Statement by David Kaye
SPECIAL RAPPORTEUR ON THE
PROMOTION AND PROTECTION OF THE RIGHT TO
FREEDOM OF OPINION AND EXPRESSION

General Assembly
Third Committee
Item 70 (a-d)

21 October 2019
New York
Mr. President, Excellencies, Ladies and Gentlemen,

In my preparation for this Interactive Dialogue, I went back to review some of my previous presentations before this Committee. Year after year, I have noted the trend of continuing deterioration of the rights to freedom of expression. This is not to say that freedom of expression is under threat everywhere; there are places where years of repression are giving way to new and open forms of governance, though the brave people who push for such change know that their progress is often tenuous and subject to pushback. There are places where freedom of expression is expanding and strengthening.

Yet I have noted regularly the ways in which threats to expression manifest themselves. In 2016, for instance, I noted, “Old tools of repression and new forms of censorship in the digital age are combining with punitive laws and policies to harass, threaten, chill, punish, and otherwise restrict expression around the world.” This is still true.

In 2017 I decried our “era of misinformation and propaganda.” This is also still our era.

Last year I spoke at a dire moment, when we still did not know the details of the murder of Saudi journalist Jamal Khassoghi, a year after the murder a brave Maltese journalist, Daphne Caruana Galizia, when the president of the United States was attacking the media, as he still does, as enemies of the people. A moment when hundreds of journalists and activists remained imprisoned by Turkey, and repressive anti-speech laws were imposed against people in Tajikistan, Iran, Belarus, Nicaragua, Myanmar, and beyond. A moment of attack on civil society in Philippines, Hungary, Cambodia, Russia, the United Arab Emirates, Bahrain, Egypt, and, against an entire population, in China. India’s repeated shut down of the internet last year looks tame compared to the massive communications blackout in Kashmir this year.

These are still unchanged.

The United Nations was founded with the purpose, as the UN Charter memorializes, of “promoting and encouraging respect for human rights and for fundamental freedoms for all.” I am proud to be a part of the special procedures mechanism of the Human Rights Council, to share an affiliation with human rights advocates and scholars and defenders from around the world, but I urge the General Assembly to do all it can to reinforce the human rights we are all bound to promote and that all States are bound to respect, to use its voice and the power of its members to hold accountable all those responsible for global attacks on freedom of expression.

Online ‘hate speech’ and international human rights law

I want to turn to the subject of my present report to the Committee, and I do so with an apology for the delay of the report’s publication and a personal note of its dedication to my father, Dr. Jerry Kaye, who passed away at the beginning of this month after a long-term illness.

The subject of the report is online ‘hate speech’, a short-hand phrase that conventional international law does not define. The phrase’s vagueness 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.
International human rights law provides standards to govern State and company approaches to online hate speech. The present report explains how those standards provide a framework for governments considering regulatory options and companies determining how to respect human rights online.

The report explains in some detail how fundamental principles of international human rights law shape the way in which States may address the broad category of ‘hate speech’ and how companies should protect the rights of their users and the public when dealing with the same kind of expression. It details Articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR), as well as its principles of anti-discrimination in Articles 2 and 26, and it review Article 4 of the International Convention on the Elimination of Racial Discrimination (ICERD). It emphasizes the value of interpretations by the Human Rights Committee and the CERD Committee and draws particular attention to the Rabat Plan of Action as providing standards for evaluating questions of online hate speech.

Online hate speech has become a justifiably global concern. I share the concern that certain kinds of online attack can lead to offline violence, cause discrimination against marginalized communities, deny people the right to expression or privacy or even the right to vote. In this connection, I want to emphasize that freedom of expression – a robust form of it, as guaranteed by the ICCPR – must be part of the solution to hateful content online. Free expression is not the problem; a failure to adjust our institutions and frame our laws according to the problem is. A failure to ensure that our laws meet the standards of legality, necessity and proportionality, and legitimacy undermine the effort to address genuinely harmful content such as incitement to discrimination, hostility and violence.

So with that in mind, let me note the following and then highlight recommendations for States and companies.

**States and online hate speech**

Beginning with States, let’s 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. Any legal evaluation of such a proposal must address the cumulative conditions of Article 19(3) to be consistent with international free expression standards.

I am concerned that State laws applying to hate speech, online and off, often fail to meet the standards of legality, necessity and legitimacy. They are often vague and leave excessive discretion in government authorities to punish expression. 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). Governments have been increasing the pressure on companies to serve as the adjudicators of hate speech.

Further, legislative efforts to incentivize the removal of online hate, and impose liability on internet companies for a failure to do so, often push 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. 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.

**Company moderation of hate speech**

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. 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.

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. 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:

- conduct periodic review of their impact on human rights,
- avoid adverse human rights impacts and prevent or mitigate those that do arise, and
- implement due diligence processes to “identify, prevent, mitigate and account for how they address human rights impacts” and have a process for remediating harm.

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). However, the guidance in international human rights law 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).

**Conclusions and recommendations**

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 the 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.

Government approaches to online hate speech should begin with two premises: First, as the General Assembly and Human Rights Council have noted repeatedly, 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 human rights law outlined in the report, States should at a minimum do the following in addressing online hate speech:

1. Strictly define the terms in their laws that constitute prohibited content under the ICCPR and the ICERD and resist criminalizing such speech except in the gravest situations, such as advocacy of hatred that constitutes incitement to discrimination, hostility or violence, and adopt the interpretations of human rights law contained in the Rabat Plan of Action;

2. 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;

3. 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;

4. 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;

5. 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;

6. 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

7. Actively engage in international processes designed as learning forums for addressing hate speech.

Meanwhile, 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:

1. 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;

2. 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;
3. 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;

4. 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;

5. 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

6. 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.

**M. President,**

Thank you very much for the opportunity to address this Committee this afternoon.