



unesco

Social Media 4 Peace

Local lessons for global practices



Published in 2023 by the United Nations Educational, Scientific and Cultural Organization (UNESCO), 7, place de Fontenoy, 75352 Paris 07 SP, France and UNESCO Liaison Office in Brussels, 35 square de Meeus, 1000 Brussels, Belgium

© UNESCO 2023

ISBN 978-92-3-100610-4



This publication is available in Open Access under the Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) license (<http://creativecommons.org/licenses/by-sa/3.0/igo/>). By using the content of this publication, the users accept to be bound by the terms of use of the UNESCO Open Access Repository (<http://en.unesco.org/open-access/terms-use-ccbysa-en>).

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

This publication was funded by the European Union under the UNESCO project 'Social Media 4 Peace.' Its content is the sole responsibility of its authors and do not necessarily reflect the views of the European Union.

Editors: Adeline Hulin

Author: João Brant

Research assistants: Diogo Moyses, Sivaldo Pereira

Graphic design: Monika Martinovic

Cover design: © UNESCO/Monika Martinovic

Illustrations: Monika Martinovic



**Funded by the
European Union**

SHORT SUMMARY

Local lessons for countering online harmful content

Harmful content, particularly hate speech and disinformation, has become pervasive in the digital realm, profoundly impacting people's lives beyond virtual interactions. It seeps into the real world, affecting human rights, social cohesion, democracies, and peace. This has corroded public discourse and fragmented societies, with marginalized communities often bearing the consequences.

Addressing these challenges requires understanding the root causes and impact of harmful content. Governments, social media companies, civil society organizations, and international bodies must collaborate to develop strategies that protect fundamental rights online while safeguarding users.

This publication, developed under the UNESCO project "Social Media 4 Peace" funded by the European Union, overviews research conducted under the project focusing on Bosnia and Herzegovina, Kenya, and Indonesia. These include analyses of the regulatory frameworks governing harmful content online in these target countries, assessments of self-regulatory tools and content moderation policies of platforms, and the mapping of the local efforts by civil society.

The publication aims to inform global discussions on countering harmful content, especially in conflict-prone environments, by delving into the complexities of these countries' political, cultural, linguistic, and societal contexts. Its insights aim to serve as guideposts for stakeholders seeking to promote freedom of expression and a safer online environment.

Only **13%**
of Facebook's moderation
budget was allocated for
developing algorithms to
detect misinformation
outside the USA in 2020.



“Since wars begin in the minds of men and women it is in the minds of men and women that the defences of peace must be constructed”



unesco

Social Media 4 Peace

**Local lessons for
global practices**

Contents

List of Acronyms -----	3
Executive Summary -----	4
<i>Main findings</i> -----	5
<i>Recommendations</i> -----	5
Introduction-----	6
The importance of analyzing the digital realm in conflict-prone countries -----	7
Chapter 1: International standards on freedom of expression and the legitimate restrictions of harmful content-----	9
1.1 What is harmful content and what are its limits -----	9
1.2 Protected rights and bounded restrictions -----	12
1.3 Paths and challenges -----	17
1.3.1 <i>Lack of uniformity in definitions</i> -----	17
1.3.2 <i>Lack of transparency</i> -----	19
1.3.3 <i>Concentration of power and decision-making</i> -----	21
1.3.4 <i>Effectiveness and enforcement</i> -----	22
Chapter 2: Overview of country reports -----	26
2.1 Bosnia and Herzegovina -----	26
2.1.1 <i>Context</i> -----	26
2.1.2 <i>Legislation addressing harmful content</i> -----	27
2.1.3 <i>Civil Society and Companies' initiatives</i> -----	29
2.1.4 <i>Analytical Synthesis</i> -----	31
2.2 Indonesia -----	33
2.2.1 <i>Context</i> -----	33
2.2.2 <i>Legislation addressing harmful content</i> -----	34
2.2.3 <i>Civil Society and Companies' initiatives</i> -----	36
2.2.4 <i>Analytical Synthesis</i> -----	37
2.3 Kenya -----	39
2.3.1 <i>Context</i> -----	39
2.3.2 <i>Legislation and main State initiatives addressing harmful content</i> -----	40
2.3.3 <i>Initiatives by Companies</i> -----	41
2.3.4 <i>Initiatives by Civil Society</i> -----	42
2.3.5 <i>Analytical Synthesis</i> -----	43

Chapter 3: Comparison points -----	45
3.1 Evidence of the impacts of hate speech and disinformation on peace and human rights at the national level-----	45
3.2 Compatibility between national legislation and international standards -----	46
3.3 Effective enforcement of legal frameworks -----	46
3.4 Presence and local contextualization of social media companies -----	47
3.5 Multi-stakeholder environment -----	48
Chapter 4: Main findings -----	50
Chapter 5: Recommendations -----	55
5.1 To international organisations -----	55
5.2 To States -----	57
5.3 To Social Media Companies -----	59
5.4 To Civil Society -----	63
5.5 To Donors -----	64
5.6 For Multi-stakeholder initiatives -----	65
References -----	67

List of Acronyms

AI	Artificial Intelligence
AIRA	Africa Internet Rights Alliance
APJII	Indonesian Service Provider Association
BAKE	Bloggers Association of Kenya
BD	Brčko District
BiH	Bosnia and Herzegovina
CEDAW	Committee on the Elimination of Discrimination against Women
CSO	Civil Society Organisation
CSRG	Civil Society Reference Group, Kenya
EU	European Union
EIT Act	Electronic Information and Transactions Law
FBiH	Federation of Bosnia and Herzegovina
ICCPR	International Covenant on Civil and Political Rights
ICERD	Convention on the Elimination of All Forms of Racial Discrimination
ICT	Information and Communications Technology
LGBTQI+	Lesbian, Gay, Bisexual, Trans, Intersex, Queer, and Other
MOCI	Minister of Communication and Informatics of the Republic of Indonesia
MRS	Ministerial Regulation 5/2020
NCIC	National Cohesion and Integration Commission, Kenya
NSC	National Steering Committee on Peacebuilding and Conflict Management, Kenya
OHCHR	UN High Commissioner for Human Rights
OSCE	Organisation for Security and Cooperation in Europe
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNHCR	United Nations High Commissioner for Refugees
RS	Republika Srpska

Executive Summary

- As part of the project Social Media 4 Peace, funded by the European Union (EU) and implemented by UNESCO in Bosnia and Herzegovina, Indonesia, and Kenya, research was conducted in these three pilot countries. The goal was to better understand the root causes, scale and impact of harmful content and the effectiveness of the regulatory and self-regulatory frameworks to address it.
- The three countries provide evidence of online hate speech and disinformation affecting human rights offline. The evidence is not comprehensive yet clear enough to raise serious concerns. Online gender-based violence is also reported as critical in the three countries.
- In the three countries, national legislation to address harmful content shows some degree of inconsistency in comparison to international standards, notably in relation to the protection of freedom of expression. The reasons for such inconsistency vary among countries.
- The effective enforcement of legal frameworks is uneven in all three countries. Social and cultural inequalities are often reproduced in government or judicial decisions, and vagueness in legislation opens space for discretionary decisions.
- Platform companies have offices in Indonesia and Kenya, but not in Bosnia and Herzegovina.
- In the three countries, there is a lack of transparency in how companies allocate the roles of moderation tasks, including the number of different language moderators and their trusted partners and sources. Companies do not process content moderation in some of the main local languages and community standards are not entirely or promptly available in local languages.
- Civil society organizations are active in all three countries to monitor, curb and respond to online harmful content, but currently they have no strong coalitions to cooperate on these activities. Kenya and Indonesia in particular have vibrant organizations and seemingly a fruitful collaborative environment. However, the relations between CSOs and social media companies need to be consolidated.

Main findings

- // Online harmful content, in particular hate speech, disinformation and gender-based violence, affects the offline world and has a negative impact on peacebuilding in the three target countries. However, the lack of transparency on the moderation of such content by social media companies creates dependence on anecdotal evidence.
-

- // The preconditions to ensure that social media companies undertake content moderation that considers local contexts are not yet in place.
-

- // Existing legislation is often being used to restrict legitimate rights, notably freedom of expression, while at the same time not sufficiently protecting vulnerable groups.
-

- // Tensions arising from countries' historical and political contexts are often reinforced by social media dynamics.
-

- // Adherence to international standards to curb online harmful content on social media while protecting freedom of expression should be strengthened. At the same time, discussions are needed on the interpretation of these standards as they apply to the information ecosystem of social media, characterized by speed and volume of circulation of potential harmful content.

Recommendations

- // Thirty-four recommendations are presented at the end of this report for international organizations, states, social media companies, civil society, donors and multi-stakeholder actions.

Introduction

This report was produced to inform the implementation of the project entitled Social Media 4 Peace funded by the European Union (EU) and implemented by UNESCO. The overall objective of the project is to strengthen the resilience of societies against potentially harmful content that is spread online, particularly hate speech that incites violence; the project also seeks to enhance the promotion of peace through digital technologies, notably social media, in conflict-prone environments.

This report considers specifically actions carried out during the first year of the project's implementation in three pilot countries – Bosnia and Herzegovina, Indonesia, and Kenya – in order to better understand the root causes, scale, and impact of potentially harmful content and the effectiveness of the tools used to address it.

During the first year of project implementation, two research assignments were undertaken in each country. UNESCO partnered with Media Centar (Bosnia and Herzegovina), Build Up (Kenya), and Center for Digital Society (Indonesia) to work on understanding the legal frameworks, the adopted forms of regulating harmful content, and the trends and concerns regarding the implementation of these laws, their effectiveness to protect the targets of harmful content, the loopholes, and the impact on freedom of expression in each country.

At the same time, UNESCO partnered with ARTICLE 19 to research the current status of content moderation in each country, particularly the self-regulatory framework and tools put in place by social media companies to curb harmful content, and their effectiveness. ARTICLE 19 simultaneously mapped the local stakeholders' actions at the national level to help curb harmful content and gathered recommendations for the creation of local multi-stakeholder coalitions to provide local expertise and to ensure local dialogues on freedom of expression and content moderation.

The goal of this global report is to gather the main findings of this research and to formulate recommendations derived from them. The report also contributes to the implementation of the United Nations Strategy and Plan of Action on Hate Speech, which has identified a series of priority areas for monitoring and analyzing hate speech, stipulating that relevant UN entities should be able to 'recognize, monitor, collect data and analyze hate speech trends.' When focusing on online hate speech, UN entities are encouraged to promote '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' as well as to 'map the emerging risks and opportunities related to the spread of hate speech posed by new technologies and digital platforms', and 'define action protocols that account for the new forms of digital hate speech.'

This report further provides a section on international standards, to help analyze how best to balance freedom of expression with restrictions intended to protect people's rights and to prevent 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence', as stated in the International Covenant on Civil and Political Rights (ICCPR). In a digital world where harmful content is spreading rapidly and at a massive scale online, the interpretation of international standards may need updating to ensure the protection of people's rights.

The importance of analyzing the digital realm in conflict-prone countries

Various countries experience prolonged crises that stem from historical, social and cultural conditions and characteristics. In 2020, 95 sociopolitical crises were identified, spread across four continents, in Africa (38 cases), Asia (25), the Middle East (12), Europe (10), and Latin America (10). At the end of that year, the UN High Commissioner for Refugees (UNHCR) registered 82.4 million people forcibly displaced worldwide. In these countries, peacebuilding is a permanent process that depends on tackling conditions bequeathed by the past, while at the same time seeking the fulfilment of human rights in the present.

The digitalization of our societies has also shifted conflict dynamics across the world. The increase in connectivity, the popularization of smart phones and the capacity for drawing attention and grouping personal and public issues on the same platforms have made social media key to our social, cultural, political and economic life.

Looking at this scenario from the perspective of human rights and peacebuilding, this new information environment has brought both opportunities and threats. On the one hand, social media platforms opened space for more people to share their opinions and ideas. That represents not only the promotion of freedom of expression and access to information, but also mobilization and participation. The public sphere is enriched with vibrant cultural expressions, intense exchanges between citizens, and all the social, economic and political opportunities that stem from these.

On the other hand, business models and service models create structural incentives to spread disinformation and hate speech. The attention economy is driven by the engagement of users, and content spreading is dependent mainly on the endorsement of those close to the user. This model creates social-validation feedback loops. Fragmentation and segmentation have reinforced strong group affiliations and isolation from diversity and contradictory perspectives. A vicious cycle ensues in which public interest criteria (pluralism, diversity, credibility, common understanding) are substituted by private and self-interested ones, which are reinforced by users' confirmation biases.

In this scenario, discourses that mobilize negative emotion (such as fear, anger or resentment) are 'rewarded' with likes, shares and forwards. This makes false or misleading news travel faster than true news. This is worsened by coordinated actions that exploit fake accounts and click or troll farms for profit.

Thus, both as a negative externality of its openness and as a consequence of economically driven choices, this new information environment has led to increased circulation of hate speech, gender-based violence and disinformation. The growing force of disinformation is especially troubling in key time periods when correct information becomes a matter of life and death, as happened during the Covid-19 pandemic. The trend is also worrying because of its capacity to affect other fundamental rights in both individual and collective dimensions.

Yet because social media can reinforce social and cultural trends, it can also be used to reduce the deleterious effects of historical problems experienced in conflict-prone countries. Understanding the root causes, scale and impact of potentially harmful content is a necessary step the better to propose effective measures to mitigate it.

False or misleading information travels faster than true information.



Chapter 1: International standards on freedom of expression and the legitimate restrictions of harmful content

1.1 What is harmful content and what are its limits



The use of the internet has significantly increased access to information in recent decades. There is no question that today we have much more available information that can be accessed in different formats by different types of audiences. This represents an important tool for strengthening the enjoyment of Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which states that ‘Everyone shall have the right to hold opinions without interference’ and ‘Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.’ At the same time, the internet has also opened the door to increased circulation of harmful content through the same channels that provide reliable information. This ambivalence certainly constitutes a considerable challenge.

We can view this growing presence of harmful content on social media as a violation of Article 19 in two ways. First, by ignoring the special duties and responsibilities prescribed by the Covenant, it affects ‘the rights or reputations of others’ and ‘the protection of national security or of public order (ordre public), or of public health or morals’. Second, in an attempt to contain the first effect, many national laws and regulations, corporate norms, and attitudes have created an environment in which attempts to curb online harmful content result in the potential violation of human rights, particularly freedom of expression. Unclear and vague definitions of the limits and boundaries of harmful content are the main reasons that such tension is generated.

