United Nations A/74/486

General Assembly

Distr.: General
9 October 2019

Original: English

Seventy-fourth session
Agenda item 70 (b)
Promotion and protection of human rights: human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms

Promotion and protection of the right to 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 the present report, the Special Rapporteur evaluates the human rights law that applies to the regulation of online "hate speech".

* The present report was submitted after the deadline so as to include the most recent information.

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

Summary

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. In the present report, submitted in accordance with Human Rights Council resolution 34/18, the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression explains how those standards provide a framework for Governments considering regulatory options and companies determining how to respect human rights online. The Special Rapporteur begins with an introduction to the international legal framework, focusing on United Nations treaties and the leading interpretations of provisions related to what is colloquially called “hate speech”. He then highlights key State obligations and addresses how content moderation by companies may ensure respect for the human rights of users and the public. He concludes with recommendations for States and companies.

Contents

		Page
I.	Introduction	4
II.	“Hate speech” regulation in international human rights law	4
III.	Governing online hate speech	12
	A. State obligations and the regulation of online hate speech	12
	B. Company content moderation and hate speech	16
IV.	Conclusions and recommendations	21

I. Introduction

1. “Hate speech”, a shorthand phrase that conventional international law does not define, has a double 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”, similar to the way in which they use “fake news”, to attack political enemies, non-believers, dissenters and critics. However, the phrase’s weakness (“it’s just speech”) also seems to inhibit Governments and companies from addressing genuine harms, such as the kind resulting from speech that incites violence or discrimination against the vulnerable or the silencing of the marginalized. The situation gives rise to frustration in 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 (A/HRC/38/35, para. 45).[1] In the present report, the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression explains how those standards provide a framework for Governments considering regulatory options and companies determining how to respect human rights online. The Special Rapporteur begins with an introduction to the international legal framework, focusing on United Nations treaties and the leading interpretations of provisions related to what is colloquially called “hate speech”. He then highlights key State obligations and addresses how content moderation by companies may ensure respect for the human rights of users and the public. He concludes with recommendations for States and companies.

3. The present report is the sixth in a series of reports published since 2015 in which the Special Rapporteur has addressed the human rights standards applicable to the freedom of opinion and expression in the information and communications technology (ICT) sector.[2] It should be read in the light of the standards and recommendations previously proposed, which are not necessarily repeated herein. As in his previous reports, the Special Rapporteur draws extensively from existing international standards and from considerable civil society input over the past several years.

II. “Hate speech” regulation 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 to allow open debate and individual autonomy and development with the also compelling obligation to prevent attacks on vulnerable communities and ensure the equal and non-discriminatory participation of all individuals in public life.[3] Governments often exploit the resulting uncertainty to threaten legitimate expression,

[1] The term “hate speech” is used in the present report to refer to obligations and limitations in human rights law in which that particular term is not used. See Susan Benesch, “Proposals for improved regulation of harmful online content”, paper prepared for the Israel Democracy Institute, 2019. Benesch coined a sibling term, “dangerous speech”, to identify a “capacity to catalyse violence by one group against another”. See also Susan Benesch, “Dangerous speech: a proposal to prevent group violence”, 2012.

[2] See A/HRC/29/32, on encryption and anonymity, A/HRC/32/38, on mapping the impact of the ICT sector on rights, A/HRC/35/22, on the digital access industry, A/HRC/38/35, on online content moderation, and A/73/348, on artificial intelligence and human rights.

[3] See, in particular, the report on hate speech of the previous Special Rapporteur Frank La Rue (A/67/357).

such as political dissent and criticism or religious disagreement.[4] However, the freedom of expression, 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 the enjoyment of all rights.

Freedom of expression

5. Article 19 (1) of the International Covenant on Civil and Political Rights protects the right to hold opinions without interference, and article 19 (2) guarantees the right to 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 regional, expressly protect the freedom of expression.[5] The Human Rights Committee, the expert monitoring body for the Covenant, 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". They "form a basis for the full enjoyment of a wide range of other human rights".[6]

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".[7] The exceptional nature of limitations is described in article 19 (3) of the Covenant, recognizing that States may restrict expression under article 19 (2) only where provided by law and necessary to respect the rights or reputations of others or protect national security, public order, public health or morals. These are narrowly defined exceptions (see, in particular, A/67/357, para. 41, and A/HRC/29/32, paras. 32–35), and the burden falls on the authority restricting speech to justify the restriction, not on the speakers to demonstrate that they have the right to such speech.[8] Any limitations must meet three conditions:

(a) **Legality**. The restriction must be provided by laws that are precise, public and transparent; it must avoid providing authorities with unbounded discretion, and appropriate notice must be given to those whose speech is being regulated. Rules should be subject to public comment and regular legislative or administrative processes. Procedural safeguards, especially those guaranteed by independent courts or tribunals, should protect rights;

(b) **Legitimacy**. The restriction should be justified to protect one or more of the interests specified in article 19 (3) of the Covenant, that is, to respect the rights

[4] Ibid., paras. 51–54.

