United Nations

# Countering and Addressing Online Hate Speech:

## A Guide for policy makers and practitioners

July 2023

# Foreword

Around the world we are witnessing alarming trends of hate speech on the rise fueling xenophobia, racism, anti-religious hatred and misogyny. Hate speech can have devastating impacts on its victims and on societies. It has resulted in hate crimes, discrimination, and even violence. In the most serious cases, hate speech can be an indicator of risk and a trigger of atrocity crimes, in particular genocide. We know this from history. This is what the history of the Holocaust and the genocides in Rwanda and Srebrenica teaches us. In Nazi Germany magazines such as Der Stürmer were used to disseminate anti-Jewish hatred and conspiracy theories aimed to demonize and dehumanize the Jewish population in Europe leading up to the Holocaust. In Rwanda, ***Radio Télévision Libre des Mille Collines*** (**RTLM**) contributed to spread hate and incite violence against the Tutsi. In many other past and present situations, various media became the vectors of death.

Today social media has become another vehicle for hate speech, with the ability to spread information at a speed never seen before, reaching potentially huge audiences within a few seconds. The manner in which many platforms operate feeds on hateful and discriminatory content, and provides echo chambers for such narratives. Online hate speech has led to real world harm. We have seen this from incidents of identity based violence where the perpetrators were instigated through online hate, to its widespread use to dehumanize and attack entire populations on the basis of identity. Unfortunately, many times the victims are those already most marginalized in society, including ethnic, religious, national or racial minorities, refugees and migrants, women and men, sexual orientation and gender identity minorities.

However, just as social media provides the means for disseminating hate speech, it can also provide the tools for tackling it. We have seen positive examples of concerted efforts to address online hate, when the necessary resources and capacities to understand the particular context, language and impact of such hate is prioritized. Unfortunately, the investment to counter online hate does not yet match the reality of its spread and impact online. Much more needs to be done.

It was with this in mind that the Secretary-General, in June 2019, launched the United Nations Strategy and Plan of Action on Hate Speech. My office – the United Nations Office on Genocide Prevention and the Responsibility to Protect – is the United Nations global focal point on the implementation of this Strategy that seeks to enhance United Nations efforts to address the root causes and drivers of hate speech as well its impact on victims and societies. In this context, my Office coordinates system-wide efforts to counter and address hate speech, including by supporting United Nations field presences, Member States, and civil

society to develop context specific and national action plans on addressing and countering hate speech.

One of the commitments of the Strategy is to use new technologies and engage with social media to address online hate speech. In this regard, my Office has a longstanding engagement with technology and social media platforms to promote policies and practices on addressing online hate speech, in line with the UN Strategy. We have organized annually, since 2020, roundtable discussions with tech and social media companies on their role and responsibilities in addressing hate speech on their platforms, in line with international human rights norms and standards. These roundtables organized in collaboration with the ESRC Human Rights, Big Data and Technology Project at the University of Essex, also included the members of the United Nations Working Group on Hate Speech as well as Special Rapporteurs and civil society organizations working on this topic. The roundtables provide a platform for dialogue engagement and action on combating online hate speech. They aimed at looking beyond the limited scope of content removal to address hate speech online holistically, including through promoting positive narratives, warnings on problematic content, as well as reducing virality of posts and countering inauthentic coordinated behavior and other forms of disinformation related to online hate speech.

We cannot ignore the dangers of online hate. We must all act to counter it. If not, hard won gains in advancing non-discrimination and equality are at risk of being eroded by those who seek to maintain or consolidate power at any cost, often manipulating identity for political gain, scapegoating and targeting vulnerable groups through hate speech, using the opportunities provided by new technologies and social media. The technology and social media companies have a crucial role in responding to these dangers on their platforms.

This document sets up the main recommendations identified through the three years of engagement and dialouge on this topic. It is anchored in the discussions from the round tables. It further builds on the experiences from UN field presences that are working to address these challenges at country level. The paper is the result of a cooperation between my Office and the ESRC Human Rights, Big Data and Technology Project at the University of Essex. I would like to especially thank Ahmed Shaheed, Deputy Director, Essex Human Rights, Big Data and Technology Project and former United Nations Special Rapporteur on Freedom of Religion or Belief for this partnership over the last several years and which has resulted in this policy paper.