‘Potentially harmful content’ is indeed a broad term that encapsulates many different formats of content (text, video, images, audio, scripts, etc.) that can cause damage. It refers to illegal content as well as to content that is not necessarily unlawful, but that might nevertheless have detrimental effects on individuals or groups (‘lawful, but awful’). This includes misinformation, disinformation, hate speech, cyberbullying and so on.



UNESCO’s World Trends in Freedom of Expression and Media Development Global Report 2021/2022 found evidence of 57 laws across 44 countries adopted or amended in the last five years.

‘Potentially harmful content’ is a broad term that includes illegal but also legal content, i.e., misinformation, disinformation, hate speech, and cyberbullying.

Given such a broad scope, this report focuses on two types of harmful content that can either be legal or illegal depending notably on the context and the intent of such speech: disinformation and hate speech. The report looks at how such harmful speech is catalysed by the online environment, particularly through social media, and how national authorities and global actors – the social media companies – are balancing freedom of expression with the curbing of such content, in a peacebuilding perspective.

UNESCO defines disinformation as 'information that is false and deliberately created to harm a person, social group, organization or country'. According to the UN Special Rapporteur on freedom of opinion and expression Irene Kahn, disinformation can be understood 'as false information that is disseminated intentionally to cause serious social harm and misinformation as the dissemination of false information unknowingly.' The European Commission has defined disinformation as 'false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit' and misinformation as 'misleading or inaccurate information shared by people who do not recognize it as such.'

This is not a new problem. The deliberate dissemination of false information to generate cultural or political effects is an age-old issue that predates the internet and even journalism. Disinformation has taken many forms throughout history (such as defamation, rumours, false correlations between facts, fictional narratives, etc.) and has been spread through various media types (oral, written, audiovisual communication). In the 21st century, disinformation has been marked by three important aspects that are directly related to the digital environment:

- // (a) Reach: it is a global problem unfettered by borders in both its production process and dissemination;
- // (b) Scale: it has enormous volume with characteristics similar to industrial production and can circulate through different kinds of media at the same time; and
- // (c) Sophistication: its production involves increasingly complex and refined techniques intended to simulate true information. Disinformation can emulate journalistic texts by mixing true and false information, producing videos or audios that imitate voices and images using techniques like deep fake, or promoting coordinated actions that looks like spontaneous communication.

+

Another important type of harmful content that directly impacts peacebuilding is hate speech. There is no exact, universally accepted definition for understanding hate speech. In the UN Strategy and Plan of Action on Hate Speech, the concept is defined as 'any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language concerning 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.' It also states that 'this is often rooted in, and generates intolerance and hatred and, in certain contexts, can be demeaning and divisive.'

The UNESCO study Countering Online Hate Speech states that ‘in common parlance, however, definitions of hate speech tend to be broader, sometimes even extending to encompass words that are insulting those in power, or derogatory of individuals who are particularly visible. Especially at critical times, such as during elections, the concept of hate speech may be prone to manipulation: accusations of fomenting hate speech may be traded among political opponents or used by those in power to curb dissent and criticism.’

International human rights law standards provide us with important elements that make it possible to determine the definition of illegal hate speech that may trigger legitimate limitations to freedom of expression. Article 20 of the ICCPR states that ‘any propaganda for war shall be prohibited by law’ and that ‘any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.’



Especially in this last passage, three words are important: ‘hatred’, ‘advocacy’ and ‘incitement’. The first one means an extreme feeling of negative bias towards groups or individuals. The latter two refer to the form and the objective: to defend, incite or stimulate implies amplifying negative feelings through speech, potentially generating violence, discrimination and hostility that can cause real damage to human rights. The term ‘discrimination’ needs to be understood as something dynamic and must incorporate the different groups that suffer from hate speech in our time. This goes beyond national, racial or religious status and includes issues involving sexual orientation and gender identity; it can, depending on the context, include phenotypes, biological features, disabilities, refugee status and other characteristics.

The International Convention on the Elimination of All Forms of Racial Discrimination requires States Parties to ‘declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof.’ In other words, the dissemination itself of these ideas is illegal. The Convention also requires States Parties to ‘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’ and ‘not permit public authorities or public institutions, national or local, to promote or incite racial discrimination’.

Additional relevant documents include the Convention on the Prevention and Punishment of the Crime of Genocide and the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW).

Although the dissemination of harmful content such as disinformation and hate speech can be detrimental to a set of individual and collective rights, not all that is defined as disinformation or hate speech should be deemed illegal, much less criminalized. As stated in the [Rabat Plan of Action](#) ‘criminal sanctions related to unlawful forms of expression should be seen as last resort measures to be applied only in strictly justifiable situations.’ The Plan makes it clear that a distinction should be made between three types of expression:

- // a) Expression that constitutes a criminal offence – those specifically provided for in articles 19 and 20 of the Covenant, for example, hate speech or disinformation that incite acts of violence, discrimination or harm to individuals or groups protected by law.
- // b) Expression that is not criminally punishable but may justify a civil suit or administrative sanctions – medium severity harmful content that causes damage but cannot be considered a crime.
- // c) Expression that does not give rise to criminal, civil or administrative sanctions, but still raises concern in terms of tolerance, civility and respect for the rights of others.

Merely defining content as harmful does not determine its treatment. Any restriction should be considered an exception and ‘in every case in which the State restricts freedom of expression it is necessary to justify the prohibitions and their provisions in strict conformity with article 19,’ as stated in the [Human Rights Committee general comment No. 34 para. 52. \(CCPR/C/GC/34\)](#). As for companies, they should comply with the [UN Guiding Principles on Business and Human Rights](#), which define their duties of respecting and promoting human rights and of remedying in cases of violations.



1.2 Protected rights and bounded restrictions

While freedom of opinion and expression and free speech are defined in broad, general terms, their restriction must be exceptional. The illegality of hate-related content is defined in Article 20 of the Covenant, and can be summarized as the simultaneous existence of the following characteristics specified in [general comment No. 34 \(2011\) of the Human Rights Committee](#): ‘first, only advocacy of hatred is covered; second, hatred must amount to advocacy which constitutes incitement, rather than incitement alone; and third, such incitement must lead to one of the listed results, namely discrimination, hostility or violence.’



The restrictions on freedom of opinion and expression must be exceptional and justified.

Therefore, it is important to clarify that not all harmful or inappropriate content constitutes a criminal offence. In their interpretation of the Convention (general recommendation No. 35 [2013]), the Committee on the Elimination of Racial Discrimination stated:

'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.'

The interpretation made by the Human Rights Committee of general comment No. 34 (2011) is mentioned in the 2019 report on the Promotion and protection of the right to freedom of opinion and expression, made by the former UN Special Rapporteur on freedom of opinion and expression, David Kaye. Referring to the CCPR/C/GC/34, he clarifies specific types of situations and discourse that cannot be considered criminal offences:

- //** a) Disrespectful speech – ‘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.’

- //** b) Attacks on religion – ‘Prohibitions of displays of lack of respect for a religion or other belief system, including blasphemy laws, are incompatible with the Covenant, except in the specific circumstances envisaged in article 20, paragraph 2, of the Covenant.’

- //** c) Interpretation of past events - Laws that ‘penalize the expression of opinions about historical facts are incompatible’ with Article 19 of the Covenant. ‘The Covenant does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events.’

- //** d) Declarations of prejudice –‘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.’

In other words, speech that expresses hatred, anger, rancour, reprobation, blasphemy or disagreement may be reprehensible, undesirable, uncivil, or inappropriate, but not necessarily illegal. 'Such advocacy becomes an offence only when it also constitutes incitement to discrimination, hostility or violence, or when the speaker seeks to provoke reactions on the part of the audience' that violate human rights and those of protected groups, as stated in the 2012 report of the former Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue ([A/67/357](#)).



When dealing with harmful content, the State must act to guarantee the right to freedom of opinion, expression and thought as a broader precept; be specific and clear in the application of criminal restrictions that can only be applied in strictly justifiable situations; and 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' ([A/HRC/22/17/Add.4](#)).

Therefore, the State needs to be cautious when applying sanctions and should note that any limitations must meet three conditions, enlisted in David Kaye's 2019 report ([A/74/486](#)). First, ***legality*** involves the formal aspects necessary to ensure that '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.' Second, ***legitimacy*** is linked to the backing of existing reasons capable of justifying the penalty. '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 or reputations of others or to protect national security, public order, public health or morals'. Third, ***necessity and proportionality*** is related to the intensity and adequacy of punishment. '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'.



Noting this set of factors is necessary to ensure that the remedy against disinformation and hate speech is applied correctly using the proper dosage and avoiding unwanted side effects, including the violation of international human rights.



We must also remember that violations of freedom of expression and the handling of harmful content do not occur only on the internet. The concepts of hate speech and disinformation and the sanctions for incitement as provided for in the Covenant must be applicable to any media, whether online or offline. As stated by former UN rapporteur David Kaye, 'penalties on individuals for engaging in unlawful hate speech should not be enhanced merely because the speech occurred online.' Therefore, the State should start with two premises when addressing online hate speech: '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 extra-legal threats – that intermediaries take action that international human rights law would bar States from taking directly.'

In order to determine a practical and concrete procedure for the application of Article 20 of the International Covenant on Civil and Political Rights, the Rabat Plan of Action stipulates that it is necessary to establish a high threshold for defining restrictions on freedom of expression. To this end, six factors must be taken into account:

● // a) Context: Every communication action can be properly interpreted only if it is analysed within the social, cultural and political context prevalent at the time it was produced and spread. Context is critical to assessing whether content may incite discrimination, violence or hostility.

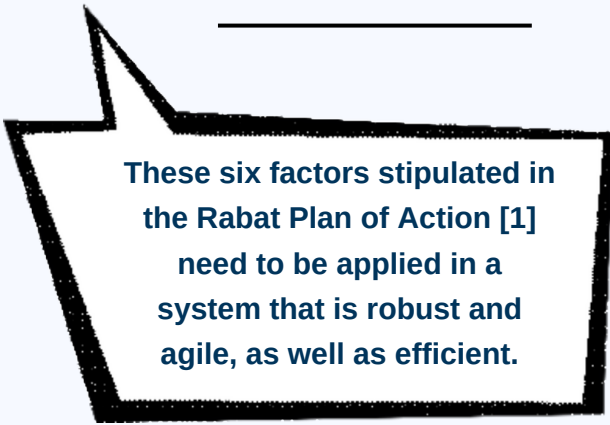
● // b) Speaker: The dissemination of discourse can have different effects, interpretations and impacts depending on who the speaker is. This element needs to be evaluated to measure the severity and scope of harmful content. 'The speaker's position or status in the society should be considered, specifically the individual's or organization's standing in the context of the audience to whom the speech is directed' (A/HRC/22/17/Add.4).

● // c) Intent: In order for speech to be considered a violation of human rights according to international standards, it is necessary to prove the existence of advocacy for national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. This is relevant mainly in the digital environment where there is often impulsivity and imprudence in the act of sharing content. 'Article 20 of the International Covenant on Civil and Political Rights anticipates intent. Negligence and recklessness are not sufficient for an act to be an offence under article 20 of the Covenant, as this article provides for "advocacy" and "incitement" rather than the mere distribution or circulation of material. In this regard, it requires the activation of a triangular relationship between the object and subject of the speech act as well as the audience' (A/HRC/22/17/Add.4).

● // d) Content and form: Discourses are complex actions that can mobilize meanings through different symbols and languages. Therefore, a detailed analysis of the form and meanings inherent to the content is a fundamental aspect of measuring and judging the real objectives of the speaker. This analysis is a critical element in proving whether there was incitement to discrimination, hostility or violence and ascertaining the qualification of the offence: 'Content analysis may include the degree to which the speech was provocative and direct, as well as the form, style, nature of arguments deployed in the speech or the balance struck between arguments deployed' (A/HRC/22/17/Add.4).

● // e) Extent of the speech act: In some cases, speech addresses a small group of people and has a limited effect. In other situations, it can reach a huge audience and generate significant impact, causing unrest and violence on a large scale. This is related to the technical sophistication used in the speech act. There is a big difference between harmful content published in a simple leaflet distributed in small quantities, and harmful content spread digitally using mass dissemination technologies and automated systems like bots. The use of participatory tools that enable the audience to replicate the incitement more efficiently is also an element linked to the magnitude of the discourse. Therefore, the material, professional and technical investment used to increase the dissemination of harmful content needs to be considered when assessing its severity.

● // f) Likelihood, including imminence: In order to prove the existence of incitement, it is not necessary to prove the existence of damage. The incitement can be unsuccessful, but if it did exist, even without having achieved its objectives, it must be punished. 'Incitement, by definition, is an inchoate crime. The action advocated through incitement speech does not have to be committed for said speech to amount to a crime. Nevertheless, some degree of risk of harm must be identified. It means that the courts will have to determine that there was a reasonable probability that the speech would succeed in inciting actual action against the target group, recognizing that such causation should be rather direct' ([A/HRC/22/17/Add.4](#)).



These six factors stipulated in the Rabat Plan of Action [1] need to be applied in a system that is robust and agile, as well as efficient.

Complementary to State regulation, the Rabat Plan also provides for the self-regulation of companies, which must ensure '(a) Taking care to report in context and in a factual and sensitive manner, while ensuring that acts of discrimination are brought to the attention of the public. (b) Being alert to the danger of furthering discrimination or negative stereotypes of individuals and groups in the media. (c) Avoiding unnecessary references to race, religion, gender, and other group characteristics that may promote intolerance' ([A/HRC/22/17/Add.4](#)).

The [Rabat Plan of Action](#) was approved in 2012, at the dawn of the social media era. Consequently, States face many challenges when applying its standards to the online world of today. The same holds true for social media companies, which have progressively taken on a regulatory role.

[1] Also explained in this short UNESCO video: <https://www.youtube.com/watch?v=ADrB32OSe3A>

The Office of the UN High Commissioner for Human Rights (OHCHR) recently (April 2022) published its [report to the Human Rights Council on The practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies](#). It establishes that ‘When undertaking human rights due diligence, companies should pay special attention to any particular human rights impacts on individuals from groups or populations that may be at heightened risk of vulnerability or marginalization, such as children, ethnic minorities, members of the lesbian, gay, bisexual, transgender and intersex community and human rights defenders, and to keep in mind gender-based risks and impacts.’



1.3 Paths and challenges

The dissemination of hate speech and disinformation occurs in an online environment that has grown increasingly complex with the centrality of social media and platforms for public life, the growing digital culture, and the intensive use of artificial intelligence. As stated in [Countering Online Hate Speech](#), ‘while hate speech online is not intrinsically different from similar expressions found offline, there are peculiar challenges unique to online content and its regulation. Those challenges related to its permanence, itinerancy, anonymity, and cross-jurisdictional character are among the most complex to address.’