[5] See, e.g., the International Convention on the Elimination of All Forms of Racial Discrimination, art. 5; the Convention on the Rights of the Child, art. 13; the Convention on the Rights of Persons with Disabilities, art. 21; the International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families, art. 13; the American Convention on Human Rights, art. 13; the African Charter on Human and People's Rights, art. 9; and the European Convention on Human Rights, art. 10.

[6] Human Rights Committee, general comment No. 34 (2011) on the freedoms of opinion and expression, paras. 2 and 4; see also ibid., paras. 5–6.

[7] Ibid., para. 21. The Human Rights Committee clarified that "restrictions must not impair the essence of the right", 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" (Human Rights Committee, general comment No. 27 (1999) on freedom of movement, para. 13).

[8] Human Rights Committee, general comment No. 34 (2011), para. 27.

or reputations of others or to protect national security, public order, public health or morals;

(c) **Necessity and proportionality**. The restriction must be demonstrated by the State as necessary to protect a legitimate interest and to be the least restrictive means to achieve the purported aim. The Human Rights Committee has referred to these conditions 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".[9]

7. States regularly assert proper purposes for imposing limitations on expression but fail to demonstrate that their limitations meet the tests of legality or necessity and proportionality (see A/71/373). For this reason, the rules are to be applied strictly and in good faith, with robust and transparent oversight. Under article 2 (3) (b) of the Covenant, States are obligated to ensure that individuals seeking remedy for a violation of the Covenant have their right thereto "determined by competent judicial, administrative or legislative authorities, or by any other competent authority provided for by the legal system of the State" (see also A/HRC/22/17/Add.4, para. 31).

Advocacy of hatred that constitutes incitement

8. Under article 20 (2) of the Covenant, States parties are obligated to prohibit by law "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence". States are not obligated to criminalize such kinds of expression. The previous Special Rapporteur explained that article 20 (2) relates to (a) advocacy of hatred, (b) advocacy which constitutes incitement, and (c) incitement likely to result in discrimination, hostility or violence (A/67/357, para. 43).

9. United Nations human rights standards offer broader protection against discrimination than that afforded through the focus in article 20 (2) on national, racial or religious hatred. Article 2 (1) of the Covenant guarantees rights to all individuals "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.[10] The scope of protection has expanded over time, such that other categories, such as age or albinism, are also now afforded explicit protection. Given the expansion of protection worldwide, the prohibition of incitement should be understood to apply to the broader categories now covered under international human rights law.

10. A critical point is that the individual whose expression is to be prohibited under article 20 (2) of the Covenant is the advocate whose advocacy constitutes incitement. A person who is not advocating hatred that constitutes incitement to discrimination, hostility or violence, for example, 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 on or raise awareness of the issue, is not to be silenced under article 20 (or any other provision of human rights law). Such expression is to

[9] Ibid., para. 22.

[10] See also, Article 19, *"Hate Speech" Explained: A Toolkit* (London, 2015), p. 14. On online violence against women, see A/HRC/38/47.

be protected by the State, even if the State disagrees with or is offended by the expression.[11] There is no "heckler's veto" in international human rights law.[12]

11. In the International Convention on the Elimination of All Forms of Racial Discrimination, adopted the year before the International Covenant on Civil and Political Rights, States are called upon to "eradicate all incitement to, or acts of" racial discrimination, with due regard to other rights protected by human rights law, including the freedom of expression (see articles 4 and 5 of the Convention). Under article 4 of the Convention, States parties are obligated, inter alia, to: (a) "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 (b) "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 International Covenant on Civil and Political Rights and article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination address specific categories of expression, often characterized as "hate speech".[13] The language in these provisions is ambiguous, compared with that of article 19 (2) of the Covenant.[14] Whereas the freedom of expression defined in article 19 (2) involves expansive 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) of the Covenant and article 4 of the Convention, 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). The Human Rights Committee has concluded that articles 19 and 20 of the Covenant "are compatible with and complement each other".[15] Even so, they require interpretation.