The importance of tackling hate speech has been reiterated in the United Nations Our *Common Agenda*, that sets out specific initiatives complimentary to the UN Strategy and Plan of Action. This guide for technology and social media companies also builds on previous initiatives such as the 2021 Global Ministers of Education Conference on the role of Education in Addressing Hate Speech which informed the 2023 Guide for poilcy makers, *Addressing hate speech through education: A guide for policy-makers*. This guide

also builds on the 2022 Policy Paper on *Combating Holocaust and Genocide Denial Protecting Survivors, Preserving Memory, and Promoting Prevention.*

It is my firm hope that the recommendations in this document will help accelerate action to tackle online hate speech and in doing so promoting and upholding our fundamental rights, in particular freedom of opinion and expression as well as to non-discrimination and equality.

**Alice Wairimu Nderitu**

*Under-Secretary General and Special Adviser on Prevention of Genocide to the United Nations Secretary General*

# I. Introduction

**On 18 June 2019, the United Nations Secretary-General launched the United Nations Strategy and Plan of Action on Hate Speech. The Strategy represents the United Nations commitment to step up its action to address this global challenge in an holistic way, including various relevant actors in society. The Strategy comprises 13 commitments of action to address and counter hate speech in line with international human rights norms and standards, the right to freedom of opinion and expression in particular. The Strategy further emphasizes the importance of partnerships in tackling hate speech, and among its guiding principles it notes: *'tackling hate speech is the responsibility of all – governments, societies, the private sector, starting with individual women and men. All are responsible, all must act'*. Two of these commitments stress the importance of engaging technology and social media companies:**

**Commitment #5** calls on the UN to establish and strengthen partnerships with new and traditional media to address hate speech narratives and promote the values of tolerance, non-discrimination, pluralism, and freedom of opinion and expression.

**Commitment #6** emphasizes that relevant UN entities 'should keep up with technological innovation and encourages more research on the relationship between the misuse of the Internet and social media for spreading hate speech and the factors that drive individuals towards violence'. This same commitment also calls on relevant UN entities to 'engage private sector actors, including social media companies, on steps they can take to support UN principles and action to address and counter hate speech, encouraging partnership between government, industry and civil society'.

Other relevant commitments include: #1 Monitoring and analyzing hate speech; #4 Convening relevant actors; and #11 Leveraging partnerships.

Engaging with technology and social media platforms is critical as they have become one of the main vectors for disseminating hate speech. Indeed, with just one click, hate speech can rapidly reach very far distances.

When launching the Strategy in 2019 the Secretary-General noted:

***"As new channels for hate speech are reaching wider audiences than ever at lightning speed, we all – the United Nations, governments, technology companies, educational institutions – need to step up our response."***

Furthermore, in the United Nations 'Our Common Agenda' report from 2021, addressing online hate speech is identified as a frontier issue to prevent harms in the digital or technology spaces. Hence, addressing these issues requires approaches that are mindful that the same rights that people have offline

must also be protected online, in particular freedom of expression.[1].

There is no legal definition of "hate speech" in international law. Under the UN Strategy and Plan of Action, hate speech is understood as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates, intolerance and hatred, and in certain contexts can be demeaning and divisive."[2]

The Strategy's definition of hate speech captures a very broad range of expression. The types of hate speech covered by the Strategy can be divided into three categories, according to the level of severity.

1. AT THE TOP LEVEL, "direct and public incitement to commit genocide" and "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" are prohibited under international law[3]. 2. AT THE INTERMEDIATE LEVEL, certain forms of hate speech may be prohibited, but only if restrictions are provided by law, pursue a legitimate aim (e.g. respect of the rights of others, or the protection of public order),are necessary for the protection of the legitimate aim and proportionate in the sense that they must be the least intrusive measure to achieve such protection. 3. AT THE BOTTOM LEVEL, expression that may not be subject to restriction, including expression that is offensive, shocking or disturbing.[4]

To implement the commitments of the UN Strategy related to tackling online hate speech, the UN Office on Genocide Prevention and the Responsibility to Protect, in its capacity as global focal point for the implementation of the UN Strategy, has established partnerships with technology and social media companies to address areas of concern and identify opportunities for collaboration.

[1] A/HRC/RES/20/8
[2] UN Strategy and Plan of Action on Hate Speech, June 2019
[3] Convention on the Prevention and Punishment of the Crimes of Genocide, Article III; International Covenant on Civil and Political Rights Articles 19 and 20; International Convention on the Elimination of Racial Discrimination Article 4 ; .