In this online environment, different perspectives and multiple actions coexist in order to reduce the negative impacts of the online dissemination of harmful content. However, they are not always convergent and synergistic and may often be inconsistent with international human rights law.

To minimize these problems and move towards building a global system that can operate in better alignment with international standards, several aspects need to be addressed. Additional major tensions and challenges around this problem can be found in four crucial dimensions: uniformity, intelligibility, diversity of players and strategies, and applicability.

▶ ▶ ▶ 1.3.1 Lack of uniformity in definitions

Lack of uniformity is a key aspect that affects actions against the spread of harmful content. First, the definition of hate speech differs across countries and regions. Usually, each country ‘has a slightly different approach to how it defines hate speech in terms of how it is expressed, who the potential targets are and what kind of harm has to happen for speech itself to be considered hateful. The lack of a unified definition is one of the major challenges when it comes to combating online hate speech, which is not necessarily confined to national borders.’ [2]

[2] CI/FEJ/2021/DP/01, UNESCO/UN Office on Genocide Prevention and the Responsibility to Protect, 2021, Addressing Hate Speech on Social Media: Contemporary Challenges, Paris, UNESCO, p. 2.

Nevertheless, the Human Rights Committee states there should be no margin of appreciation when invoking restrictions for freedom of expression. According to the Human Rights Council, to analyse the need for restrictions in a given circumstance, '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 that has caused it to restrict freedom of expression' ([CCPR/C/GC/34](#)).

This lack of uniformity in understanding hate speech and violations of freedom of expression generates broad concepts that give rise to differing interpretations. [3] According to the Rabat Plan, this produces violations of international human rights law in two ways. On one hand, the non-prosecution of 'real' incitement cases and, on the other, the persecution of minorities in ways that violate their legitimate right to freedom of expression, under the guise of domestic incitement laws. As stated in the Plan: 'Anti-incitement laws in countries worldwide can be qualified as heterogeneous, at times excessively narrow or vague. Jurisprudence on incitement to hatred has been scarce and ad hoc, and while several States have adopted related policies, most of them are too general, not systematically followed up, lacking focus and deprived of proper impact assessments' ([A/HRC/22/17/Add.4](#)).



Lack of uniformity is evident both in states' definitions of various types of harmful content and the ways companies approach the issues of harmful content, including when it comes to design choices behind digital platforms.

This heterogeneity is also present in how companies are treated in attempts to solve the problem. Internet intermediaries have developed different definitions of hate speech and guidelines to regulate it. The term 'hate speech' is not always mentioned, as some companies prefer to act upon clear conducts or behaviour, as mentioned in the UNESCO 2021 report "[The Hate Speech Policy of "Major" Platforms during the Covid-19 Pandemic.](#)" Each company has its own definition of what is harmful content, often removing content without compliance with the law. The report shows that the pandemic changed companies' norms and practices regarding hate speech moderation, leading to more detailed policies and stricter applications. However, the lack of uniformity did not change, and it affects the way each company reports its actions and its modus operandi, making it impossible to have a global view of the problem. As stated in [UNESCO's report on transparency for platforms](#): 'Information about actual practices, not only of moderation, but especially of curation, the approach to trade-offs between rights, and the role of company interests, is usually less forthcoming....At present, the levels of transparency do not generally allow for the possibility for verification of the data presented; therefore, much depends upon what the companies choose to share, and how they interpret it, which reflects largely how they wish to set the agenda of debate.'

[3] See I. Gagliardone, D. Gal, T. Alves and G. Martinez, 2015, *Countering Online Hate Speech*, Paris, UNESCO, pp.23-26.

This heterogeneity is also exacerbated by differences in the designs of each platform. ‘The architectures on which these platforms are based, however, may vary significantly and have important repercussions on how hate speech spreads and can be countered,’ as stated in [Countering Online Hate Speech](#).

That does not mean we should ignore that digital speed and scale, along with coordinated actions, can change the extent of the speech act, as defined in the Rabat Plan of Action. One piece of lawful but harmful content, even if it goes viral, does not have the same effect as thousands of different pieces reproducing the same discourse targeting specific groups. Incitement to harm is certainly influenced by a high volume of similar content spread over a short period of time, especially when involving coordinated action.

1.3.2 Lack of transparency

The second crucial dimension for dealing with harmful content is related to the availability of information needed to understand its aspects, dynamics and way of functioning. An effective solution to a problem is possible only when it becomes intelligible: that is, we have enough knowledge to attack its weak points. This is a relevant premise because ‘efforts to understand hate speech not primarily with the instrumental goal to counter or eliminate it, but also to grasp what it is the expression of, are particularly difficult – yet continue to be very important’ ([Countering Online HateSpeech report](#)).

To this end, the [Rabat Plan of Action](#) stipulates that ‘States should ensure the necessary mechanisms and institutions in order to guarantee the systematic collection of data in relation to incitement to hatred offences’. However, putting this into practice is still very problematic. As the discussion paper [‘Addressing Hate Speech on Social Media: Contemporary Challenges’](#) by UNESCO and the UN Office on Genocide Prevention and the Responsibility to Protect indicates, ‘From a technological perspective, online hate speech is difficult to study due to the inconsistent reliability of detection systems, opaque nature of proprietary algorithms, lack of access to data held by companies and so forth. Clarity on how these challenges can be addressed is indispensable for producing further understanding of how online hate speech emerges and proliferates, and subsequently for formulating effective responses’.

Although there have been significant advancements, such as companies publishing more meaningful transparency reports in recent years, the scenario is still far from ideal. Effective monitoring and access to disaggregated data depends on companies’ goodwill. For example, according to the [transparency reports of social media companies](#), ‘it is clear that in 2020, so-called hate speech and the removal of such posts grew significantly on social media. There is not enough disaggregated data to understand what each platform classifies as hate speech, the decision-making processes, and error rates, making it more difficult to understand the root causes of this growth.’

Data availability is also affected by such factors as:

- // a) the predominance of English in detection methods and existing tools that cannot always operate in more than one language;
- // b) the vast majority of research and monitoring of hate speech on social media platforms is concentrated in a few places like the United States and Europe;
- // c) methodological issues like the different definitions used to frame the phenomena, the social and historical contexts, the linguistic subtleties, the variety of online communities, and the forms of online hate speech (text, videos, images, audios, etc.);
- // d) concerns over privacy and the potential misuse of user data; [4]
- // e) the huge and growing volume of data produced every minute, which demands sophisticated big data tools able to capture and analyse at the same speed as content is produced and spread; and
- // f) the increasing use of automated systems like machine learning and deep learning that are inherently opaque.

This set of issues demonstrates the enormous task of developing a system based on transparency and accountability to monitor and track the dissemination of online harmful content. It also shows that this processed information should additionally provide feedback in order to leverage changes in algorithmic governance, and thus enhance the system and deliver better outputs over time. An intelligence system capable of preventing excesses while respecting international standards for freedom of expression can be created with tools designed to respect different social contexts and to be open to public scrutiny and permanent collaborative development that accompanies the problem's evolution.

Companies need to spell out the concepts behind their approach to content moderation in order to make their actions more transparent and intelligible for all interested parties. For example, the UN Special Rapporteur on freedom of opinion and expression has suggested that social media companies develop a human rights-compliant framework for handling online hate speech by answering the following questions:

- **What are protected persons or groups?**
- **What kind of hate speech constitutes a violation of company rules?**
- **Is there specific hate speech content that the companies restrict?**
- **Are there categories of users to whom the hate speech rules do not apply?**



[4] See UNESCO, 2021, *Letting the Sun Shine In: Transparency and Accountability in the Digital Age*, Paris, UNESCO; and UNESCO/UN Office on Genocide Prevention and the Responsibility to Protect, 2021, *Addressing Hate Speech on Social Media: Contemporary Challenges*, Paris, UNESCO.

Currently, there is little clarity and transparency in how companies operate and how they deal with potential violations in their content moderation process. 'There is a significant barrier to external review (academic, legal and other) of hate speech policies as required under principle 21' and stipulated in the [OHCHR's 2011 Guiding Principles on Business and Human Rights](#): 'In order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders. Business enterprises whose operations or operating contexts pose risks of severe human rights impacts should report formally on how they address them.'

The lack of information about the criteria companies use in their content moderation process keeps us from understanding how far companies have gone to implement the aspects listed by the Rabat Plan of Action (context, speaker, intent, content and form, extent of the speech act, likelihood).



1.3.3 Concentration of power and decision-making



Currently, the spread of disinformation and hate speech is a complex, two-pronged phenomenon. On one side, we have conceptual and cultural aspects: the concept of truth, narrative disputes, cultural and psychological elements, and abstract and symbolic aspects that are underscored by philosophical, social and political beliefs. On the other, we have economic aspects, especially relevant for the pervasiveness of digital technology that is increasingly present in all areas of public life (concentration of power, oligopolies, data economy, ubiquity of autonomous systems based on artificial intelligence).

The requirements for dealing with complex issues are diverse points of view and the active participation of various stakeholders. Multiple aspects can thus be considered in the conceptual and procedural definitions, and greater efficiency and legitimacy in the problem-solving process can be produced. According to the [Rabat Plan of Action](#), 'States should have in place a public policy and a regulatory framework which promotes pluralism and diversity of the media, including new media, and which promotes universal and non-discrimination in access to and use of means of communication.' In addition to the role of States, the [Plan](#) also refers to a collective role: 'States, media and society have a collective responsibility to ensure that acts of incitement to hatred are spoken out against and acted upon with the appropriate measures, in accordance with international human rights law' and 'any related legislation should be complemented by initiatives from various sectors of society geared towards a plurality of policies, practices and measures nurturing social consciousness, tolerance and understanding change and public discussion.'

Today the dissemination of online harmful content has been widely debated and analysed by many actors, but there is a clear concentration of power in the hands of a small number of companies, due to their market domination and the lack of regulations in this field. This has 'resulted in scant attention being paid, and little budget share allocated, towards monitoring challenges, creating guardrails like independent oversight, or commissioning human rights impact assessments....Challenges arise from the logics of the companies' architectures, and the role of users, although this does not mean there is a symmetry of power or obligation on the two sides.'

[5]

In many cases, companies have acted as summary courts, with the power to suppress speech based only on their own criteria. To prevent such arbitrary judgements, solutions may come from regulation defining public criteria and independent supervising processes, but also from discussions with affected groups and relevant stakeholders. As stated in the 2019 report from the former UN Special Rapporteur on freedom of opinion and expression (A/74/486): '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.' In addition, social media companies would need to consider external scrutiny of their structures and processes, as provided for in principle 18 of Guiding Principles on Business and Human Rights: 'a) Draw on internal and/or independent external human rights expertise; b) Involve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation'.



The lack of regulation and market domination enables the concentration of power in the hands of a small number of companies that can suppress speech on their own criteria.

▷ ▷ ▷ 1.3.4 Effectiveness and enforcement

Finally, the fourth crucial dimension for dealing with harmful content is related to the practical aspect of the problem: enforcement. The architecture of the digital environment has introduced new aspects to disinformation and hate speech, namely volume, speed, various types of content, super users (influencers), sophistication in the production of disinformation, anonymity, cross-jurisdictional flow, multiple stakeholders, new governance parameters, and others. Clearly, the effectiveness of any rights protection system faces numerous challenges.

[5] See UNESCO, 2021. [Letting the Sun Shine In: Transparency and Accountability in the Digital Age](#), Paris, UNESCO, p. 5.

The Rabat Plan stipulates that States should adopt comprehensive anti-discrimination legislation that includes preventive and punitive actions to effectively combat incitement to hatred. Punitive actions that restrict freedom of expression need to correspond to the real existence of a criminal offense. Therefore, it is important during the process to observe the criteria that are applied and their gradations. 'Between introducing new and potentially intrusive regulation of content, and a completely laissez-faire approach, a third way is increasingly being proposed: to focus more on issues of process, rather than content, and especially to focus on greater transparency of the processes used by the platform companies.' [6]

One of the biggest challenges is the enormous volume of harmful content produced daily that circulates rapidly on a large scale. The social and political impact of this massive phenomenon has put States in a position to demand more forceful actions from companies, whether through liability laws or the threat of banning their services. As for companies, they run automated systems in the moderation process that are not always able to detect degrees of severity and levels of corresponding penalties, which generates other violations during the process. There is pressure to implement automated tools that would serve as a form of pre-publication censorship. The former UN Special Rapporteur on freedom of opinion and expression, in his report on hate speech, states that an upload filter requirement for this kind of content '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.'

Between difficulties from the private and the public sector, individuals and groups affected by hate speech lack effective mechanisms for defending their rights. As stated in the Rabat Plan of Action: 'In many instances, victims are from disadvantaged or vulnerable groups and case law on the prohibition of incitement to hatred is not readily available. This is due to the absence or inadequacy of legislation or lack of judicial assistance for minorities and other vulnerable groups who constitute the majority of victims of incitement to hatred. The weak jurisprudence can also be explained by the absence of accessible archives, but also lack of recourse to courts owing to limited awareness among the general public as well as lack of trust in the judiciary.'

In this context, the debate on how to create effective and enforceable systems of dealing with harmful content gains more relevance, as does the question of formulating the clear definition of the role of the different stakeholders. This requires improvement in the following areas:



Public officials need to be trained and they must understand their responsibilities in law enforcement. The Rabat Plan stipulates that 'States should build the capacity to train and sensitize security forces, law-enforcement agents and those involved in the administration of justice on issues concerning the prohibition of incitement to hatred.' At the same time, politicians, government officials and public figures '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.'

[6] See UNESCO, 2021. Letting the Sun Shine In: Transparency and Accountability in the Digital Age, Paris, UNESCO, p. 7.



Companies need to play an active but limited role in countering harmful content, based on the application of transparent and intelligible criteria. As stated by the former UN Special Rapporteur on freedom of opinion and expression David Kaye, ‘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.’

As stated in the Rabat Plan of Action, ‘Any related legislation should be complemented by initiatives from various sectors of society geared towards a plurality of policies, practices and measures nurturing social consciousness, tolerance and understanding change and public discussion....States, media and society have a collective responsibility to ensure that acts of incitement to hatred are spoken out against and acted upon with the appropriate measures, in accordance with international human rights law.’