13. In its general comment No. 34 (2011), the Human Rights Committee found that whenever a State limits expression, including the kinds of expression defined in article 20 (2) of the Covenant, it must still "justify the prohibitions and their provisions in strict conformity with article 19".[16] In 2013 , a high-level group of human rights experts, convened under the auspices of the United Nations High Commissioner for Human Rights, adopted an interpretation of article 20 (2).[17] In the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, key terms are defined 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,

[11] Human Rights Committee, general comment No. 34 (2011), para. 11.

[12] See Evelyn M. Aswad, "To ban or not to ban blasphemous videos", *Georgetown Journal of International Law*, vol 44, No. 4 (2013).

[13] See Jeremy Waldron, *The Harm in Hate Speech* (Harvard University Press, 2012).

[14] The ambiguity is not surprising, considering the negotiating history. See Jacob Mchangama, "The sordid origin of hate-speech laws", *Policy Review* (December 2011 and January 2012).

[15] Human Rights Committee, general comment No. 34 (2011), para. 50.

[16] Ibid., para. 52, and, in the context of art. 20 (2) of the Covenant in particular, see para. 50.

[17] See, e.g., Committee on the Elimination of Racial Discrimination, general recommendation No. 35 (2013) on combating racist hate speech.

> hostility or violence against persons belonging to those groups (A/HRC/22/17/Add.4, appendix, footnote 5).[18]

14. A total of six factors were identified in the Rabat Plan of Action to determine the severity necessary to criminalize incitement (ibid, para. 29):

(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, meaning that "negligence and recklessness are not sufficient for an offence under article 20 of the Covenant", 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 communications, whether the audience had the means to act on the incitement";

(f) Its likelihood, including imminence, meaning that "some degree of risk of harm must be identified", including through the determination (by courts, as suggested in the Plan of Action) of a "reasonable probability that the speech would succeed in inciting actual action against the target group".

15. In 2013, the Committee on the Elimination of Racial Discrimination, the expert monitoring body for the International Convention on the Elimination of All Forms of Racial Discrimination, followed the lead of the Human Rights Committee and the Rabat Plan of Action. It clarified the "due regard" language in article 4 of the Convention as meaning that strict compliance with freedom of expression guarantees is required.[19] In a sign of converging interpretations, the Committee emphasized that criminalization under article 4 should be reserved for certain cases, as follows:

> The criminalization of forms of racist expression 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.[20]

16. The Committee on the Elimination of Racial Discrimination explained that the conditions defined in article 19 of the International Covenant on Civil and Political Rights also apply to restrictions under article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination.[21] With regard to the qualification

[18] The previous 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* (London, 2012), pp. 24–25.

[19] Committee on the Elimination of Racial Discrimination, general recommendation No. 35 (2013), para. 19. The Committee understands the due-regard clause as having particular importance with regard to freedom of expression, which, it states, is "the most pertinent reference principle when calibrating the legitimacy of speech restrictions".

[20] Committee on the Elimination of Racial Discrimination, general recommendation No. 35 (2013), para. 12.

[21] Ibid., paras. 4 and 19–20.

of dissemination and incitement as offences punishable by law, the Committee found that States must take into account a range of factors in determining whether a particular expression falls into those prohibited categories, including the speech's "content and form", the "economic, social and political climate" during the time the expression was made, the "position or status of the speaker", the "reach of the speech" and its objectives. The Committee recommended that States parties to the Convention consider "the imminent risk or likelihood that the conduct desired or intended by the speaker will result from the speech in question".[22]

17. The Committee also found that the Convention 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".[23] The terms "ridicule" and "justification" are extremely broad and are generally precluded from restriction under international human rights law, which protects the rights to offend and mock. Thus, the ties to incitement and to the framework established under article 19 (3) of the Covenant help to constrain such a prohibition to the most serious category.

18. In the Rabat Plan of Action, it is also clarified that criminalization should be left for the most serious sorts of incitement under article 20 (2) of the Covenant, and that, in general, other approaches deserve consideration first (A/HRC/22/17/Add.4, appendix, para. 34). These approaches include public statements by leaders in society that counter hate speech and foster tolerance and intercommunity respect; education and intercultural dialogue; expanding access to information and ideas that counter hateful messages; and the promotion of 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 as to whether States may restrict advocacy of hatred that does not constitute incitement to discrimination, hostility or violence. In other words, the question is whether States may restrict hate speech when defined, as was done recently in the United Nations Strategy and Plan of Action on Hate Speech, 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, ethnicity, nationality, race, colour, descent, gender or other identity factor".[24] Clearly such language stops short of the article 20 (2) and article 4 meanings of 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 in the United Nations Strategy on Hate Speech, that is, speech that is hateful but does not constitute incitement, article 19 (3) of the Covenant 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. Given its vagueness, language similar to that used in the Strategy, if meant

[22] Ibid., paras. 15–16.

