and require human evaluation. Human evaluation, moreover, must be more than an assessment of whether particular words fall into a particular box. It must be based on real learning from the communities where hate speech may be found, people who can understand the ‘code’ that language sometimes deploys to hide incitement to violence, evaluate the speaker’s intent, consider the nature of the speaker and her audience, evaluate the environment in which hate speech can lead to violent acts. None of these things is possible with AI alone, and the definitions and strategies should reflect the nuances of the problem. The largest companies should bear the burden of these resources and share their knowledge and tools widely, as open source, to ensure that smaller companies, and smaller markets, have access to such technology.

Necessity and proportionality

51. Companies have tools to deal with content in human rights-compliant ways, in some respects a broader range than enjoyed by States. Their range of options enables them to tailor their responses to specific problematic content, according to its severity and other factors. They can delete content, restrict its virality, label its origin, suspend the relevant user, suspend the organization sponsoring the content, develop ratings to highlight a person’s use of prohibited content, temporarily restrict content while a team is reviewing, demonetize, friction and warnings, broader blocking capacity, minimize its amplification, interfere with bots and coordinated online mob behaviour, adopt geolocated restrictions, and even promote counter-messaging. Not all of these tools are appropriate in every

[75] A/73/348.

circumstance, and they may require limitations themselves, but these show the range of options short of deletion that may be available to companies in given situations. In other words, just as States should evaluate whether a limitation on speech is the least restrictive approach, so too should companies carry out this kind of evaluation. And in carrying out the evaluation, companies should bear the burden of publicly demonstrating necessity and proportionality when so requested by affected users – whether the user is the speaker, the alleged victim, another person who came across the content, or a member of the public.

52. Evelyn Aswad identifies three steps that a company should undertake in the necessity framework: evaluate the tools it has available to protect a legitimate objective without interfering with the speech itself; identify the tool that least intrudes on speech; and assess and demonstrate that the measure it selects actually achieves its goals.[76] This kind of evaluation tracks the UN Guiding Principles' call for businesses to ensure that they prevent or mitigate harms, particularly because such an approach enables the companies to evaluate the two sets of potential harms involves – the restrictions on speech caused by implementation of their rules and the restrictions on speech caused by users deploying hate speech against other users or the public. An approach that draws from this framework enables the companies to determine how to respond not only to genuine incitement but also the kinds of expression that is common online – borderline hate speech and non-incitement.

Remedy

53. The mechanisms of international human rights law provide a wealth of ideas for remediation of online hate speech. The ICCPR and ICERD require the availability of remedies for violations of their provisions, as does the UN Guiding Principles.[77] The 2018 content moderation report highlighted company responsibility to remedy adverse human rights impacts under the UN Guiding Principles and so need not be repeated in detail here.[78] In short, the process of remediation must begin with an effective way for individuals to report potential violations of hate speech policies as well as protections against abuse of the reporting system as a form of hate speech. It should include a transparent and accessible process for appealing platform decisions, with companies providing a reasoned response that may also be publicly accessible.

54. At a minimum, the companies should publicly identify the kinds of remedies they will impose on violations of their hate speech policies. It may be that user suspension is insufficient. They should have graduated responses according to the severity of the violation or the recidivism of the user. They should develop strong products that protect user autonomy, security and free expression to remedy violations. Their approaches may involve de-amplification and de-monetization of problematic expression that they do not want to ban, for whatever reason, but they should, again, make the policies clear and known in advance to all users, based on accessible definitions, with warnings for all and the opportunity to withdraw and, if necessary, remedy the consequences of an offending comment. They may develop programs that require suspended users who wish to return to the platform to engage in kinds of reparations, like apology, or other forms of direct engagement with others they harmed. They should have remedial policies of education, counter-speech, reporting, and training. Remedy should also include, for the most serious lapses, post-violation impact assessments and the development of policies to end the violations.

55. The Rabat Plan of Action and Human Rights Council Resolution 16/18 also provide ideas that companies may draw upon in providing remedies for hateful content. The Rabat Plan of Action indicates, "States should ensure that persons who have suffered actual harm as a result of incitement to hatred have a right to an effective remedy, including a civil or non-judicial remedy for damages." Such remedies could include pecuniary damages, the

[76] Aswad, *The Future of Freedom of Expression Online*, at 47-52.

[77] Article 2, ICCPR; Article 6, ICERD.

[78] A/HRC/38/35, para. 59.