The response to violations must be applied quickly, but in a safe and legitimate way. Summary and hasty judgments of potentially harmful content can lead to violations of the right to freedom of expression. On the other hand, delayed actions can amplify the damage of potentially harmful content. ‘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.’ [7]



The application of artificial intelligence in content moderation should be balanced with a human-based approach and underpinned by international standards. Algorithms are not neutral. Codes are structures of repeated values and concepts; when used to moderate hate speech, they should be designed to incorporate the aspects listed in the Rabat Plan of Action (context, speaker, intent, content and form, extent of the speech act, likelihood). Despite the sophistication of autonomous digital systems, there will always be inaccuracies that can generate rights violations. As stated in the Santa Clara Principles [8] on Transparency and Accountability in Content Moderation, ‘companies should only use automated processes to identify or remove content or suspend accounts, whether supplemented by human review or not, when there is sufficiently high confidence in the quality and accuracy of those processes.’ Therefore, every sanction application must have human evaluation as the main revision-making instance. This process should take into account contextual knowledge, as stated by the former UN Special Rapporteur on freedom of opinion and expression (A/74/486):

‘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.’

[7] See A/74/486, para. 48.

[8] Principles for meaningful transparency and accountability around Internet platforms’ content moderation, developed by fourteen civil society organizations and endorsed by twelve major companies.

On the other hand, 'whilst manual approaches have the distinct advantage of capturing context and reacting rapidly to new developments, the process is labor-intensive, time-consuming and expensive, limiting scalability and rapid solutions.' [9] This has led to creating more sophisticated techniques for automated analysis: machine learning, natural language processing, keyword-based approaches, distributional semantics, sentiment analysis, source metadata and deep learning. [10]



The application of sanctions must be appropriate to the severity. 'Companies should have graduated responses according to the severity of the violation or the recidivism of the user'. Deleting content and muting the speaker should be a last resort, an exception applied only in accordance with article 20 of the Covenant. In this sense, '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....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.' [11]



The occurrence of violations must be accompanied by options for remedy. 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.' International human rights standards list several possibilities and forms of remediation. '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.' [12]

These four crucial dimensions of uniformity, transparency, concentration of power and decision-making, and effectiveness and enforceability, contain fundamental tensions and challenges that surround the spread of harmful online content. They must be addressed in an integrated and collaborative way to build systems aligned with international standards for the guarantee and protection of human rights.

[9] See UNESCO/JN Office on Genocide Prevention and the Responsibility to Protect, 2021, *Addressing Hate Speech on Social Media: Contemporary Challenges*, Paris, UNESCO, p.5.

[10] Ibid.

[11] See A/74/486, para. 50-54.

[12] See A/74/486, para. 53.

Chapter 2: Overview of country reports

2.1 Bosnia and Herzegovina

▶▶▶ 2.1.1 Context

Internet penetration and the number of social media users in Bosnia and Herzegovina have been rising consistently in recent years. In 2020, the internet penetration rate was around 95% of the country's approximately 3.5 million residents, a sharp increase when compared to 2011 when that percentage was 55%. Almost two-thirds of households (72.8%) have internet access, and close to 90% of the population has access to the internet via mobile phones. [13] Despite the country's geographical and ethnic diversity, there is no significant digital divide, even when considering gender and rural communities. Studies also show that 52% of the local population report accessing information through social media, and 73% use at least one social media platform. The most popular is Facebook, used by three-quarters of the adult population (73%), followed by Instagram (39%), YouTube (38%), TikTok (8%), Snapchat (8%), Twitter (4%), Pinterest (2%) and LinkedIn (2%). [14]

In this online ecosystem, the spread of harmful content has become a growing issue with significant impacts on the political and social dynamics in Bosnia and Herzegovina (BiH). This is particularly true of inter-ethnic hatred content, war crime and genocide denial, glorification of war criminals, inflammatory narratives, disinformation campaigns, and gender-based harassment and discrimination against under-represented groups like the LGBTIQ+ community and refugees, [15] especially during pre- and post-election periods, as in 2018 and 2020, when an intensive frequency of hate speech was identified. [16] In 2020 and 2021, the Covid-19 pandemic was associated with increased disinformation related to public health, in particular disinformation designed to discourage people from getting vaccinated.

Internet penetration in BiH

55% in 2011

95% in 2020



73% of Facebook users



39% of Instagram users



38% of YouTube users



8% of TikTok users



8% of Snapchat users



4% of Twitter users



2% of Pinterest users



2% of LinkedIn users

[13] Data of the Communications Regulatory Agency in BiH.

[14] Council of Europe, 2021, [Media Habits of Adults in BiH](#).

[15] A. Sokol and M. Čalović, 2022, [Regulation of harmful content online in Bosnia and Herzegovina: Between freedom of expression and harms to democracy](#), Sarajevo, Mediacentar; and B. Kostić, 2022, [Content moderation on social media and map of stakeholders in Bosnia and Herzegovina](#), ARTICLE 19.

[16] Sokol and Čalović, op. cit.

Curbing such harmful content by regulation or self-regulation without restricting freedom of expression is challenging, as it necessitates understanding the country's political and historical context. The complexity results primarily from the historical process of building a national identity, which was permeated by severe ethnic and geopolitical conflicts leading up to the country's independence in the 1990s and the constitution of a democratic republic organized into two-state entities, the Federation of Bosnia and Herzegovina (FBiH) and Republika Srpska (RS), and the semi-autonomous Brčko District (BD). This political system generated political party structures divided according to ethnicity, resulting in a multipolarized social and political system that is strongly reflected in the digital sphere.

In this context, local reports reveal an environment where the freedom of expression of ordinary citizens is often limited and, at the same time, the circulation of harmful content is intense – including hate speech and glorification of war criminals – despite actions taken in recent years by authorities, civil society and social media companies to curb it.

▶ ▶ ▶ 2.1.2 Legislation addressing harmful content

In BiH, there are legal, regulatory and self-regulatory tools to use against harmful content, which includes hate speech and hate narratives, war crime denial and the glorification of war criminals, ethnonational and politically biased reporting, disinformation, defamation and threats, attacks, and smear campaigns. Most of these tools are found fragmented in different laws, are not entirely aligned with international standards, and are inconsistently applied in the online sector.

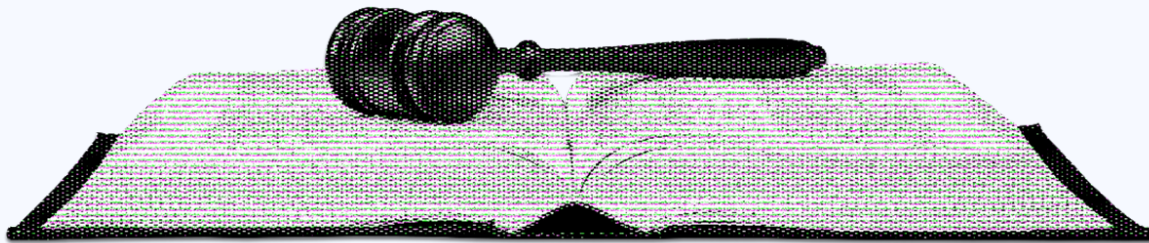
Freedom of expression is officially guaranteed by the Constitution of Bosnia and Herzegovina and by the constitutions of the two-state entities (FBiH and RS). Criminal codes prohibit public incitement and inflaming national, racial and religious hatred, discord or hostility among people. Hate speech is also prohibited by the Code on Audio-Visual Media Services of the Communications Regulatory Agency, although media regulatory tools do not apply to online content.

BiH and FBiH criminal codes, however, do not include other categories such as skin colour, gender, sexual orientation, disability, or others included in international standards [17] that are protected in the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) and the Council of Europe Convention on Cybercrime and its additional protocol, both ratified by BiH. Moreover, cases tried by the judiciary indicate that important cases may not be subject to prosecution, once criminal code sanctions are limited to violations committed directly (and only) against specific groups. If the targets are groups that are considered as non-specified (such as 'immigrants'), [18] the cases are outside the reach of the criminal code.

[17] RS, in particular, includes these other protected categories.

[18] Such as the Antimigrant.ba case, in which the verdict concluded that statements about migrants did not refer to any particular nation, race, religion or other specific group and were within the framework of a political, journalistic, free narrative, protected under freedom of speech (Sokol and Čalović, op. cit.)

Furthermore, hate speech cases brought to court are rare overall, particularly involving hate speech on the internet. This lack of engagement in protecting citizens from harmful content is also reflected in the low number of major court decisions related to online harmful content. Local data reveal that, so far, cases have been initiated mostly against ordinary citizens, leaving out the cases that involve powerful political figures.



More recently, after regional standards were published by the Council of the European Union in 2021, [19] the High Representative of BiH amended its Criminal Code to prohibit the public condoning, denial, gross diminution or justification of genocide, crimes against humanity and war crimes as determined by final court decisions that might incite violence or hatred against a group of persons or a member of such a group [20]. These amendments were considered an important step to 'restore mutual understanding about past events and toward a common future' by the UN Special Adviser on the Prevention of Genocide [21]. According to local reports, the implementation of the law is recent and the results observed in 2021 are not conclusive, considering a growing number of online incidents glorifying war crimes and convicted war criminals in early 2022. [22] Adopting the amendment of the Criminal Code prohibiting the justification and denial of digital, crimes against humanity and war crimes was very controversial – precisely because of the historically conflictive environment of the country – and the RS National Assembly adopted a Law of Non-Application of this decision. [23]

All these categories of crime addressed in the criminal codes appear to require regulatory instruments to better guide the sanctioning powers such as the judiciary when incorporating different standards of acceptability, tolerance, and proof. This should be done in accordance with the case law of the European Court of Human Rights, and have as references the International Covenant on Civil and Political Rights, the European Convention on Human Rights, and especially the Rabat Plan of Action.

[19] Council of the European Union. Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

[20] But only – as hate speech regulation – if it is directed against a group of persons or a member of a group defined by reference to race, colour, religion, national or ethnic origin, when the conduct is carried out in a manner likely to incite violence against such a group or one or more of its members.

[21] 23 July 2021. Note to Correspondents: Statement by Alice Wairimu Nderitu, Special Adviser on the Prevention of Genocide, on the introduction of amendments to the Criminal Code of Bosnia and Herzegovina.

[22] Srebrenica Genocide Denial Report, 2021. Researchers identified 234 instances of genocide denial online in Serbia (142), BiH (60), and Montenegro (19).

[23] In addition, other amendments were approved to stipulate prison sentences for publicly expressing ridicule, contempt or grossly disparaging Republika Srpska, its flag, coat of arms, emblem or anthem, which is also not in line with international standards.

In the civic sphere, there are laws against defamation and a specific law for election periods. The law against defamation is perceived by local stakeholders as too complex to be implemented, and cases described in local reports suggest that it creates a culture of self-censorship. The Election Law has been used to sue political candidates who provoke or incite someone to violence or who spread hatred. However, reports produced under the Social Media 4 Peace project also state that the law's provisions are too imprecise to create a strong basis for combating harmful ethnonational and political online content, as revealed in cases of sentence evasion by using 'private' social media accounts and profiles as opposed to official campaign ones. Additionally, provisions are limited to official election campaign periods (30 days). Since the campaigns are typically longer, spreading such content outside of the official election periods can also be a way of evading sentences, as shown in concrete cases outlined in local reports.

Lastly, as previously highlighted, there is no regulatory body with the power to oversee online content, since the regulations of the BiH Communications Regulatory Agency do not apply to internet content.

▶ ▶ ▶ 2.1.3 Civil Society and Companies' initiatives

In Bosnia and Herzegovina, social media companies seem not to act assertively to mitigate the problems generated by harmful content. As local reports show, most companies do not even have a local office, and community standards and terms of service are not entirely available in local languages or are poorly translated. Transparency reports also do not provide country-level figures for relevant indicators, in particular the amount of hate speech detected and removed in BiH or in any country. The fact that BiH languages (Bosnian, Serbian and Croatian) are used mostly only in the country – and are not widespread internationally, as are other languages – makes adapting to local contexts more difficult. Therefore, the implementation of company policies is limited and allows harmful content to remain accessible to social media users.



Despite the amendments to the Criminal Code in BiH prohibiting the public condoning, denial, gross diminution or justification of genocide, crimes against humanity and war crimes, there is a growing number of online incidents glorifying war crimes and convicted war criminals in early 2022.

The implementation of social media companies' policies is limited and allows harmful content to remain accessible to social media users.

Regarding disinformation, emphasis is placed on self-regulatory actors, specifically the Press Council of Bosnia and Herzegovina and the Raskrinkavanje fact-checking initiative. The Press Council self-regulates online and in print media and is restricted to mediation and non-binding decisions related to media content that violates the standards of the Press Code, regarding, for example, hate speech, disinformation and editorial responsibility. The Press Code was amended in 2021 to cover the overall content of online media, including user-generated comments, and to introduce provisions on the use of information technologies and disinformation, among other issues. The number of users' complaints to the council has been rising in recent years, and a significant number of them have been resolved by mediation. [24] Local reports point out that the council and its code have a positive impact on raising concerns among professional producers. But the council is limited in scope and in multisectoral composition, and it is unable to change the ethnonationally biased environment in traditional media or to have a positive impact on social media content moderation. Once it has expanded its scope, it remains restricted to professional journalistic content. In addition, reports point out that automated reporting mechanisms for flagging or asking for remedy are unresponsive, not user-friendly, and consequently misunderstood by users. Furthermore, users who are not able to reach platform staff through personal or professional connections. Finally, as in other countries, there is a lack of harmony between the standards and policies of different platforms, resulting in content moderation practices that vary across the most used social media platforms.

The main fact-checking organization in BiH is Raskrinkavanje, which is run by Zašto ne. Since its establishment in 2017, it has uncovered thousands of examples of problematic social media content, in particular disinformation, especially during the Covid-19 pandemic. In 2020, platform members of IFCN began working with Facebook so that after fact-checkers marked content as fake, Facebook also marked it to reduce its reach, similarly to what happens in other jurisdictions. This partnership between Facebook and Zašto ne is the only institutional relationship that Meta has in BiH, which reinforces the perception held by different stakeholders that content moderation is not conducted sufficiently in a broad and multisectoral manner. Additionally, local reports identify differences in perspective between Zašto ne and the Press Council, which, despite their having the same goals, reduce and weaken the impact of both initiatives.

There are other media watchdog organizations making various contributions to help curb harmful content:

- // **Mediacentar** - publishes thematic articles on issues related to the media and social networks, including disinformation and hate speech.
- // **Analiziraj.ba** - monitors the content of television broadcasters.
- // **Sarajevo Open Center** - monitors hate speech against the LGBTIQ+ community.
- // **Association for Democratic Initiatives** - provides an online form for reporting hate crimes and hate speech and gives free legal advice.