[23] Ibid., para. 13.

[24] In the Rabat Plan of Action, reference is made to speech that is below the thresholds established under article 20 (2) of the Covenant but that either "may justify a civil suit or administrative sanctions" or gives rise to no sanctions but "still raises concern in terms of tolerance, civility and respect for the rights of others" (A/HRC/22/17/Add.4, para. 20).

to guide prohibitions under law, would be problematic on legality grounds, although 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. In most instances, the rights of others, as defined in article 19 (3), may provide the appropriate basis, focused on rights related to discrimination or interference with privacy, or protecting public order. However, 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.[25]

21. Some restrictions are specifically disfavoured under international human rights standards. As a first example, the Human Rights Committee noted that "prohibitions of displays of lack of respect for a religion or other belief system, including blasphemy laws, are incompatible with the Covenant", except in cases in which blasphemy also may be defined as advocacy of religious hatred that constitutes incitement of one of the required sorts.[26] To be clear, anti-blasphemy laws fail to meet the legitimacy condition of article 19 (3) of the Covenant, given that article 19 protects individuals and their right to freedom of expression and opinion; neither article 19 (3) nor article 18 of the Covenant 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 Governments to show preference for the ideas of one religion over those of other religions, beliefs or non-belief systems (see, in particular, A/HRC/31/18, paras. 59–61).

22. Second, laws that "penalize the expression of opinions about historical facts are incompatible" with article 19 of the Covenant, calling into question laws that criminalize the denial of the Holocaust and other atrocities and similar laws, which are often justified through references to hate speech. The Human Rights Committee noted that opinions that are "erroneous" and "an incorrect interpretation of past events" may not be subject to general prohibition, and any restrictions on the expression of such opinion "should not go beyond what is permitted" under article 19 (3) or "required under article 20" of the Covenant.[27] In the light of these and other interpretations, the denial of the historical accuracy of atrocities should not be subject to criminal penalty or other restrictions without further evaluation under the definitions and context noted above. The application of any such restriction under international human rights law should involve the evaluation of the six factors noted in the Rabat Plan of Action.

23. A third kind of non-incitement speech may involve a situation in which a speaker is "individually targeting an identifiable victim" but not seeking to "incite others to take an action against persons on the basis of a protected characteristic".[28] Again, in reference to article 19 (3) of the Covenant, such speech may be subject to restriction in order to protect the rights of others or to protect public order. Often States restrict such expression under the general rubric of "hate crimes", whereby the

[25] The public morals exception provided under article 19 (3) of the Covenant 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" (Human Rights Committee, general comment No. 34 (2011), para. 32, quoting Human Rights Committee, general comment No. 22 (1993) on the right to freedom of thought, conscience and religion, para. 8).

[26] Human Rights Committee, general comment No. 34 (2011), para. 48. In this case, the blasphemy would be beside the point; only the advocacy constituting incitement would be relevant.

[27] Human Rights Committee, general comment No. 34 (2011), para. 49. See Sarah Cleveland, *Hate Speech at Home and Abroad*, in Lee C. Bollinger and Geoffrey R. Stone, eds., *The Free Speech Century* (New York, Oxford University Press, 2019). See also A/67/357, para. 55.

[28] Article 19, *"Hate Speech" Explained*, p. 22.

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 that may 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 International Covenant on Civil and Political Rights or the International Convention on the Elimination of All Forms of Racial Discrimination, and other restrictions or adverse actions would require an analysis of the conditions provided under article 19 (3) of the Covenant. The six factors identified in the Rabat Plan of Action for criminalizing incitement 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, such as government condemnation of prejudice, education, training, public service announcements and community projects, to counter such intolerance and ensure that public authorities protect individuals against discrimination rooted in these kinds of assertions of hate.

25. Finally, the Convention on the Prevention and Punishment of the Crime of Genocide requires States to criminalize incitement to genocide. In some situations, such as in 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.[29]

Human rights norms at the regional level

26. Human rights systems in Europe, the Americas and Africa also articulate standards related to hate speech. The European Court of Human Rights has emphasized that freedom of expression protects the kinds of speech that may "offend, shock or disturb".[30] However, the Court has adopted relatively deferential attitudes towards States that continue to ban blasphemy by law on the grounds of prohibiting hate speech or continue to criminalize genocide denial, in contrast to trends observed at the global level.[31] Often the Court avoids the hate speech question altogether, relying not on freedom of expression but on "abuse of rights" grounds to find that claims of violation are inadmissible.[32] European norms may be in flux when it comes to making intermediaries liable for hate speech on their platforms.[33] By contrast, standards in the Inter-American Commission on Human Rights have tended to be

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

[30] European Court of Human Rights, *Handyside v. the United Kingdom*, application No. 5493/72, Judgment, 7 December 1976, para. 49. See Sejal Parmer, "The legal framework for addressing 'hate speech' in Europe", presentation at the international conference on addressing hate speech in the media, Zagreb, November 2018.