[4] Detailed Guidance on UN Strategy and Plan of Action, June 2020; Also refer to the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, UN Doc. A/HRC/22/17/Add.4, appendix; 'Hate Speech' Explained: A Toolkit by Article 19.

# II. Addressing Online Hate Speech

In this context, between 2020 and 2023, this Office, in partnership with the ESRC Human Rights, Big Data and Technology Project at the University of Essex, organized four roundtable consultations to discuss the implementation of the United Nations Strategy and Plan of Action on Hate Speech in the online space. The roundtables engaged technology and social media companies on their role and responsibility to address online hate speech in line with international human rights norms and standards, and with the view to building partnerships for the implementation of the UN Strategy and Plan of Action on Hate Speech. The roundtable discussions, held under Chatham House Rule, also included the members of the United Nations Working Group on Hate Speech[5], civil society and relevant Special Procedures mandate holders.

Each roundtable identified a number of challenges and priorities in regard to addressing online hate and resulted in a set of recommendations addressed to technology and social media companies, Member States, United Nations, and civil society, including academia and media.

The primary challenges identified included:

- Non-compliance by States with their obligations to respect, protect and promote freedom of expression and prohibit incitement to discrimination and violence, resulting in either overly broad restrictions on freedom of expression, including take-down demands on intermediaries or insufficient regulations to protect communities targeted by hate speech. Similarly, the lack of a non-discrimination lens in initiatives aimed at tackling online hate.
- Inadequate investment by social media companies in efforts to counter online hate speech, given the speed, volume and diversity of online postings. In particular lack of investment in tackling hate speech that does not reach the threshold of incitement, requiring multipronged approaches, beyond content removal.

[5] The United Nations Working Group on Hate Speech includes the following entities: Department of Global Communications, Department of Peace Operations, Department of Political and Peacebuilding Affairs, Envoy of the Secretary-General on Youth, Executive Office of the Secretary-General of the United Nations, Global Pulse, International Organization for Migration, Office of Counter-Terrorism, the United Nations Office on Genocide Prevention and the Responsibility to Protect, Office of the United Nations High Commissioner for Human Rights (OHCHR), Office of the United Nations High Commissioner for Refugees (UNHCR), United Nations Alliance of Civilizations, United Nations Children's Fund (UNICEF), United Nations Development Programme (UNDP), United Nations Educational, Scientific and Cultural Organization (UNESCO) and United Nations Entity for Gender Equality and the Empowerment of Women (UN-Women).

- Limited transparency both from companies regarding their policies and from States in their demands and rulings. More transparency is needed on how companies moderate and curate content, their processes including for the implementation of their community guidelines, as well as information on their responses to due diligence, redress, human rights impact assessments, and ways to appeal content moderation decisions, among other issues. Similarly, States should be more transparent on their requests for content removal and avoid passing legislation that unduly prevents companies from disclosing governments' requests for users' data or content removal.

In follow-up to the roundtables, the UN Office on Genocide Prevention and the Responsibility to Protect has supported action in follow up to the recommendations, engaging with Member States, social media companies and other relevant actors. In this same vein, and based on expert consultations, in June 2022, the Office published a policy paper on *Combating Holocaust and Genocide Denial Protecting Survivors, Preserving Memory, and Promoting Resilience* which included specific recommendations on addressing Holocaust and genocide denial online. The Office has also supported, with the United Nations Working Group on Hate Speech, consultations on the impact of online gender-based hate speech. Moreover, in October 2021, the Secretary-General convened a Global Ministers of Education Conference on the role of Education in Addressing Hate Speech. This conference, co-organized by the Office and UNESCO, resulted in 'conclusions of the chairs' which set out key recommendations areas.

The present policy paper builds on these initiatives and aims at disseminating the main recommendations from three years of dialogue to a wider audience and to enable a broad-based mobilization on concrete action needed to further enhance efforts to tackle online hate. The policy paper will enable UN Country Teams, Member States and civil society to engage with technology and social media companies on the importance of addressing hate speech online, in line with the UN Strategy and Plan of Action and under the overarching umbrella of international human rights norms and standards.