[24] In 2021, the Press Council received 1,073 complaints, of which 559 referred to texts published in print and online media; 505 complaints referred to user-generated comments, mostly hate speech, on online media. Sokol and Čalović, op. cit.

- // Association of BiH Journalists - receives complaints from journalists and gives legal advice on various issues including online harassment.*
- // Interreligious Council of BiH - is a non-governmental organization that connects different religious communities in BiH and has an online form for reporting attacks that include hate speech and hate crimes against religious objects.*
- // Organization for Security and Cooperation in Europe's (OSCE) mission to Bosnia and Herzegovina - has a mechanism for monitoring hate crimes based on police records and court proceedings. [25]*

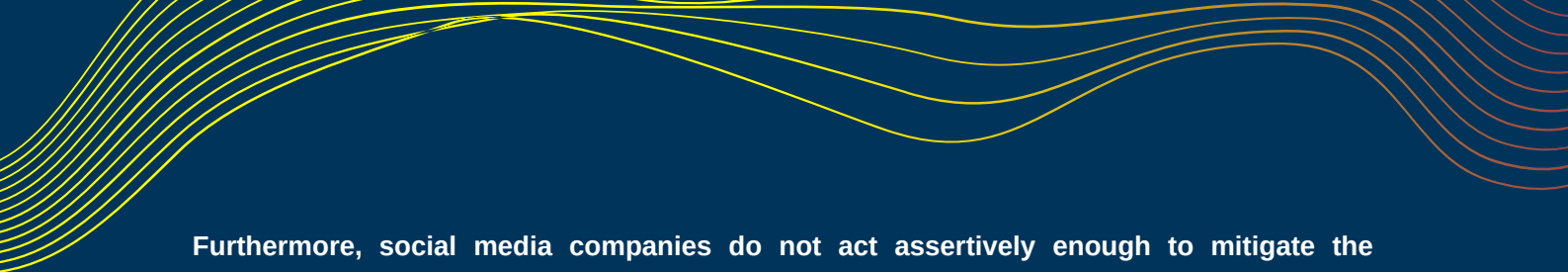
It is also important to mention the existence of coalitions specifically formed to address hate speech. Although informal and without specific mandates, such initiatives can be very relevant to the development of a more structured coalition as part of the project Social Media 4 Peace. The Coalition to Combat Hate Speech and Hate Crimes, founded in 2013, advocates for the improvement of the legislative framework on hate speech and public awareness campaigns, and it comprises, among others, Civil Rights Defenders, Media Centar Sarajevo, Journalists Association, and the Press Council. More recently, the SEE Digital Rights network was established, coordinated by Share Foundation and BIRN Hub, whose members include Zašto ne.

Although the number, strength and quality of initiatives such as these have grown in recent years, local reports produced by the Social Media 4 Peace project emphatically point out a lack of broader dialogue and cooperation among all stakeholders, and note that although a number of actors are working on monitoring harmful content, such initiatives are limited, notably in terms of impact, and fragmented.

▶ ▶ ▶ 2.1.4 Analytical Synthesis



Linked to the country's history of ethnic conflict, BiH communities are extremely vulnerable to harmful and hateful narratives, and this landscape has serious implications for social cohesion and peace-building processes. The legislative framework regarding hate speech in BiH is fragmented: FBiH and RS legal instruments often conflict. Criminal codes are limited and do not include categories of groups that are normally protected by international and regional legal instruments, such as skin colour, gender, sexual orientation and disability. There is an evident lack of gradation and levels of damage extent, tolerance and acceptability. Nonetheless there is no specific regulation or institutional body responsible for objectively protecting BiH's population from harmful content since the Communications Regulatory Agency does not focus on online content. Cases described in local reports also show that existing legislation is often used to limit freedom of expression and to prosecute journalists and independent content producers. In this context, developing a legal framework according to regional and international standards is key. In a scenario that gives social media a central role to play in limiting harmful online content, it seems relevant regulatory adjustments should be made.



Furthermore, social media companies do not act assertively enough to mitigate the problems generated by harmful content. As local reports show, community standards and terms of service are either not entirely available in local languages, or are poorly translated. As is true for most countries, transparency reports do not provide domestic figures for most of the relevant indicators, in particular the amount of hate speech removed and detected in the country, the types and targets.

In this scenario, the effects of company content moderation policies can be considered limited, allowing harmful content to be easily accessible to social network users. Practices orchestrated by specific groups seeking to stop or reduce the circulation of content produced by different political or ethnic groups – such as inauthentic coordinated behaviour – are not correctly identified by social media companies. Additionally, limitations of Artificial Intelligence (AI) are also reported, which result in both the blocking of legitimate content and non-detection of hate speech, especially due to the use of techniques to evade moderation. This tendency becomes even more accentuated due to the country's linguistic specificity.

In this context, civil society initiatives to curb harmful content and promote the factors that encourage pluralism and a culture of peace end up being scattered and fragmented, which reduces their effectiveness. Similarly, the role of companies in content moderation is limited, as well as their dialogue with civil society and public authorities to curb harmful content. These shortcomings highlight the need for these actors to engage more deeply in multisectoral dialogues, and for companies to invest in local content moderation policies.

2.2 Indonesia

▶ ▶ ▶ 2.2.1 Context

Indonesia is the fourth most populated country in the world (270 million inhabitants) and the internet penetration in the country is at 73.7% (202.6 million people), according to the Indonesian Internet Service Provider Association (APJII); 96.4% of all Internet users (195.3 million) use mobile internet. [25] The most used social media platforms are YouTube (93.8%), Instagram (86.6%), Facebook (85.5%), and Twitter (63.6%). Of the social messaging apps, WhatsApp is the leader by far, with 87.7% of users compared to 52.4% of Facebook Messenger users.

The country has the largest number of Muslims (231 million in 2021) in the world, but also has five other official religions, namely Protestantism, Catholicism, Hinduism, Buddhism and Confucianism; and it has over 300 ethnic groups living together. Indonesia is characterized by social and cultural diversity, but also deep historical roots of divisions, especially of ethnic and religious minorities.

In this context, online harmful content is characterized by the frequent occurrence of hate speech against specific ethnic minorities, notably the Ahmadiyah, Shi'a, and Chinese Indonesians, and against the LGBTIQ+ communities. Besides, 'the spread of disinformation is used to deepen the existing social, racial, and religious divisions in the country, and such efforts are more aggressive during election periods.' [26] A clear example is the circulation of hoaxes on social media claiming that protesters were shot by Chinese police during demonstrations that happened just after the 2019 presidential elections. Although most of these messages did not contain explicit incitement to violence, by promoting anti-Chinese perceptions they contributed to triggering racial hatred that resulted in chaos.

Although the State has adopted severe legislation that makes all harmful content illegal and subject to criminal prosecution, the circulation of such content notably targeting minorities remains an issue while the vague legal terminology makes room for discretionary applications against ordinary citizens. On the other hand, social media companies have the same 'text-based' policies for hate speech that they have for the rest of the world, but CSOs representatives allege that this is not enough. According to one interviewee, the diversity found in the country would justify a close dialogue with civil society on the enforcement processes of social media policies. [27]

[25] Kemp, S., 18 February 2020, Digital 2020: Indonesia report; and APJII internet survey report 2019 – 2020, November 2020.

[26] Haristya, S., 2022, [ARTICLE 19 report on Indonesia for the Social Media 4 Peace Project](#).

[27] As stated in Haristya, op. cit., p. 14: 'In response to a question on whether it is possible for community guidelines to specify how to handling 'grey area' speech, a platform representative in Indonesia underlined the limitations of text-based policies to reflect the complexity of how platforms design and operate their content moderation processes, all the more in consideration of the enormous challenges of tackling the prevalence of 'grey area' content in a large and diverse country as Indonesia. Accordingly, they concluded that text-based policies should be complemented by dialogue with local groups in the enforcement processes.'

Internet penetration in Indonesia

73.7%



93.8% of YouTube users



86.6% of Instagram users



85.5% of Facebook users



63.6% of Twitter users

▶ ▶ ▶ 2.2.2 Legislation addressing harmful content

The Indonesian Constitution is aligned with the Universal Declaration of Human Rights and the ICCPR, but legislation related to content production and enforcement mechanisms conflict with international standards for freedom of expression.

The main legal instruments that affect content moderation are the Criminal Code and the Electronic Information and Transactions Law (EIT Act) approved in 2008 and amended in 2016. The Amendment expressly authorized the Indonesian Government to terminate access or order a provider to terminate access to electronic information or documents containing content defined as illegal. This includes gambling; slander or defamation; extortion and/or threats; false and misleading news that cause consumer losses in electronic transactions; hatred or hostility based on ethnicity, religion, race and class; and threats of violence or intimidation directed at individuals.

Almost all following content is criminalised in Indonesia:
gambling; slander or defamation; extortion and/or threats; false and misleading news that cause consumer losses in electronic transactions; hatred or hostility based on ethnicity, religion, race and class; and threats of violence or intimidation directed at individuals.



Almost all content listed above is classified as criminal and is subject to criminal sanctions. In the EIT Act, all prohibited acts are subject to criminal sanctions. The same happens with the Indonesian Criminal Code, Pornography Act, and Elimination of Racial and Ethnic Discrimination Act.

The Criminal Code defines slander and/or defamation as a crime that: 'deliberately attacks someone's honour or reputation by accusing someone of something, with the obvious intent to give publicity'. The Supreme Court Decision No. 183 K/Pid/2010 allowed institutions to be considered 'victims' of slander. A recent Joint Decree by administrative, judicial and police authorities [28] has tried to eliminate this perspective, but it is not clear whether its application will affect the Supreme Court's understanding. Indonesia's Criminal Code also defines spreading 'false news or information which can cause trouble among the people' as a crime. The EIT Act narrows the criminalization of fake news to those cases 'resulting in consumer losses in electronic transactions'.

Broad interpretations of terminology and the prohibition of hate speech have negative effects on freedom of expression in the country, as seen in District Court decisions. In 2018, the District Court of Kebumen expanded defamation to include the honour of a legal entity or state institutions. In 2019, the District Court of South Jakarta categorized supporters of presidential and vice president candidates and the organizer of the general election as a 'group'. In 2020 the District Court of Jayapura broadly interpreted the nation of Indonesia as part of a 'group' according to Article 28 paragraph (2) of the EIT Act, and the District Court of Kendari held that harsh criticisms and negative comments against state institutions be considered hate speech under the same Article 28 paragraph (2) of the EIT Act. [29]

[28] Joint Decree of the Minister of Communications and Informatics, the Attorney General and the Chief of the Indonesian National Police No. 229 of 2021; No. 154 of 2021; No. KB/2/VI/2021.

[29] As stated in F. Rahman, S. H. Nasution, A. Firdharizki, N.O. Aletha and A. Putrawidjoyo, 2022, [Regulating Harmful Content in Indonesia, Report for the Social Media 4 Peace Project](#), Jakarta, CfDS.

Another fact that puts Indonesian legislation at odds with international standards is the government's responsibility to apply content moderation directly. The government is legally bound to prevent prohibited content as defined in laws and regulations and is authorized to terminate access and/or instruct the Electronic System Operator to terminate access to content that violates the laws and regulations. In other words, the service provider is obliged to remove illegal and harmful content when ordered by government institutions.

A problematic step in this direction was the Ministerial Regulation 5/2020 (MR5), which defines the obligation of social media companies to remove content in four hours in case of urgent takedown requests, without enabling any kind of due process. When responding to orders from the Minister of Communication and Informatics of the Republic of Indonesia (MOCI), social media companies have no other option than to remove the content in question. Alternative measures such as reducing dissemination or flagging content are not allowed. Noncompliance makes companies subject to fines as well as sanctions that can go as far as to terminating access to the platform.

After the MR5 enactment, Meta Transparency Reports show an increase in content restriction for local law violation. In the second half of 2021, there were restrictions for 3,380 pieces of content based on local law, a sharp increase considering the previous years. The same trend is shown in the Google Transparency Report, which found 253,633 requests coming from the Ministry in the period ending in June 2021, compared to 227 in the previous period ending in December 2020 and 26 in the period ending June 2020.

As stated by ARTICLE 19, 'the overly broad definition of prohibited online content in Indonesia's Internet-related regulations along with the compliance of platforms with the government's requests for securing their presence and expansion in the country may further undermine the protection of freedom of expression in the country.'

Reports from Southeast Asia Freedom of Expression Network (SAFENet), a civil society organization, show that the EIT Act is usually applied by government officials against ordinary citizens, journalists and social movements, [30] which indicates it can stifle freedom of expression and enable power abuse. In 2020, SAFENet's report noted 84 cases of criminalization, compared to 24 cases the previous year. The EIT Act was the primary restriction regulation used.

A coalition of 24 civil society organizations in Indonesia put together a report urging several reforms to Indonesia's content regulation and content moderation regime. The EIT Act is seen as neither enabling due process nor providing robust accountability. Data shows a 96.8% conviction rate and an 88% incarceration rate in applying the EIT Act, what is deemed quite high. The restrictive legal environment, however, does not prevent the dissemination of harmful content, according to the local reports produced for this project.

The restrictive legal environment in Indonesia does not prevent the dissemination of harmful content.



[30] As stated in Rahman et al., Regulating Harmful Content in Indonesia: '70% of reporting on the EIT Act to the police from 2017 to 2019 are conducted by people with power, including officials, businessmen, and the police themselves. Meanwhile, the other 29% are carried out by citizens. Moreover, according to SAFENet's report in 2021, out of the 84 subjects reported, 50 are civilians, 15 are activists, four laborers, three private employees, two students, and a journalist.'

▶ ▶ ▶ 2.2.3 Companies and Civil Society initiatives

The presence of social media companies in Indonesia increased after the 2017 Jakarta gubernatorial elections. Because of the difficulties involved in contacting the offices of US-based social media companies, the Ministry of Communication and Information Technology started putting pressure on the companies. In July 2017, a government request for content removal on Telegram that received no response led to the app's blocking. In August of that year, Facebook opened an Indonesian office. One year after, TikTok also faced blockages in the country for hosting 'pornography, immorality, religious harassment, and other negative content.' As in the case of the two other countries studied, Indonesia has no available information on how social media companies employ moderators or how many local moderators are employed. Moreover, community guidelines are not updated in local languages. [31]

As stated, companies apply global rules, but acknowledge that they should be complemented by dialogue with local groups in the enforcement processes. An example of the relatively little knowledge companies possess can be seen in the attacks against online social movements in a village called Wadas. In February 2022, there was a clash between the police and some residents of this village in Central Java who were protesting a planned mine, which led to the forceful arrest of 40 residents. Twitter accounts of Wadas residents and youth activists who sided with the residents were suspended due to mass flags. It took two days for the online situation to be clarified through the intervention of civil society organizations. Twitter eventually unsususpended them and ended up applying its blue sign of verification to the accounts of users who represented the movement.