[31] See Council of Europe, "Hate speech", fact sheet, October 2019; and Evelyn M. Aswad, "The future of freedom of expression online", *Duke Law and Technology Review*, vol. 17 (August 2018).

[32] For an overview of practice, see Council of Europe, "Guide on article 17 of the European Convention on Human Rights: prohibition of abuse of rights", updated 31 August 2019.

[33] Compare European Court of Human Rights, Grand Chamber, *Delfi AS v. Estonia*, application No. 64569/09, Judgment, 16 June 2015, with European Court of Human Rights, Fourth Section, *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary*, application No. 22947/13, Judgment, 2 February 2016. See also Article 19, "Responding to 'hate speech': comparative overview of six EU countries", 2018.

similar to the international standards explained above, while standards in the African system are at a comparatively early stage.[34] Regional human rights norms 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".[35] The Committee does not grant discretion to the State simply because the national authorities assert that they generally are better placed to understand their local context.

Summary of United Nations 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, interdependent 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 in article 20 of the International Covenant on Civil and Political Rights and article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination 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 the consistency of such restrictions with international law; prohibitions under article 20 of the Covenant and article 4 of the Convention must be subject to the strict and narrow conditions established under article 19 (3) of the Covenant, and States should generally deploy tools at their disposal other than criminalization and prohibition, such as education, counter-speech and the promotion of pluralism, 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) of the International Covenant on Civil and Political Rights, such as Internet shutdowns and the criminalization of online political dissent or government criticism (see A/HRC/35/22). Penalties on individuals for engaging in unlawful hate speech should not be enhanced merely because the speech occurred online.

30. It is useful to contemplate a hypothetical State that is considering legislation that would hold online intermediaries liable for the failure to take specified action

[34] See Inter-American Commission on Human Rights, "Hate speech and incitement to violence", in Inter-American Commission on Human Rights, *Violence against Lesbian, Gay, Bisexual, Trans and Intersex Persons in the Americas* (2015).

[35] Human Rights Committee, general comment No. 34 (2011), para. 36.

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, sometimes with a view to fulfilling the obligation under article 20 (2) of the Covenant. Any legal evaluation of such a proposal must address the cumulative conditions established under article 19 (3) to ensure consistency with international standards on free expression.[36]

Legality

31. Article 19 (3) of the Covenant requires that, when imposing liability for the hosting of hate speech, the phrase itself and the factors involved in identifying the instances of hate speech must be defined. In a proposal to impose liability for a failure to remove "incitement", the content of such incitement must be defined consistent with article 20 (2) of the Covenant and article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination, including by defining the key terms in the Rabat Plan of Action noted above. If a State wishes to regulate hate speech on grounds other than those provided under article 20 of the Covenant and article 4 of the Convention, it must define the content that is in fact unlawful;[37] the precision and clarity required under article 19 (3) of the Covenant mean that State laws should constrain the excessive discretion of government actors to enforce the rules or of private actors to use the rules to suppress lawful expression and must provide for individuals to be given appropriate notice to regulate their affairs.[38] 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 those established under article 20 (2) of the Covenant.

32. Several States have adopted or are considering adopting rules that require Internet companies to remove "manifestly unlawful" speech within a particular period, typically within 24 hours or even as brief as 1 hour, or otherwise to remove unlawful content within a lengthier period. The most well-known of these laws, the Network Enforcement Act of Germany, 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.[39] For example, 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

[36] For a statement on the principles that should apply in the context of intermediary liability, see Electronic Frontier Foundation, "Manila principles on intermediary liability", 2015.

[37] 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" is often used as a substitute for "hate speech", albeit as a term that is not rooted in law. The term "violent extremism" does not add much clarity. Governments that use the term "extremism" in good faith in an online context seem to focus on the problem of the virality of "terrorist and violent extremist ideologies" and seem to have as their goal to counter "extremist" narratives and "prevent the abuse of the internet" (Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online).

[38] 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. However, defining the expression that may cause legally redressable harm is required under article 19 of the Covenant.