# III. Recommendations

### *Ensure Respect for Human Rights and the Rule of Law when Countering Online Hate Speech, and Apply these Standards to Content Moderation, Content Curation and Regulation*

**To States:**

- **Regulation of digital communications must always comply with their [States] obligations under international human rights law, in particular article 19 of the International Covenant on Civil and Political Rights (ICCPR) on the right to hold opinions without interference and the right to freedom of expression.** The right to freedom of expression under international human rights law, includes "the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers". Any restriction to the right to freedom of expression shall only be such as are provided by law and are necessary, as established under article 19.3. It should also take into consideration their obligations related to non-discrimination and equality, including article 2.1 of the ICCPR, as well as the freedom to seek, receive and impart information and ideas of all kinds. For any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, States have an obligation to prohibit this by law (ICCPR, article 20.2). Also, States must ensure that technology and social media companies comply with their responsibility to conduct human rights due diligence in accordance with the UN Guiding Principles on Business and Human Rights (UNGPs). In this regard, guidance has been provided, in particular on restrictions to the right to freedom of expression, by the UN Human Rights Committee in General Comment no. 34 on article 19, 'Freedoms of Opinion and Expression', the UN Committee on the Elimination of Racial Discrimination in General Recommendation no. 35 on 'Combating Racist Speech', and 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, including the latter's six-part threshold test.[6]
- **Legislative efforts** to prohibit hate speech that reaches the threshold identified in article 20.2 of the ICCPR must be developed through participatory efforts, in particular with the participation of groups who are subject to such incitement to hatred.

[6] UN Doc. A/HRC/22/17/Add.4, appendix, para. 29.

- **Policies** must also be formulated to counter and address hate speech that does not reach this threshold through similar participatory efforts.
- Measures taken to address hate speech must aim at **building systems that effectively address the problems holistically**, offline and online. Such measures should not encourage mass surveillance[7], criminalization of the exercise of freedom of expression as guaranteed under international law, undermine trust or attempt to regulate each and every piece of content.
- Ensure that there is no impunity for advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence (ICCPR, article 20.2) as well as public and direct incitement to genocide as set out under the Convention on the Prevention and Punishment of the Crime of Genocide**.**
- Refrain from imposing internet shutdowns[8], including as a response to hate speech. Internet shutdowns is damaging to a range of human rights. Keeping the internet on, and expanding connectivity for those on the other side of the digital divide are essential.
- Ensure that official requests for **takedowns and removal of online** content follow existing guidelines and are compliant with human rights norms and standards and enhance transparency of such requests, in line with the Rabat Plan of Action.
- States institutions and public authorities should not weaponize social media to spread hate speech.
- Refrain from imposing obligations on social media platforms to monitor online content generally, and require them to establish mechanisms to address specific content such as hate speech, especially when it amounts to incitement to discrimination, hostility or violence.

**To Technology and Social Media Companies:[9]**

- Conduct human rights due diligence to identify the risks the use of their services may pose to people, and take all reasonable steps to prevent or mitigate such risks, in accordance with the UN Guiding Principles on Business and Human Rights. Effective due diligence has to be a continuous, ongoing and iterative process; supported by efforts to embed human rights into policies and management systems; and aimed at

[7] A/HRC/51/17
[8] A/HRC/50/55

[9] While these recommendations should be considered by all technology and social media companies, they were developed especially with larger online platforms in mind.

enabling companies to remediate adverse impact that they cause or contribute to[10].

- Adopt **community standards and frameworks** for content moderation that are in line with international human rights norms and standards, in particular the guarantees of freedom of thought, opinion and expression, and the rights to equality, non-discrimination and privacy.
- Ensure that frameworks and policies are transparent and are applied consistently. In addition, ensure accessible and consistent notice and review procedures, and effective remedies[11].
- Carefully consider **contextual factors while applying community standards on hate speech, including history, language and socio-economic elements, building on the six-part threshold test of the Rabat Plan of Action (context, speaker, intent, content, extent, and likelihood of harm).[12]** Alternatives to content removal should be considered, such as labels, demonetization, limiting "reach", or counterspeech. Adopt a human **rights-respecting business model[13]** that ensures the protection of the rights of users and others affected by content on the platform, including non-discrimination, right to privacy and right to freedom of thought, opinion and expression. Content moderation and the design of content curation and recommender systems should be done in consultation with human rights experts.
- Demonstrate that **policies and decision-making processes** draw on international human rights norms and standards and associated guidance, including the UN Strategy and Plan of Action on Hate Speech, and improve **communication** regarding key concepts and definitions, including those related to hate speech.
- Ensure that there is always a 'human in the loop' for any automated system used for **content moderation and ensure** that it is a competent human oversight, capable of understanding, rectifying, and reporting automated decisions.