Social media platforms acknowledge that their global rules should be complemented by dialogue with local groups in the enforcement processes.

Civil society in Indonesia is strong, and there are several NGOs focused on digital issues and on the defence of various groups affected by content moderation. Examples of such organizations are ICT Watch; Mafindo (the Indonesian Anti-Hoax Community); and SAFEnet (Southeast Asia Freedom of Expression Network), which has teams dealing with different kinds of harmful content.

Some human rights organizations also focus on digital issues, namely ELSAM, Human Rights Watch Indonesia and Tifa Foundation. Other organizations are more focused on defending the rights of individuals and groups affected by content moderation, such as LBH Apik (women), Arus Pelangi (LGBTIQ+) and ECPAT (children), in addition to election watchdogs, such as DEEP Indonesia and Perludem.

[31] As an example, in March 2022, the Indonesian page for Facebook's misinformation guidelines announced that 'Some of the content on this page is not yet available in Indonesian language.'

Some of these civil society organizations (for example ICT Watch, Mafindo, and SAFEnet) are recognized by the companies as trusted flaggers, acting upon Facebook, YouTube, Twitter, and TikTok. Meta claims to have twelve trusted flaggers in the country. Trusted flagger groups are considered to be closely engaged with social media companies, but an ongoing process of dialogue between the two that can produce meaningful results seems to be lacking.

While coalitions on freedom of expression and social media ethics already exist, there is space for more collaborative and coordinated efforts on content moderation. The establishment of a coalition on this issue could improve the dialogue between civil society and social media companies at the national level. The ARTICLE 19 report for Indonesia identifies the Tifa Foundation, SAFEnet and Perludem as having the potential to lead a coalition-building process.

▶ ▶ ▶ 2.2.4 Analytical Synthesis

Indonesia has reportedly seen several cases of harmful yet lawful ('grey area') speech acts that have resulted in real-world violence. The case study of the 2019 elections, for instance, shows that the election authority argued in favour of moderating anti-Chinese and anti-Communist content, but social media companies resisted, arguing that the content did not contain incitement to violence. The elections were marked by riots triggered by this type of content. Other riots incited by social media debates were reported in North Sumatra, Singkil, Aceh and Tolikara, Papua. Local interviewees report that social media companies were consistently reluctant to deal with harmful but lawful content. This is considered one of the triggers for more restrictive legislation from the government.



In 2016, massive protests from Indonesian Muslim groups happened after a video with Ahok (Basuki Tjahja Purnama), the Jakarta governor at the time, had been edited to seem as if he were insulting the Koran as a Christian and Chinese Indonesian, Ahok has a double minority background and was actually criticizing the use of Islam as a campaign tool. The protests led President Jokowi to allow Ahok to be charged and prosecuted. He ended up being jailed for blasphemy.

Reports from the Oxford Internet Institute provide evidence of coordinated disinformation actions in Indonesia conducted by extremist groups. Paid commenters and automated accounts are said to be used as 'buzzer strategies' to distort the public debate, by increasing positive or negative interactions over determined content.

Hate speech against LGBTIQ+ communities and against Ahmadiyah, Shi'a and Chinese Indonesians is frequent. The Indonesian Government has often ordered the removal of LGBTIQ+ content. Facebook has misapplied hate speech policy for the use of the Indonesian word 'queer' by a gay activist (Hartoyo case). YouTube removed a webinar of the Indonesian Journalists Union for Diversity following massive flagging of its account as 'sensitive content'. These cases are better described in the local report of ARTICLE 19 produced as part of the project, which affirms that the acceptance and tolerance of LGBTIQ+ in Indonesia cannot be separated from sociocultural attitudes and religious values. 'Generally speaking, all of the religions that are accepted in Indonesia are against the LGBTIQ+ practice,' stated in the 2022 report Regulating Harmful Content in Indonesia.

The act of exposing personal data as a harassment strategy (also known as doxing) is reported as being used against journalists, activists and human rights defenders. Online gender-based violence has increased in the last years, as shown in the ARTICLE 19 local report. According to the report: 'The National Commission on Violence Against Women (Komnas Perempuan) recorded 940 reported online gender-based violence cases in 2020, an increase of 241 cases from 2019. The Legal Aid Foundation of the Indonesian Women's Association for Justice (LBH APIK) dealt with 307 cases in 2020, while before the pandemic, it handled only 17 cases in 2019. Moreover, while in 2019 the Digital At-Risks (DARK) Subdivision of SAFENet assisted 45 victims of online gender-based violence, it received 169 filed cases from March to June 2020.

2.3 Kenya

▶ ▶ ▶ 2.3.1 Context

With an estimated population of approximately 50 million people, Kenya has seen an intense expansion of the internet in recent years, despite its still having significant inequalities in terms of access. Data from 2021 show approximately 64 million mobile subscriptions (density of 132%), 46.7 million internet subscriptions, and 27.5 million broadband subscriptions. [32] The number of active monthly social media users was calculated at 11 million, increasing the social media penetration rate to 20.2%. Among the open social media platforms, Facebook and YouTube command the highest overall usage with 9.5 million and 7.8 million active users respectively. This is followed by LinkedIn (2.5 million), Instagram (2.3 million), Twitter (1.1 million), and Snapchat (1.3 million).

In this vibrant digital environment, the spread of potentially harmful content on social media has increasingly become a key challenge for the country. An emblematic milestone was Kenya's 2007 elections when SMS messages were spread to undermine social cohesion and create a violent environment. Communication channels changed in the following years as social media emerged, although digital divides linked to gender, geography, and class are still persistent. As local reports point out abundantly, social media has been used, especially during recent election periods (2013 and 2017), to fuel political tensions, reaffirm existing prejudices and increase political divisions. [33]

The research commissioned under the Social Media 4 Peace project that was conducted in Kenya also shows that groups or individuals that are already targeted or marginalized by the society – women, LGBTQI+ people, minority ethnicities or nationalities, individuals with serious diseases – are more likely to be disproportionately affected by stereotypes, prejudice and discrimination, including on social media. [34]

Social media users in Kenya

11 million



9.5 M of Facebook users



7.8 M of YouTube users



2.5 M of LinkedIn users



2.3 M of Instagram users



1.3 M of Snapchat users



1.1 M of Twitter users



[32] Authority of Kenya, 2021, Fourth Quarter Sector Statistics Report for the Financial Year 2020/21. In general, UNESCO uses International Telecommunication Union (ITU) data as a reference for describing local scenarios of internet access and use. However, this report took as reference data produced by the national regulatory agencies of the researched countries, as they were, at the time the report was completed, more up-to-date than ITU data.

[33] An emblematic example took place in 2017, when Cambridge Analytica, working for the ruling party, mined Kenyan voter data from Facebook and used it to manipulate users with apocalyptic attack ads and smear campaigns against the incumbent's opponent, depicting him as violent, corrupt and dangerous (Build Up, 2022).

[34] V. Kapiyo, 2022, Content Moderation on Social Media and Local Stakeholders In Kenya; and Mapping of Legal Framework and Responses by Actors to Address Harmful Content Online In Kenya, Build Up 2022.

The effective curbing of online harmful content, notably hate speech and disinformation, through both legal and non-legal instruments, requires a good understanding of the history and political context of the country where political forces are linked to ethnic groups. It also demands comprehension of cultural and linguistic diversity, another central element of the national context with more than 70 distinct ethnic communities speaking close to 80 different dialects and practicing different cultural traditions. [35]

Despite the enactment of Kenya's 2010 Constitution and the new legislation on harmful content, as well as the many civil society and social media company initiatives launched in the past decade, the research for this project shows that these new tools have not solved the problem of online harmful content. There are fears about the impact of disinformation and hate speech on conflict dynamics, particularly at the time of elections. The problems identified include, on one hand, the lack of transparency of social media companies regarding the implementation of their content moderation policies to curb harmful content in Kenya, and the lack of coordination efforts from the civil society side to limit the spread of harmful content; and on the other hand, the misuse of the legal framework to curb harmful content as a means to curtail freedom of expression. All of the above is revealed in several examples of posts by politicians or groups of supporters making inflammatory remarks and encouraging ethnic violence that nonetheless remain online, while there are reports of numerous cases of legitimate speech that have been either punished or removed from social media.



2.3.2 Legislation and main State initiatives addressing harmful content



The central legal framework for freedom of expression and harmful content in Kenya is based on (1) the national Constitution, (2) the National Cohesion and Integration Act, and (3) the Computer Misuse and Cybercrime Act. Kenya's 2010 Constitution protects freedom of expression, freedom of media and freedom of information. It states that freedom of expression does not extend to hate speech, incitement to violence, war propaganda, and hatred advocacy that constitutes ethnic incitement, vilification of others or incitement to cause harm; or is based on any ground of discrimination specified or contemplated in the Constitution. These definitions go beyond the legitimate restrictions laid out under article 20 of the ICCPR.

Hate speech is formally addressed in the 2008 National Cohesion and Integration Act that defines its limits and accordingly sets fines and possible imprisonment of up to three years in case of breach. It also establishes the National Cohesion and Integration Commission (NCIC) to promote harmony and the peaceful coexistence of different communities in Kenya. While the National Cohesion and Integration Act addresses issues of hate speech by individuals, hate speech disseminated through media and journalism is covered by the Media Act. The NCIC also coordinates a multi-agency approach in partnership with other state bodies like the National Steering Committee on Peacebuilding and Conflict Management (NSC), to fast-track hate speech cases and support investigations. Prior to the 2017 elections, the NCIC issued guidelines to regulate political messaging on social media, including hate speech. At the institutional level, another important initiative was the Uwiano Platform for Peace created by the NSC and NCIC that was recently relaunched to help promote a peaceful election in 2022 by enhancing coordination between a wide range of partners at county and national levels.

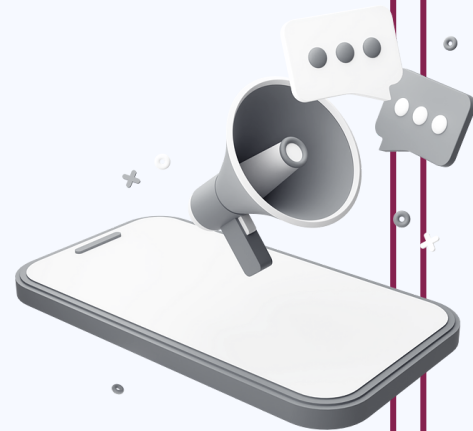
[35] English and Swahili are the country's official languages, and Swahili is spoken by the majority. The largest ethnic communities recorded are the Kikuyu, Luhya, Kalenjin, Kamba and Luo, with additional other minority ethnicities and indigenous communities.

According to experts and local reports, the limits on hate speech in Kenya's legislation are aligned with international standards, albeit vague in terms of application, since the legislation does not distinguish between levels of intentionality and impact, as specified in the Rabat Plan of Action. This lack of clarity about how to differentiate insults, abuses, threats, and other behaviours can put people at risk of unnecessary punitive responses and create a 'chilling effect'. [36] The NCIC's position in Kenya's institutional design is recognized as crucial to addressing harmful content. However, although the number of cases it has assessed is significant, the regulatory body has come under criticism for not sufficiently trying to limit the harmful speech of political leaders. [37]

Misinformation and disinformation were addressed in the 2018 Computer Misuse and Cybercrime Act, which came into full effect in 2020. This allowed the judiciary to impose prison sentences in cases of the intentional spread of misinformation. Local reports and stakeholders, however, indicate that neither the Computer Misuse and Cybercrime Act nor the National Cohesion and Integration Act are in line with international standards and in particular the International Covenant on Civil and Political Rights (ICCPR). These reports find that the laws' definitions are vague and can lead to violations of freedom of expression because of excessive punitive responses and the lack of gradation and specificity related to context, speaker, intent, content, form and extent of the online speech act, as specified in the six-point threshold test outlined in the Rabat Action Plan.

▶ ▶ ▶ 2.3.3 Initiatives by Companies

The role played by social media companies at the national level is key to overcoming challenges regarding content moderation in Kenya. Although Google, Meta and TikTok have offices in Kenya, which is their hub for East Africa, local reports show that these and other social media companies are inconsistent and opaque in enforcing their policies in the country. The uneven application of rules, particularly to disable or suspend users' accounts, was emphasized by local stakeholders, as was the limited independent oversight to assess how companies apply rules and content moderation decisions.



Another central issue is that company policies are not adapted or responsive to local complexities and linguistic diversity. A substantial part of the population, whose primary language is Swahili or a minority language, does not have full access to platform community standards because they are mostly available only in English. [38] This same linguistic issue has an impact on content moderation, performed by human moderators or automated systems, since these processes are applied primarily to English content: there is no data on or reliable information about the existence or intensity of content moderation in Swahili or other local languages. Fact-checking is done in only two languages. There is also a reported lack of data on the trusted flagger programmes, including content categories used, amount, and intensity of content removed, and actions taken by social media companies in the moderation process.

[36] Reported cases in which politicians have sought to remove online content, as well as episodes of electoral manipulation involving Cambridge Analytica, corroborate this perception.

[37] Currently, the NCIC has over 300 hate speech cases that are under investigation around the country. It also has ten cases that are pending in courts across the country where the offence is hate speech around ethnic content ([Build Up research](#)).

[38] Only the Facebook Community Guidelines are available in Kiswahili, while YouTube and Twitter rules are not.

In general, it is unclear how companies manage content moderation in Kenya, or even how many different language moderators there are. In addition, local reports point out that automated reporting mechanisms for flagging or requesting remedy are unresponsive, not user-friendly and consequently misunderstood by users. Furthermore, users who are not able to reach platform staff through personal or professional connections find it even more difficult to enforce their rights. Finally, as in other countries, there is a lack of harmony between the standards and policies of different platforms, resulting in content moderation practices that vary across the most used social media.

Local reports point out that automated reporting mechanisms for flagging or requesting remedy are unresponsive, not user-friendly and consequently misunderstood by users.