[39] Germany, Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act) (2017), sect. 1 (3).

them".[40] The law evidently does not define its key terms (especially "incite" and "hatred"),[41] and yet, through the Network Enforcement Act, it imposes significant fines on companies that fail to adhere to its provisions. The underlying law is problematically vague. While the Network Enforcement Act 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).[42] 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 of the impact on human rights.[43]

Necessity and proportionality

34. Legislative efforts to incentivize the removal of online hate speech and impose liability on Internet companies for the failure to do so must meet the necessity and proportionality standards identified above. In recent years, States have pushed companies towards a 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".[44] Because such filters are notoriously unable to address the kind of natural language that typically constitutes hateful content, they can cause significant disproportionate outcomes.[45] Furthermore, there is research suggesting that such filters disproportionately harm historically underrepresented communities.[46]

35. The push for upload filters for hate speech (and other kinds of content) is ill-advised, as it drives the platforms towards the regulation and removal of lawful

[40] Similar references are made in the French bill concerning online hate speech. See communication FRA 6/2019 and the response of the Government of France, available at https://spcommreports.ohchr.org/Tmsearch/TMDocuments.

[41] See, however, Germany, Federal Court of Justice, Judgment of 3 April 2008, Case No. 3 StR 394/07.

[42] The previous Special Rapporteur noted that "any restriction imposed 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).

[43] See communication AUS 5/2019 and the response from the Permanent Mission of Australia to the United Nations Office and other international organizations in Geneva, available at https://spcommreports.ohchr.org/Tmsearch/TMDocuments.

[44] Communication OTH 71/2018, available at https://spcommreports.ohchr.org/Tmsearch/TMDocuments. See also, European Commission, recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, para. 36, which calls for "proactive measures, including by using automated means, in order to detect, identify and expeditiously remove or disable access to terrorist content".

[45] See Center for Democracy and Technology, "Mixed messages? The limits of automated social media content analysis", 28 November 2017.

[46] Regarding the serious concerns about freedom of expression raised on the matter of upload filters, see Daphne Keller, "Dolphins in the Net: Internet content filters and the Advocate General's *Glawischnig-Pieczek v. Facebook Ireland* opinion", Stanford Center for Internet and Society, 4 September 2019.

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 national 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,[47] some states in India have 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 the role of content police to distant companies.[48]

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 was 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 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.[49]

38. This approach deserves further development and consideration, as it addresses 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 those contained in 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 articles 19 or 20 (2) of the Covenant. In addition, legal terms that restrict incitement that, for

[47] See communications IND 7/2017 and IND 5/2016, available at https://spcommreports.ohchr.org/Tmsearch/TMDocuments; see also Office of the United Nations High Commissioner for Human Rights, "United Nations rights experts urge India to end communications shutdown in Kashmir", press release, 22 August 2019.

[48] Chinmayi Arun and Nakul Nayak, "Preliminary findings on online hate speech and the law in India", 8 December 2016, p. 11.

[49] France, "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.

example, instigates “hatred against the regime” or “subversion of State power” are unlawful bases for restriction under article 19 (3) of the Covenant (A/67/357, paras. 51–55). Overly broad definitions of hate speech, for example proscribing incitement of “religious discord” or speech that might subject a country to violent acts,[50] typically enable speech restrictions for illegitimate purposes, or, in the case of government regulation of online intermediaries, demands on those 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.[51] The largest companies deploy “classifiers”, using artificial intelligence software to identify proscribed content, with perhaps only intermittent success, on the basis of specific words and analysis. The companies operate across jurisdictions, and the same content in one location may have a different impact elsewhere. Online hate speech often involves unknown speakers, with coordinated bot threats, disinformation and so-called deep fakes, and mob attacks.[52]

41. Internet companies shape their platforms’ rules and public presentation (or brand).[53] They have an enormous 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 (A/HRC/42/50, paras. 70–75). 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’ right to freedom of expression.[54]

42. In previous reports, it has been argued that all companies in the ICT sector should apply the Guiding Principles on Business and Human Rights of the United Nations and integrate human rights into their products by design and by default. However, companies manage hate speech on their platforms almost entirely without reference to the human rights implications of their products.[55] This is a mistake, as it deprives 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. The Special

[50] See communication JOR 3/2018, available at https://spcommreports.ohchr.org/Tmsearch/TMDocuments.

[51] See Tim Wu, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (New York, Vintage Books, 2016).

[52] 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, October 2017).

[53] See Kate Klonick, “The new governors: the people, rules, and processes governing online speech”, *Harvard Law Review*, vol. 131, No. 6 (April 2018); and David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (New York, Columbia Global Reports, 2019).