[10] For further information about Human Rights Due Diligence in the technology sector, please see Key Characteristics of Business Respect for Human Rights: A B-Tech Foundational Paper at https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf and A/HRC/50/56

[11] See B-tech Foundational paper on access-to-remedy-company-based-grievance-mechanisms.pdf (ohchr.org)

[12] See the Rabat threshold test, which is available in 32 languages online at www.ohchr.org/en/freedom-of-expression.

[13] For further information on addressing risks to people connected to business models,seehttps://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/B_Tech_Foundational_Paper.pdf

- Invest in improving the capacity and quality of **content moderation** in all languages in which their platforms can be used.
- Dedicate more resources to moderating content that could constitute incitement to discrimination, hostility or violence, or is particularly harmful, in countries or contexts that are fragile or high-risk, with an increased focus on dedicating resources to languages other than English.
- Engage with the relevant entities at the United Nations, including the UN Office on Genocide Prevention and the Responsibility to Protect as the UN Focal Point on Hate Speech, dedicate sufficient capacity, and consider the use of heat maps for early warning and early action to prevent violence, especially around major political events.
- Continue to strengthen efforts to detect and address non-verbal hate speech that appears including through **videos, music pictures, memes and other media**, as well as through **coded language**, that might be harder to detect.
- Assess the impact of hate speech by **political and religious leaders and other public figures** on their respective platforms and adopt policies and guidelines in relation to such content in line with international human rights norms and standards[14] and with the Plan of Action for Religious Leaders and Actors to Prevent Incitement to Violence that Could Lead to Atrocity Crimes (Fez Plan of Action)[15].
- **Ensure that content curation is informed by and compliant with international human rights norms and standards and that particular attention is paid to avoid any undue restrictions of freedom of opinion and expression.** Undertake periodic and publicly available **human rights impact assessments in line with the UN Guiding Principles on Business and Human Rights with regard to content moderation, content curation and prioritization practices.**

**To Civil Society:**

- Root hate speech analysis and assessment in international human rights norms and standards to **monitor online hate speech trends**, in collaboration with affected communities; and promote regulations that are in line with international human rights norms and standards.

---

[14] See Rabat Plan of Action, UN Doc. A/HRC/22/17/Add.4, appendix; Beirut Declaration and its 18 commitments on "Faith for Rights", UN Doc. A/HRC/40/58, annexes I and II; Plan of Action for Religious Leaders and Actors to Prevent Incitement to Violence that Could Lead to Atrocity Crimes, available online at www.un.org/en/genocideprevention/documents/publications-and-resources/Plan_of_Action_Religious-rev5.pdf.

[15] For further information please see https://www.un.org/en/genocideprevention/documents/publications-and-resources/Plan_of_Action_Religious-rev5.pdf

### *Enhance Transparency of Content Moderation, Content Curation and Regulation*

**To States:**

- Remove undue legal barriers for disclosure of **transparency reports**, including government-prompted content removals and requests on user data. Refrain from issuing orders and restrictions on reporting of government-requested content removal.
- Require technology and social media companies to be more transparent on their **content moderation, ranking practices and algorithms**, as well as their operations related to due process and redress, personal data gathering and use, and implemented business model.

**To Technology and Social Media Companies:**

- Continue to increase **transparency** and strengthen **granularity of reporting**, particularly regarding the operation of automated systems for content moderation and curation as well as information on human rights **due diligence, redress, and human rights impact assessments**. These measures cannot be applied as a blanket one-size fits-all approach to all companies but must be tailored to the scale and engineering of each platform. Platforms must provide more consistent and disaggregated information on the interventions they make, the use of automation in content moderation, and on interventions motivated by government requests, including government-prompted content removals and requests on user data.
- Increase transparency to users and to other stakeholders (such as researchers, journalists, civil society) on how they address online hate speech, including greater **transparency on policies and definitions, use of automation for content moderation and ranking algorithms**.[16]
- Clearly state how artificial intelligence (AI) technologies and automation are used on their platforms, and their known risks to individuals.