▶ ▶ ▶ 2.3.4 Initiatives by Civil Society

In the last fifteen years, Kenya's civil society has implemented a wide range of initiatives – capacity-building, raising awareness and monitoring – to restrain online harmful content while still protecting freedom of expression. For example, Amnesty International, ARTICLE 19 Eastern Africa and KICTANet are members of the Africa Internet Rights Alliance (AIRA) coalition, which promotes digital rights and seeks, among other goals, transparent moderation policies and consistent company performance throughout the region. FIDA Kenya, Act! and ARTICLE 19 Eastern Africa are members of the Civil Society Reference Group (CSRG). Other coalitions and networks work on different human rights issues such as peace (Peace Net), defending civic space (CSO Reference Group), digital identity (NIIMS Coalition), freedom of information (FOI Network), election observations, and monitoring (Election Observation Group).



In Kenya, there is a vibrant civil society environment that has made positive contributions to the curbing of harmful content.

In addition, there are a number of initiatives with goals more related to monitoring. Two of the most important are The Elephant, which publishes content on a diverse range of issues, including the impact of social media and harmful content on political and cultural context; and the Umati project, which monitors dangerous online speech. Other organizations like Watoto Watch Network, Bloggers Association of Kenya (BAKE), Pesa Check, Africa Check and AFP Fact Check are trusted flaggers and/or fact-checking institutions. Efforts like Maskani Commons, I Have No Tribe/ Mashada.com and the Sentinel Project – this last one also implemented in five other countries – seek to promote peace on social media.

These and other initiatives reveal a vibrant civil society environment that has made positive contributions to curbing the circulation of harmful content.

▶ ▶ ▶ 2.3.5 Analytical Synthesis

The present digital media landscape in Kenya makes addressing and countering potentially harmful content an extremely challenging task. Social media's exponential growth has allowed it to be used in different ways to disseminate content that contributes to maintaining an environment of tension between different political forces and ethnicities, despite important efforts to adapt legislation and create institutional instruments to strengthen a culture of peace.

Although the State has adopted specific legislation and created a regulatory authority to deal with harmful content, particularly hate speech, various reports highlight the need for making the rules more precise so that they specify the regulatory approach and sanctions for each type of content, considering the intentions involved and the impact of the messages according to the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) and the Rabat Plan of Action. This would help prevent potential censorship of content that is problematic yet legal under international standards. Extending legal protections in line with international standards to all segments potentially affected by harmful content, such as women, LGBTIQ+ people and individuals with serious diseases, is also a significant challenge.

Although the NCIC is recognized as a central institution attempting to ensure that different state actors follow a consistent approach, local reports produced by the Social Media 4 Peace project indicate that multiple institutions have overlapping mandates on digital content regulation without a coordinated approach. Strengthening this ecosystem and determining appropriate responsibilities for each state entity and for other stakeholders are actions recognized as essential to mitigate the effects of harmful content.

Civil society plays a central role in Kenya, especially during election periods when historically a large volume of hate speech and disinformation has been generated. However, civil society initiatives are limited in scope and coverage.

Various reports highlight the need for making the rules more precise so that they specify the regulatory approach and sanctions for each type of content, considering the intentions involved and the impact of the messages according to the international standards.



The performance and cooperation of social media companies is another crucial issue that is particularly important in multilingual, culturally diverse and polarized countries such as Kenya. While the companies' institutional presence is already more concentrated there than in other East African countries, reports recommend that companies exponentially increase their presence and their investments in content moderation, in order to offer greater transparency in their practices and create a transparent environment for dialogue and effective joint efforts with local stakeholders. A considerable number of problems must be addressed, including algorithms that amplify extreme and polarizing content or that remove legitimate content; low public awareness and limited access to content rules in local languages, along with lack of consideration for various language dimensions within local contexts; ineffective complaint mechanisms and remedies; and inconsistent and contradictory enforcement of content rules.

Finally, despite the efforts of state bodies and civil society, there is a discernible need to bring together different stakeholders, whose work, despite existing points of contact between them, remains fragmented. A sustainable and collaborative engagement between local stakeholders could maximize and harmonize existing efforts and help social media companies integrate a deeper understanding of the various dimensions of the local context into their content moderation systems.

This need for greater and sustainable engagement calls for a local multi-stakeholder group or coalition for freedom of expression and content moderation to promote international human rights standards and ensure local contexts are taken into account in content moderation.

Chapter 3: Comparison points

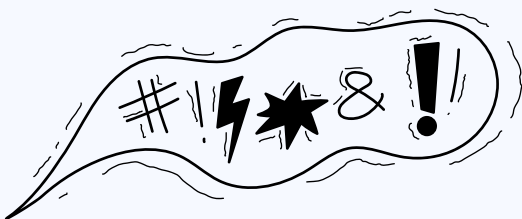
3.1 Evidence of the impacts of hate speech and disinformation on peace and human rights at the national level

Bosnia and Herzegovina, Indonesia, and Kenya provide evidence of online hate speech and disinformation affecting human rights, democracy, and peace and stability offline. Though the evidence is not comprehensive and based on a collection of individual cases, it is clear enough to raise some serious concerns.

In Kenya, social media has grown exponentially and has enabled the dissemination of content that reinforces tensions between different political forces and ethnicities. Already marginalized groups or individuals are disproportionately targeted by such content, in particular hate speech, and the reports refer to several examples of politicians or groups of political supporters making inflammatory speeches and encouraging ethnic conflicts. Election periods in Kenya are particularly problematic. In 2007, the spread of some SMS messages largely undermined social cohesion, and in subsequent elections (2013 and 2017), evidence showed that social media was widely used to fuel political tensions and reaffirm prejudices. While legislation has been adopted to curb online harmful content with the creation of specific agencies to enforce the law, such legislation does not yet resolve the issues, while at the same time it is having a chilling effect on freedom of expression.

In Bosnia and Herzegovina, the country's political structure is based on ethnicity, and its multipolarized social and political system is clearly reflected in the digital sphere. In this complex landscape, the impact of the intensive use of social media is largely visible in the dissemination of inter-ethnic hatred content, inflammatory narratives, disinformation campaigns and an environment that permits gender-based harassment and discrimination against under-represented groups like the LGBTIQ+ community and refugees. A characteristic that is specific to BiH is the high rate of online content that denies war crimes and glorifies war criminals. This clearly opens the door for online content to affect offline conflict dynamics. The legislation to address such content is scattered; while not effective in curbing such harmful content, it is often used by the powerful to limit freedom of expression.

Indonesia follows a similar pattern, with national elections exacerbating existing problems. In the 2019 elections, the country saw the growth of online hate speech that led to riots in the streets. As in Kenya, LGBTIQ+ groups and religious and ethnic minorities are especially affected by hate speech. Cases of harmful yet lawful ('grey area') speech have often led to real-world violence. Online gender-based violence has increased in recent years, and doxing cases against human rights defenders and journalists have been reported.



3.2 Compatibility between national legislation and international standards

In the three countries, national legislation shows some degree of inconsistency with international standards on freedom of expression, though the reasons for that vary from one country to another.

The main features of Kenya's legal framework on freedom of expression and harmful content are its excessively punitive approach and its vague application, given that legislation provides for prison sentences but does not distinguish between levels of intentionality, impact, or scope of narratives. These characteristics do not align with the International Covenant on Civil and Political Rights and the Rabat Plan of Action. As in the other assessed countries, Kenya's legal ecosystem has significant gaps in relation to other segments of the population potentially affected by harmful content, such as women, LGBTIQ+ people, and individuals with serious diseases. These omissions are in disagreement with the International Convention on the Elimination of All Forms of Racial Discrimination.

In Bosnia and Herzegovina, legislation to curb harmful content is different from one entity to another and is ineffective in curbing harmful content. The regulatory instruments do not incorporate different levels of gradation and specification that would align them with international standards, and not all the categories that need legal protection are sufficiently considered. It is specifically noted that, due to BiH's institutional design of two-state entities and a semi-autonomous district, there is contention in domestic legislation regarding the approach to war crimes, genocide, and the use of national symbols.

Indonesia has broad and detailed legislation that criminalizes several kinds of harmful online content that range from hate speech to disinformation, gambling and defamation. Its criminal code has a broad definition of disinformation, and the legislation applied specifically to the digital realm states that all harmful content can be criminally prosecuted. Indonesia's domestic legislation does not align with international standards because its definitions of harmful content are vague or broad, and its criminal approach is widely adopted, in addition to its enforcement mechanisms, as stated below.

3.3 Effective enforcement of legal frameworks

The effective enforcement of legal frameworks is uneven in all three countries. Since social and cultural inequalities are often reproduced in government or judicial decisions, socially vulnerable groups are more likely to receive sanctions than powerful ones. In addition, vagueness in the terms used in the legislation opens space for discretionary decisions.

In Bosnia and Herzegovina, judicial cases of hate speech are rare, and the number of court decisions are reported to be insufficient to prevent the dissemination of such content. Cases have been mostly initiated against ordinary citizens, and existing legislation is often used to limit freedom of expression. During election periods, which are very sensitive, gaps in legislation allow sentence evasion, which weakens their enforcement.

In Indonesia, the government directly enforces removal of harmful content. The main reasons for removal are pornography and gambling, but hate speech and disinformation are also listed as relevant issues. The Ministry of Communication and Informatics directly orders content moderation measures and is empowered to block access to content deemed illegal. Evidence shows that, as in BiH, legislation is being used against ordinary citizens and journalists, especially by public officials. In judicial cases, vague hate speech prohibitions negatively affect freedom of expression.

Kenya is the only country of the three studied in which the authorities have established dedicated institutions to deal with online harmful content; here the role of the National Cohesion and Integration Commission (NCIC) is crucial. Although the institution has assessed a significant number of cases, it has come under criticism for not seeking to limit the harmful speech of political leaders who are the main source of online hate speech. Additionally, the lack of a coordinated approach among institutions is seen by local researchers as a barrier to tackling hate speech. Strengthening this ecosystem and establishing appropriate responsibilities for different state entities and stakeholders is a key challenge.

3.4 Presence and local contextualization of social media companies

The main social media companies have offices in Indonesia and Kenya, but not in Bosnia and Herzegovina. Despite the historical ethnic conflicts in the country, the companies' operations for BiH are managed from other offices in Eastern European countries. However, the presence of social media companies in the other countries, namely Indonesia and Kenya, do not necessarily mean that companies are adapted to local contexts. There are some concerning conclusions of the reports produced by ARTICLE 19 and by the local researchers in the three researched countries:

- // There is a lack of transparency in how companies distribute the roles of moderation tasks, including the number of different language moderators and the amount of human resources and financial investments in each of the countries.
- // Companies do not process content moderation in some of the main local languages, such as Swahili in Kenya.
- // Community standards are not entirely or promptly available in local languages.

- // There is no possibility to contact representatives from the platform if content is removed, unless you know someone directly (issue of redress mechanism).
- // There is no data about how much harmful content is removed at national level and no granular data about the types of harmful content removed (including who are the targets).
- // There is no transparency about the trusted sources of information (number and names of local fact-checkers and in which languages they work).
- // The application of artificial intelligence in content moderation is often not balanced with a human-based approach and not underpinned by international standards.

Although companies have local offices in Kenya and often use them as a hub for Eastern Africa, social media companies enforce their national policies in inconsistent and opaque ways. A key local issue is that companies ignore local complexities and linguistic diversity, given that a substantial part of the population speaks Swahili but does not have full access to platform community standards, which are mostly in English.

Indonesia follows a similar pattern, with outsourced firms conducting content moderation in the country. Platform guidelines are not always translated into Indonesian, and there is no available data on how companies invest in content moderation.

In BiH, civil society organizations criticize social media platforms for their negligence of the issues of harmful content moderation in their country.

3.5 Multi-stakeholder environment

Civil society organizations are active in all three countries, but as harmful content is a wide area, those CSOs often act in silos. There are currently no specific coalitions to exchange knowledge and expertise on content moderation, although collective action and coalitions on other issues linked to digital rights exist, especially in Kenya and Indonesia, which have vibrant organizations and a fruitful collaborative environment. In each of the pilot countries, relations between CSOs and social media companies have yet to be strengthened.

In Kenya, public bodies and social media companies also make specific efforts to protect and promote freedom of expression and other human rights in the online environment. Although the relationship between different stakeholders is still fragmented, civil society organizations and some companies' representatives have manifested the will to build sustainable and collaborative engagement with local stakeholders. Such engagement can maximize and align existing efforts and help social media companies better understand the various dimensions of local contexts and apply this knowledge to their content moderation systems.

Indonesia has a reasonably strong civil society capable of monitoring social media and government actions in the country. Trusted flaggers maintain constant dialogue with social media companies, but CSOs that work on content moderation have not yet formed a coalition. Both social media companies and CSOs have shown a willingness to engage in a dialogue. A coalition in the country could conduct this dialogue between companies and various Indonesian stakeholder groups, 'pushing for accountability and transparency of social media platforms and the involvement of Indonesian stakeholder groups in the content moderation decision-making process'. [39]

In Bosnia and Herzegovina, the monitoring of online potentially harmful content is ensured by a variety of CSOs, in particular the media self-regulatory bodies, the fact-checking initiatives, and some groups jointly monitoring hate speech content. As in the other countries, however, the country lacks broader cooperation among all stakeholders dealing with online harmful content. Building an environment or open space for civil society, companies, academia, and public agents to conduct a dialogue and address potential violations of freedom of expression and the dissemination of harmful content must be a medium-term goal.

[39] Haristya, op.cit.

Chapter 4: Main findings



Finding 1



Online harmful content, in particular hate speech, disinformation and gender-based violence, affects the offline world and has a negative impact on peacebuilding. However, the lack of transparency regarding content moderation of such content by social media companies creates dependence on anecdotal evidence.



- // The three countries provide evidence that hate speech, disinformation and gender-based violence that happen in the online world affect the offline world. Vulnerable groups, such as ethnic and religious minorities, women and LGBTIQ+ groups are the most affected.
- // The absence of granular localized data on the scope, target and volume of disinformation and hate speech makes anecdotal evidence the only tool for analysing the impact of harmful online content on the offline world and creates barriers to accurate assessments.
- // Platform transparency and collaboration with researchers are fundamental requirements for better assessment of the impact of online content conflicts that reach the offline world.











Finding 2



The preconditions to ensure that social media companies undertake content moderation that considers local contexts are not yet in place.