[54] See A/HRC/32/38, paras. 87–88; see also Business for Social Responsibility and World Economic Forum, “Responsible use of technology”, white paper, August 2019.

[55] 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 Monika Bickert, “Updating the values that inform our community standards”, Facebook, 12 September 2019.

Rapporteur reiterates the call for companies to implement human rights policies that involve mechanisms to:

(a) Conduct periodic reviews of the impact of the company products on human rights;

(b) Avoid adverse human rights impacts and prevent or mitigate those that arise;

(c) Implement due diligence processes to "identify, prevent, mitigate and account for how they address their impacts on human rights" and have a process for remediating harm.[56]

43. There will always be difficult questions about how to apply United Nations human rights standards to a wide range of content, just as there are difficult questions about national laws and regional human rights law.[57] However, the guidance mentioned above can help to 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 Guiding Principles on Business and Human Rights (principle 12).[58]

Human rights due diligence and review

44. Dealing with hate speech should start with due diligence at the product development stage. Unfortunately, it seems likely that few if any major Internet companies have conducted a rights-oriented product review related to hate speech; if so, it has not been made public. However, products in the ICT sector are constantly being updated and revised, 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 Guiding Principles on Business and Human Rights, businesses should, among other things, have an ongoing process to determine how hate speech affects human rights on their platforms (principle 17), including through a platform's own algorithms (see A/73/348). 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 and other) of hate speech policies as required under principle 21: while the rules are public, the details of their implementation, at the 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, that is, those conducting the actual work of restriction (principle 16, commentary). As part of the training, the norms of human rights law that the content moderation is aimed at protecting and promoting should be identified. In particular, companies should assess whether their hate speech rules infringe upon freedom of expression by assessing the legality, necessity and legitimacy principles identified above.

[56] Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework (A/HRC/17/31, annex), principles 12 (with commentary), 13 and 15.

[57] See Benesch, "Proposals for improved regulation".

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

Legality standard

46. Company definitions of hate speech are in general difficult to understand, although companies vary on this matter. In some companies such definitions are non-existent, and in others they are vague. For example, the Russian social network VK prohibits content that "propagandizes and/or contributes to racial, religious, ethnic hatred or hostility, propagandizes fascism or racial superiority" or "contains extremist materials".[59] The Chinese messaging app 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".[60] Other definitions are dense and detailed, involving serious efforts to spell out exactly the kind of content that constitutes hate speech subject to restriction, yet the density paradoxically can create confusion and a lack of clarity. The policies of the three dominant American companies – YouTube, Facebook and Twitter[61] – 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 the 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 (A/HRC/38/35, para. 26).

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 mentioned above and provide answers to the following:

(a) What are protected persons or 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 "the promotion … of a positive sense of group identity" in particular 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;[62]

(b) What kind of hate speech constitutes a violation of company rules? Companies should develop hate speech policies by considering the kinds of interference users may face on the platform. Human rights law provides guidance, in particular by noting the legitimacy of restrictions to protect the rights of others. For example, companies could consider how hateful online expression can incite violence that threatens life, infringes upon the freedom of expression and access to information of others, 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 on legal orders from States, themselves subject to the strict conditions established under article 19 (3) of the Covenant;

(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 under article 20 (2) of the Covenant and article 4 of the Convention. In defining such prohibited expression, they should draw from the instruments identified

[59] See https://vk.com/terms.

[60] See www.wechat.com/en/acceptable_use_policy.html.

[61] See https://support.google.com/youtube/answer/2801939?hl=en, www.facebook.com/communitystandards/hate_speech and https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[62] Article 19, Camden principles on freedom of expression and equality, principle 12.

above. However, incitement is only one part of the problematic content that may constitute hate speech. Companies should identify what that category includes in addition to incitement, as some companies have already done through their evolving policies. 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 (A/HRC/38/35, para. 71);

(d) Are there categories of users to 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 adverse actions taken against their accounts. Moreover, an application of the context standards of the Rabat Plan of Action would lead to the protection of such content. Politicians, government and military officials and other public figures are another matter. Given their 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. The evaluation of context may lead to a decision to make an exception in some instances, when the content must be protected as, for example, political speech. However, 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. When company rules differ from international standards, the companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation. For example, were a company to decide to prohibit the use of a derogatory term to refer to a national, 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; in particular in environments of rising tension, 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 to protect fundamental rights.

49. The companies should define how they determine when a user has violated the hate speech rules. At the present time, 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. A set of factors is identified in the Rabat Plan of Action that is applicable to the criminalization of incitement under article 20 (2) of the Covenant, 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 as they would be applied in a criminal context. However, they offer a valuable framework for examining when the specifically defined content – the posts or the words or images that comprise the post – merits a restriction.