**To Civil Society:**

- Continue advocacy for increased **transparency** from technology and social media companies and Member States on content regulation and moderation; continue to advocate for enhanced transparency by technology and social media companies, in their products, operations as well as in the data that they make available.

### *Promote Positive Narratives to Counter Online Hate Speech, and Foster User Engagement and Empowerment*

**To the United Nations:**

- Enhance the use of digital platforms to promote **positive content to counter online hate speech** and seek the support of technology and social media companies to amplify these messages.

---

[16] Also refer to the UNESCO Internet for Trust initiative: https://unesdoc.unesco.org/ark:/48223/pf0000384031.locale=en

**To States:**

- Engage in promoting **positive messages and the use of counter-narratives** to address and combat hate speech, including by promoting diverse and pluralistic journalism.

- Develop **comprehensive anti-discrimination legislation**[17] and promote holistic equality measures to address root causes of hate.

- Promote media and information literacy for the entire population and increase **digital inclusion**, as a means to enhance resilience and empower individuals to identify and counter online hate speech.

**To Technology and Social Media Companies:**

- Develop and implement **guidance on tackling online hate speech** in times of crisis such as conflict and humanitarian crisis, health emergencies where impact of hate speech is higher and may have more serious consequences on victims and groups affected.

- Promote explicit and consistent policies to address online **Holocaust and genocide denial** as well as denial of other atrocity crimes, that are in line with international law guarantees on freedom of opinion and expression and that protect victims affected by the content. Broaden existing good practices of linking to factual information and authoritative voices on the Holocaust and other instances of genocide when users engage with content that seeks to deny established facts.

- Explore the possibility of expanding and developing **plurality of approaches** to tackle hate speech beyond content removal, **de-amplification and user warnings** and through promoting **alternative and positive narratives**. However, content curation and prioritisation policies and practices should be transparent to users and compliant with international human rights law.

- Foster partnerships with governments, multilateral organizations and civil society and any other relevant stakeholder to promote **digital literacy and inclusion**.

- Increase user-participation through finding ways to more effectively engage **local communities and civil society** without overburdening them, and ensure more transparency into how technology and social media companies use consultation processes.

**To Civil Society:**

- Continue to promote **alternative and positive narratives** to counter online hate speech.

- Increase advocacy for **user detection of online hate speech** and promote awareness-raising projects and **overall media literacy**.

- Where appropriate, **connect individuals and groups with governmental authorities and technology and social media companies**. Based

[17] OHCHR/Equal Rights Trust, *Protecting Minority Rights: A Practical Guide to Developing Comprehensive Anti-Discrimination Legislation* (2022), available online at www.ohchr.org/en/minorities/minority-rights-equality-and-anti-discrimination-law.

on the engagement with individuals and groups, they should further promote conversations with technology and social media companies and State authorities on the shaping of protective policies, through knowledge from working closely with populations.

### *Ensure Accountability, Strengthen Judicial Mechanisms and Enhance Independent Oversight Mechanisms*

**To United Nations Actors:**

- Strengthen recommendations and advocacy to **hold technology and social media companies accountable** to commitments made in their terms of service and their responsibilities under the UN Guiding Principles on Business and Human Rights.
- Reach out to newer or smaller technology and social media companies where policies on harmful content are being developed and hate speech and genocide denial is rife and encourage them to adopt **robust anti-hate speech policies**. Reach out to other far-reaching online platforms where such content is widespread.
- Collect and disseminate good practice standards on the discharge of the responsibilities of digital platforms to address hate speech.

**To States:**

- Promote and support the **role of institutions** in tackling online hate speech, in particular by strengthening the role of national courts. This would require specific **training for judges** on international human rights norms and standards, in particular issues such as freedom of expression and artificial intelligence. Such trainings could be carried out in partnership with relevant UN agencies or technology and social media companies and other stakeholders, including civil society organisations.

**To Technology and Social Media Companies:**

- Ensure that users have effective opportunities to appeal or request the review of content moderation decisions. Policies governing such appeal and redress processes should be publicly available in relevant languages. Companies should have in place systems by which users are expressly notified about content interventions, where legally permissible.
- Effective redress channels must be accessible, including in terms of language diversity and age appropriateness, and transparent to all. Companies should offer equal redress opportunities in every country where they have users. Effective remedies should be available for when actions by companies or States undermine the rights of users.
- Establish **independent and transparent multistakeholder oversight mechanisms** to ensure that content moderation policies and practices are compliant with international human rights norms and standards. Support the establishment of independent governance models, based on a multi-stakeholder approach, bringing together a range of expertise, and entrusted with the review of content moderation decisions made by technology and social media platforms.