"right of correction" and "right of reply".[79] Resolution 16/18 identifies such tools as training of government officials and promoting the right of minority communities to manifest their belief.[80] The previous Special Rapporteur urged procedural remedies ("access to justice and ensuring effectiveness of domestic institutions") and substantive ones ("reparations that are adequate, prompt, and proportionate to the gravity of the expression, which may include restoring reputation, preventing recurrence and providing financial compensation.")[81] But he also urged a set of non-legal remedies, which the companies should evaluate and implement given their responsibility as creators of platforms on which hateful content thrives. Such remedial action could include educational efforts concerning the harms of hate speech and the way in which hate speech often aims to push vulnerable communities off the platforms (i.e., to silence them); mechanisms for responses to hate speech to be promoted and given greater visibility; public denunciation of hate speech, such as promoting public service announcements and statements of public figures; and stronger collaborations with social science researchers to evaluate the scope of the problem and the tools that are most effective against proliferation of hateful content.[82]

IV. Conclusions and Recommendations

56. **International human rights law should be understood as a critical framework for the protection and respect of human rights when combatting hateful, offensive, dangerous or disfavoured speech. Online 'hate speech', the broad category of expression described in this report, can result in deleterious outcomes. When the phrase is abused, it can provide ill-intentioned States with a tool to punish and restrict speech that is entirely legitimate and even necessary in rights-respecting societies. But some kinds of expression can cause real harm. It can intimidate vulnerable communities into silence, particularly when it involves advocacy of hatred that constitutes incitement to hostility, discrimination, or violence. Left unchecked and viral, it can create an environment that undermines public debate and can harm even those who are not users of the subject platform. It is thus important that States and companies address the problems of hate speech with a determination to protect those at risk of being silenced and to promote open and rigorous debate on even the most sensitive issues in the public interest.**

For States

57. **State approaches to online hate speech should begin with two premises: First, offline human rights protections must also apply to online speech. There should be no special category of online hate speech for which the penalties are higher than offline. Second, governments should not demand – through legal or extra-legal threats – that intermediaries take action that international human rights law would bar States from doing directly. In keeping with these foundations, and with reference to the rules outlined above, States should at a minimum do the following in addressing online hate speech:**

(a) Strictly define the terms in their laws that constitute prohibited content under Article 20(2) of the ICCPR and Article 4 of the ICERD and resist criminalizing such speech except in the gravest situations, such as advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, and adopt the interpretations of human rights law contained in the Rabat Plan of Action;

(b) Review existing laws or develop 'hate speech' legislation that meet the requirements of legality, necessity and proportionality, and legitimacy, and subject such rulemaking to robust public participation;

[79] Rabat Plan of Action, paras. 33-34.

[80] A/HRC/RES/16/18, paras. 5-6.

[81] A/67/357, para. 48.

[82] Id., paras. 56 – 74.

(c) Actively consider and deploy good governance measures, including those recommended in Human Rights Council Resolution 16/18 and the Rabat Plan of Action, to tackle hate speech with the aim of reducing the perceived need for bans on expression;

(d) Adopt or review intermediary liability rules to adhere strictly to human rights standards and do not demand that companies restrict expression that they would be unable to do directly, through legislation;

(e) Establish or strengthen independent judicial mechanisms to ensure that individuals may have access to justice and remedies when suffering cognizable Article 20(2) or Article 4 harms;

(f) Adopt laws that require companies to describe in detail and in public form how they define hate speech and enforce their rules against it, and to create databases of hate speech actions the companies take, and to otherwise encourage companies to respect human rights standards in their own rules; and

(g) Actively engage in international processes designed as learning forums for addressing hate speech.

For Companies

58. **Companies have for too long avoided human rights law as a guide to their rules and rulemaking, notwithstanding the extensive impacts they have on human rights of their users and the public. In addition to the principles adopted in earlier reports and in keeping with the UN Guiding Principles on Business and Human Rights, all companies in the ICT sector should:**

(a) Evaluate how their products and services impact the human rights of their users and the public, doing so through periodic and publicly available human rights impact assessments;

(b) Adopt content policies that tie their hate speech rules directly to international human rights law, indicating that the rules will be enforced according to the standards of international human rights law, including the relevant UN treaties and interpretations of the treaty bodies and Special Procedures and other experts, including the Rabat Plan of Action;

(c) Define the category of content they consider to be 'hate speech' with reasoned explanations for users and the public, and approaches that are consistent across jurisdictions;

(d) Ensure that any enforcement of hate speech rules involves an evaluation of context and the harm that the content imposes on users and the public, including by ensuring that any use of automation or Artificial Intelligence tools involve human-in-the-loop;

(e) Ensure that contextual analysis involves communities most affected by content identified as 'hate speech' and that communities are involved in identifying the most effective tools to address harms caused on the platforms; and

(f) As part of an overall effort to address 'hate speech,' develop tools that promote individual autonomy, security and free expression, and involve de-amplification, de-monetization, education, counter-speech, reporting, and training as alternatives, when appropriate, to account bans and content removals.