50. Companies may find that detailed contextual analysis is difficult and resource-intensive. The largest companies rely heavily on automation in order to do at least the first-layer work of identifying hate speech, which requires having rules that divide content into either one category (ignore) or another (delete). They use the power of artificial intelligence to drive these systems, but the systems are notoriously bad at evaluating context (see A/73/348). 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 category. It must be based on real learning from the communities in which hate speech may be found, that is, 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 audience and evaluate the environment in which hate speech can lead to violent acts. None of these things are possible with artificial intelligence 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 of tools than that enjoyed by States. This 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 conducting a review, preclude users from monetizing their content, create friction in the sharing of content, affix warnings and labels to content, provide individuals with greater capacity to block other users, minimize the amplification of the content, 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 circumstance, and they may require limitations themselves, but they 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 take under 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 whether and demonstrate that the measure it selects actually achieves its goals.[63] This kind of evaluation is in line with the call made in the Guiding Principles on Business and Human Rights for businesses to ensure that they prevent or mitigate harms, in particular because such an approach enables the companies to evaluate the two sets of potential harms involved: the restrictions on speech caused by the 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 to the kinds of expression that are common online – borderline hate speech and non-incitement.

Remedy

53. The mechanisms of international human rights law provide a wealth of ideas for the remediation of online hate speech. Article 2 of the International Covenant on Civil and Political Rights and article 6 of the International Convention on the Elimination of All Forms of Racial Discrimination require that remedies be available for violations of the provisions contained therein, and the Guiding Principles on Business and Human Rights also require access to remedy. In his 2018 report on content moderation, the Special Rapporteur highlighted the responsibility of companies to

[63] Aswad, "The future of freedom", pp. 47–52.

remedy adverse human rights impacts under the Guiding Principles (A/HRC/38/35, para. 59), and therefore it need not be repeated in detail in the present report. In short, the process of remediation must begin with an effective way for individuals to report potential violations of hate speech policies and must ensure 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 should also publicly accessible.

54. At a minimum, the companies should publicly identify the kinds of remedies that they will impose on those who have violated their hate speech policies. It may be that user suspension is insufficient. Companies 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 the de-amplification and de-monetization of problematic expressions that they do not want to ban, for whatever reason, but companies 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 programmes that require suspended users who wish to return to the platform to engage in kinds of reparations, such as 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 on in providing remedies for hateful content. According to the Rabat Plan of Action, “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 “right of correction” and “right of reply” (A/HRC/22/17/Add.4, appendix, paras. 33–34). In its resolution 16/18, the Human Rights Council identifies tools such as training government officials and promoting the right of minority communities to manifest their belief. The previous Special Rapporteur urged procedural remedies, such as “access to justice and ensuring effectiveness of domestic institutions”, and substantive ones, such as “reparations that are adequate, prompt and proportionate to the gravity of the expression, which may include restoring reputation, preventing recurrence and providing financial compensation” (A/67/357, para. 48). However, 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 is often aimed at pushing vulnerable communities off the platforms (i.e., to silence them); promoting and giving greater visibility to mechanisms for responses to hate speech; 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 the proliferation of hateful content (ibid, paras. 56–74).

IV. Conclusions and recommendations

56. **International human rights law should be understood as a critical framework for the protection and respect for human rights when combating hateful, offensive, dangerous or disfavoured speech. Online hate speech, the**

broad category of expression described in the present 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. Some kinds of expression, however, can cause real harm. It can intimidate vulnerable communities into silence, in particular 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 therefore 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.

Recommendations for States

57. **State approaches to online hate speech should begin with two premises. First, human rights protections in an offline context must also apply to online speech. There should be no special category of online hate speech for which the penalties are higher than for offline hate speech. Second, Governments should not demand – through legal or extralegal threats – that intermediaries take action that international human rights law would bar States from taking 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 International Covenant on Civil and Political Rights and article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination 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 legislation on hate speech to meet the requirements of legality, necessity and proportionality, and legitimacy, and subject such rule-making to robust public participation;**

(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 the States 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 harms relating to article 20 (2) of the Covenant or article 4 of the Convention;**

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

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

Recommendations for companies

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

(a) **Evaluate how their products and services affect the human rights of their users and the public, 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 United Nations treaties and interpretations of the treaty bodies and special procedure mandate holders and other experts, including the Rabat Plan of Action;**

(c) **Define the category of content that 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;**

(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 the banning of accounts and the removal of content.**