- Abide by human rights commitments, including human rights **due diligence responsibilities**, and explore the possibility of adopting independent **multi-stakeholder oversight models** with relevant and competent bodies to provide and advocate for digital, media, and information literacy for all which could lead to enhanced visibility of groups in vulnerable or marginalized situations.
    - Improve **appeals and remedial processes** in a way they are available, accessible, acceptable, and transparent.
- Put in place appropriate procedures to **prevent trauma in staff members** caused by content moderation (including from external providers) or radicalization through exposure to hateful rhetoric and strengthen responses to **support staff affected by online hate speech**.

### *Strengthen Multilateral and Multi-Stakeholder Cooperation*

**To United Nations Actors:**

- The UN Strategy and Plan of Action on Hate Speech provides a framework for all stakeholders, including technology companies, and common entry point to engage with the UN, through the Office on Genocide Prevention and Responsibility to Protect as the UN focal point. The UN should adopt a **coordinated and coherent approach in engaging with technology and social media companies** on addressing and countering hate speech.
- Continue dialogue in an open-ended multi-stakeholder process, at global, regional, and national levels, within the framework of the **UN Strategy and Plan of Action on Hate Speech**, under the leadership of the UN Office on Genocide Prevention and the Responsibility to Protect.
- Strengthen **multi-stakeholder engagement and collaboration**, and facilitate dialogue between governments, national human rights institutions, civil society, academia, and technology and social media companies. Involve existing and newly established regulatory bodies which have been entrusted with oversight responsibilities over technology and social media companies, as well as more technology and social media companies, including small and medium-sized companies around the world, and national human rights institutions, parliamentarians, politicians, journalists, and traditional media actors.
- Facilitate collaboration of civil society organizations, and technology and social media companies to work with each other so that CSOs with localized contextual understanding can support **monitoring and safely promote alternative positive narratives** on these online platforms.
- Enhance collaboration with inter-governmental organizations, governments and technology and social media companies to create and disseminate **joint campaigns** to counter online hate speech.

- Continue to promote and expand a **multi-stakeholder rights-based and holistic approach** to tackle online hate speech, including through **education,[18] peer-to-peer learning,[19] awareness-raising, and fostering peaceful, inclusive and just societies**.
- Convene stakeholders to develop guidance for member States and technology and social media companies. This might include, for example, initiating a **multi-stakeholder working group** to develop an international framework on hate speech for the digital age, in line with the UN Strategy and Plan and Action on Hate Speech and international standards to respect freedom of expression, while prohibiting the advocacy of hatred that constitutes incitement to discrimination, hostility or violence, and the promotion of counter narratives. It should also build on existing mechanisms and platforms such as the UN Working Group on Hate Speech.

**To States:**

- Promote a multi-stakeholder approach to tackle online hate speech and continue dialogue in an open-ended multi-stakeholder process within the framework of the **UN Strategy and Plan of Action on Hate Speech**, under the leadership of the UN Office on Genocide Prevention and the Responsibility to Protect, and with high- level buy-in and commitment from the key stakeholders.
- Foster **engagement with social media companies, civil society, and affected communities** in order to shape laws, policies and strategies towards addressing online hate speech consistent with international human rights law.
- Ensure the creation of mutual **platforms of communication and collaborative networks**, including at country level in conflict-prone situations, which will enable the inclusive participation of different stakeholders and consultation with civil society organizations and technology and social media companies on hate speech issues.

**To Technology and Social Media Companies:**

- Create and strengthen concrete **channels of collaboration** with external stakeholders, including civil society and affected communities, in order to have a more in-depth insight on how hate, misinformation, disinformation and harassment are manifesting in local contexts and ensure moderation standards pay attention to context and is conducted with **trained human support**.
- Engage with governments to understand context, including obstacles to public freedoms, and contribute to **national policies and strategies** towards addressing online hate speech consistent with international human rights law.
- Engage with **civil society organizations and affected communities** as they seek to address and counter online hate speech. It is also necessary to

[18] See for example UNESCO, *Manual for developing intercultural competencies: story circles* (2020), available online at https://unesdoc.unesco.org/ark:/48223/pf0000370336/PDF/370336eng.pdf.multi.
[19] See for example OHCHR, *#Faith4Rights toolkit* (2023), available online at www.ohchr.org/Documents/Press/faith4rights-toolkit.pdf.

involve those most affected by hate speech in the design, development and use of effective tools to address harm caused on and by the platforms.

### *Advance Community-Based Voices and Formulate Context-Sensitive and Knowledge-Based Policymaking and Good Practice to Protect and Empower Groups and Populations in Vulnerable Situations to Counter Online Hate Speech*

**To United Nations Actors:**

- Engage and support **community-based and local voices** to address and counter hate speech.
- Explore the possibility of an **expedited channel** between UN field missions and technology and social media companies, including the Trusted Partner System, to support the identification of online hate speech and to flag situations that could potentially lead to physical harm, including atrocity crimes through online advocacy of hatred.
- Provide technical assistance and capacity-building to civil society organizations and media actors to strengthen their capacity to address hate speech and promote positive messages at local level and contribute to the development of **new tools/localized solutions** tailored to local context to curb online hate speech.
- Continue to support civil society organizations to engage with relevant authorities, technology and social media companies, and **raise recommendations, in particular at local levels**.

**To States:**

- Recognise the important role played by offline measures that build societal resilience against online incitement to hatred and foster such measures such as strengthen the rule of law, non-discrimination and inclusion. In this vein support implementation of the recommendations for the 2021 Global Education Ministers Conference on the role of education in addressing hate speech.[20]
- Rescind laws and policies that discriminate against groups on the basis of identity and ensure equal protection of the law to all, especially those who face exclusion, marginalization and stigmatization.
- Strengthen the capacity of National Human Rights Institutions, **civil society organizations** and other relevant actors to support their efforts in addressing and responding to online hate speech and the offline root causes thereof.

**To Technology and Social Media Companies:**

---

[20] Addressing Hate Speech through Education: Global Education Ministers Conference, 26 October 2021: conclusions by the Conference Chairs, available online at https://unesdoc.unesco.org/ark:/48223/pf0000379729.

- Strengthen efforts to understand and tackle online hate speech in **local contexts**, including language moderation, and enhance cooperation with UN on the ground and with national and local CSOs.
- Tackle online hate speech in a way that is commensurate to the challenges posed by its volume, speed, and diversity. Pay increasing attention to **targeted individuals and groups**, diverse contexts and community standards, efforts and resources in place (including access to tools and channels to assist) where they are most needed to protect those individuals and groups mainly targeted by hate speech in social media.
- Continue and strengthen emphasis on countering online hate speech against **groups particularly at risk**: minorities, migrants, refugees and internally displaced persons, women, sexual orientation and gender identity minorities, dissenters, human rights defenders, journalists, and other civil society voices representing targeted groups, as well as other groups in vulnerable positions (e.g. children).
- Provide **disaggregated data** to independent researchers, journalists and relevant bodies, while respecting the privacy of users, to better understand how much content is managed automatically and how successfully/unsuccessfully the automated tools address key concerns on incitement to hatred, in order to enable further understanding of hate speech to inform **evidenced-based policy responses**.

**To Civil Society:**

- Amplify and support voices of **local individuals and communities** active in addressing and countering online and offline hateful rhetoric, and work with journalists to produce **inclusive reporting**.
- Conduct comprehensive research on the **link between the offline root causes of hatred and online manifestations of hate** as well as between the online hate and the offline harm and identify good practices in order to design multifaceted and grounded responses to hate speech.
- Enhance research on existing **policies, interventions, and impact** of measures adopted by technology and social media companies and online platforms to combat hate speech.

# IV. Conclusion

Online hate speech remains a critical challenge to advancing the objectives set out in the UN Strategy and Plan of Action on Hate Speech and the pillars of the United Nations work, namely Peace and Security, Human Rights and Sustainable Development. The recommendations outlined in this policy paper, provide a framework for countering hate speech, in line with international human rights norms and standards. The recommendations are based on three years of consultations and dialogue, including with the technology and social media companies, experts, the UN Working Group on Hate Speech and civil society. Their implementation should be part of broader efforts to address hate speech globally, including its root causes and impact offline, in line with the UN Strategy and Plan of Action. The pursuit of these recommendations should also prioritize participation and engagement directly with the victims of hate speech, underpinned by the principles of non-discrimination and leaving no one behind.