-  // Social media companies do not yet make the necessary investments to ensure content moderation that is consistent with the volume or with the culture, language and local complexity of the content to be moderated.
-  // Guidelines are uniform and often not responsive to local context. There is no transparency regarding how many local language moderators and local fact-checkers are working in domestic contexts. Companies do not process content moderation in some of the main local languages, and community standards are not entirely or promptly available in local languages. This leads to inconsistency, opacity, inequality and other problems in applying the rules.
-  // Companies do not have offices in Bosnia and Herzegovina, where tensions related to ethnonational hate have been historically present. And in the countries where they do have offices, there is no ongoing organized and transparent dialogue with civil society organizations dedicated to peacebuilding or human rights.
-  // Data about moderation of harmful content is not offered at national level granularity.
-  // Part of the problem is related to global rules, such as the lack of transparency on the content moderation process and the lack of redress mechanisms and algorithmic accountability. Terms of use and community guidelines are not fully consistent with international standards, specifically the Rabat Plan of Action.
-  // The application of a human rights-balanced approach depends on the consistent understanding of local realities, which requires sufficient investments in local offices and moderation efforts and a multisectoral approach at the national level.



Finding 3



Existing legislation is often being used to restrict legitimate rights, notably freedom of expression, while at the same time it is not sufficiently protecting vulnerable groups.



-  // Parts of the legislation dealing with harmful online content in all three target countries are not in line with international standards and are often used by politically or economically powerful groups or public officials to restrict legitimate rights of ordinary citizens and journalists.
-  // The absence of thorough detail in the legal frameworks related to harmful content creates additional complexities for law enforcement. Vague legal frameworks allow for a distorted enforcement of the laws, in favour of those in power.
-  // Excessively punitive legal systems stifle freedom of expression while leaving vulnerable groups unprotected.
-  // The lack of incorporation of specific segments of the society to be protected by law often leaves women, LGBTIQ+ people, and religious and ethnic minorities legally unprotected.
-  // Restorative justice mechanisms and peacebuilding policies are lacking. Regulations and public policies are focused on punishing perpetrators without provisions for protecting and defending victims of hate speech.



Finding 4



Tensions arising from countries' historical and political contexts are often reinforced by social media dynamics.



-  // **Peacebuilding is best achieved through a balanced human rights approach, and this has been a challenge in countries prone to conflicts linked to cultural and historical factors.**
-  // **Instead of helping mitigate the tensions arising from historical conflicts, social media platforms create a fertile environment for reinforcing them. [40]**




Finding 5



Adherence to international standards to curb online harmful content on social media while protecting freedom of expression should be strengthened. At the same time, discussions are needed on the interpretation of these standards as they apply to the information ecosystem of social media, characterized by excessive speed and volume of circulation of harmful content.



-  // **We should adhere to the normative instruments of the international system for protecting human rights, in order to strike a balance between freedom of expression and other rights. However, some reinterpretation may be needed to incorporate the changes related to the speed and scale of the circulation of borderline harmful content.**

[40] One hypothesis is that polarization trends established by the economy of attention drivers (such as content traction defined by users' engagement) and the lack of mechanisms to protect human rights online imply a vicious cycle that amplifies historical oppression against minorities and vulnerable groups.

- // Coordinated and collective digital attacks targeting individuals who should not be under special scrutiny (ordinary citizens, journalists, human rights defenders, etc.) affect individual and collective rights. This creates 'grey area' cases that, when looked at separately, should not be considered infringements, but when looked at jointly, and when aimed at vulnerable persons or groups, may justify restrictions.
- // The lack of uniformity on the definition of hate speech and disinformation at the international level and in the policies of social media companies creates challenges for enforcing human rights provisions.

[40] One hypothesis is that polarization trends established by the economy of attention drivers (such as content traction defined by users' engagement) and the lack of mechanisms to protect human rights online imply a vicious cycle that amplifies historical oppression against minorities and vulnerable groups.

Chapter 5: Recommendations

Based on the main findings and on the analysis of the proposals made by stakeholders in the three countries [41] and on current developments of the related debates at international level, we present recommendations to be considered by international organizations, States, social media companies, civil society organizations and donors. In the last subsection, some recommendations listed as 'multi-stakeholder' represent those that would need to be discussed among different sectors in order to be implemented as a common agenda.



5.1 To international organisations

- // **Through UNESCO, use the project Social Media for Peace to convene a multi-stakeholder dialogue on the governance of harmful content, aimed at promoting a common understanding of hate speech and disinformation trends and occurrences and how to counter them. This would also ensure sharing of the lessons of the project at the national level, feeding into discussions at the global level. [42]**

- // **Organize discussion at the UN level within the UN Strategy and Plan of Action on Hate Speech to provide guidelines to the application of the Rabat Plan of Action to social media content moderation, considering also the recently launched report The practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies. The promotion of a unified definition of hate speech should be considered.**

[41] Found in the six reports on the three countries (three reports from ARTICLE 19 and three from local groups).

[42] At the moment of concluding this report, UNESCO announced a conference to be held in February 2023 on internet regulation, which can respond to the recommendation presented here.

- // **Discuss normative policy guidelines for content moderation on social media, considering the instruments of the international system for the protection of human rights, especially the ICCPR, the Rabat Plan of Action, the International Convention on the Elimination of All Forms of Racial Discrimination and the Convention on the Protection and Promotion of the Diversity of Cultural Expressions. These guidelines should interpret the instruments taking into account how recent changes in speed, scale and ‘swarming-alike’ dissemination of harmful content in social media affects individual and collective rights. [43]**

- // **Promote open debates and moot courts in different parts of the world on the challenges of applying international standards to platform content at local levels.**

- // **Through UNESCO, create a framework for regional offices to facilitate collaboration between social media companies and civil society groups focused on digital rights to ensure that content moderation and removal processes are aligned with community needs.**

- // **Through UNESCO, maintain existing programmes that provide training for public officials and judges on freedom of expression and harmful content in a digital environment.**

[43] Coordinated actions and swarming effects are triggered by the network effects. One piece of lawful but harmful content, even though made viral, does not have the same effect as thousands of different pieces reproducing the same discourse targeting specific groups. Harm is definitely influenced by a high volume of similar content spread over a short period of time. Besides, disinformation and hate speech do not only affect targeted groups, but also society as a whole. The effect of having a substantial part of available information be false or misleading affects directly and negatively the social dimension of freedom of expression.

- // **Develop media and information literacy programmes aimed at providing online users with the skills to critically examine online content and identify disturbing, hateful content and misinformation. Prioritize preventive educational approaches that alert to the harmful effects of online hate speech and foster media and information literacy alongside mitigation and counter efforts. [44]**



5.2 To States

- // **Reform legislation so that it is adapted to international standards, especially those laid out in the ICCPR, the Rabat Plan of Action, the ICERD, and the Convention on the Protection and Promotion of the Diversity of Cultural Expressions, as well as the official interpretation documents produced by the implementing bodies of these instruments. [45] Legislation to curb online harmful content should specifically protect the most vulnerable groups while safeguarding freedom of expression.**
- // **Ensure that liability regimes do not define a general objective responsibility [46] for social media companies, so as to avoid precautionary strict content moderation that has negative consequences on freedom of expression.**

[44] As stated in UNESCO/UN Office on Genocide Prevention and the Responsibility to Protect, 2021, Addressing Hate Speech on Social Media: Contemporary Challenges, Paris, UNESCO.

[45] National chapters offer a thorough analysis on the most problematic provisions and bring recommendations.

[46] When some part is considered liable if proved a misconduct, the harm and the causal nexus, regardless of being directly guilty.

- // Use legislation to promote transparency, due process, appeal and redress rights for users in the content moderation process.

- // Consider legally acknowledging multi-stakeholder initiatives, such as Social Media Councils, as voluntary-compliance bodies to be part of the regulatory system in dialogue with statutory regulators.

- // Ensure that companies are not prohibited from publishing information detailing requests or demands for content moderation, account removal or enforcement coming from state actors, except where such a prohibition has a clear legal basis and is a necessary and proportionate means of achieving a legitimate aim. [47]

- // Report their involvement in content moderation decisions and avoid ‘gag orders’ for company reporting, including data on demands or requests for content to be actioned or an account suspended, explained by the legal basis for the request. Reporting should account for all state actors and, where applicable, include subnational bodies, preferably in a consolidated report. [48]

- // Provide judicial assistance and restorative mechanisms for minorities and other vulnerable groups that constitute the majority of victims of incitement to hatred.

[47] As stated in Santa Clara Principles 2.0.

[48] Ibid.

- // Develop media and information literacy programmes aimed at providing online users with the skills to critically examine online content and identify disturbing, hateful content and misinformation. Prioritize preventive educational approaches that alert to the harmful effects of online hate speech and foster media and information literacy alongside mitigation and counter efforts. [49]



5.3 To Social Media Companies



- // Ensure transparency at the national level, offering granular data on:
 - Number of users active in the country or accessing content from that country.
 - Number of actions applied to accounts and content related to hate speech, disinformation, terrorism, violence, harassment and the types of moderating classifiers that affect peacemaking. This data should be clearly separated according to community standards, government requests, and judicial requests, as well as separated by country.
 - Appeals requested by users in those cases, appealing mechanisms used, and consequences.
 - Reach of infringing content.
 - Advertisements and boost investments in infringing content.
 - Aggregated data on targeted groups in the infringing content.
 - Aggregated data on the local law used as main reference for withdrawal requests by government or judicial orders.
 - General characteristics of teams involved in content moderation (number, language, qualification, nationality, and diversity aspects).



[49] As stated in UNESCO/UN Office on Genocide Prevention and the Responsibility to Protect, 2021, Addressing Hate Speech on Social Media: Contemporary Challenges.

- List of authoritative sources and trusted flaggers and fact-checkers, with access to their profiles, action mechanisms and criteria.
- Details of any rules or policies, whether applying globally or in certain jurisdictions, which seek to reflect requirements of local laws. [50]
- Details of any formal or informal working relationships and agreements the company has with state actors when it comes to flagging content or accounts, or any other action taken by the company. [51]
- Details of the process by which content or accounts flagged by state actors are assessed, whether on the basis of the company's rules or policies or local laws. [52]
- Details of state requests to action posts and accounts. [53]



Companies should comply with the [transparency recommendations defined by UNESCO](#), especially regarding the issues related to Social Media 4 Peace:

General

- || Principle 1.** Companies should explicitly recognize they have an obligation to protect human rights, and particularly freedom of expression and access to information, and the privacy of their users.
- || Principle 2.** Companies should recognize the need for the proactive disclosure of information as well as the need to respond to requests for information.

Content and process transparency

- || Principle 5.** Companies should be transparent about any terms and standards they enforce on their own platforms, setting out the limits of what they deem to be acceptable behaviour, and indicating how these parameters align to respect for international standards for freedom of expression.

[50] As stated in Santa Clara Principles 2.0.

[51] Ibid.

[52] Ibid.

[53] Ibid.

- || Principle 6. Companies should be transparent about any processes they have in place to identify, remove or reduce the impact of disinformation and hate speech, including pre- and post-publication measures; and about how such processes respect the free exchange of ideas and opinions.*
- || Principle 8. Companies should be transparent about any processes they have in place to identify and act against inauthentic behaviour and false identities when these are used to undermine human rights.*
- || Principle 9. Companies should disclose whether their processes for removing content and prohibiting behaviour are periodically subject to third party assessment as to human rights compliance, carried out by a respected external independent institution or oversight body; and consider whether such assessments should themselves be transparent as well as the company's own response to any recommendations made.*

Due diligence and redress

- || Principle 10. Companies should be transparent as to whether they have processes to enable people to raise concerns about content, including that which appears to violate human rights or advocates incitement to violence, hostility or discrimination, as well as inaccurate content; and they should be transparent about the implementation of such processes in terms of numbers and types of complaints and actions taken.*
- || Principle 11. Companies should be transparent about whether they conduct risk assessments for their operations, such as in contexts of upcoming elections or in countries in conflict, highlighting any serious potential threats to freedom of expression, privacy and other human rights, as well as their proposals for mitigating those threats.*
- || Principle 12. Companies should disclose whether they have risk assessments of any algorithms whose application can have the potential to discriminate against people unfairly, and whether any proposed mitigation measures exist.*
- || Principle 13. Companies should publish guidelines on how they will develop ethical AI processes to make consequential decisions that can impact on human rights.*

Data access

- || Principle 25. Companies should, in an analogous fashion to many public statistical bodies, have a process to allow researchers access to personal data they hold, where this will advance important public interest goals such as open access and open science, while guaranteeing users' privacy through the range of necessary measures.*

- // Ensure that reports, notices and appeals processes are available in the language in which the user interacts with the service, and that users are not disadvantaged during content moderation processes on the basis of language, country or region. [54]

- // Prioritize countries prone to conflict when defining financial and human resources investments in content moderation.

- // Establish local focal points to be contacted by vulnerable groups or individuals when affected by infringing content.

- // Hire appropriately sized teams to moderate content in the local language, with workers who know the local, social and political context and who are adequately trained to handle content moderation processes that address local conflicts.

- // Offer updated versions of community guidelines in all relevant languages in the domestic contexts.

- // Engage a list of locally informed signals of authority to identify potential harmful content.

- // Tackle coordinated actions aimed at attacking individuals or vulnerable groups based on hate speech, disinformation or gender-based violence.

- // Tailor algorithms so as to consider peace, human dignity and the rule of law as the key driving values. [55]

[54] Santa Clara Principles 2.0.

[55] Based on UN Under-Secretary-General for Global Communications recent statements. See, for instance: <https://melissa-fleming.medium.com/ukraine-is-a-moment-of-reckoning-for-social-media-5ac7c4f13d8c>

- // **Tailor algorithms to undermine the economy of attention drivers that do not consider human rights goals.**
- // **Adapt policies to international standards that strike a balance between freedom of expression and other human rights, while tailoring policies to national contexts so as to consider cultural, social, and political specificities. [56]**
- // **Offer open workshops and training on the tools offered by the guidelines and terms of services to counter hate speech and disinformation.**



5.4 To Civil Society

- // **Facilitate the creation of civil society coalitions on freedom of expression and content moderation, gathering groups with different expertise in relation to various types of harmful content and approaches. [57] Coalitions can play an effective role in bridging the gap between local civil society organizations and companies that operate on a global scale.**
- // **Gather qualitative data on individuals targeted by hate speech to better understand the scope and nature of harms, while respecting personal data protection [58], so as to foster evidence-based policies.**

[56] Caveat: tailoring policies should not imply in weakening principles against hate speech or complying with norms not aligned to the international human rights standards.

[57] ARTICLE 19 reports for Indonesia, BiH and Kenya provide thorough recommendations on the steps to create and implement such coalitions.

[58] As stated in UNESCO/UN Office on Genocide Prevention and the Responsibility to Protect, 2021, Addressing Hate Speech on Social Media: Contemporary Challenges.

- // Promote training to enhance the use of social media for spreading peace narratives and for promoting media and information literacy knowledge.



5.5 To Donors

- // Ensure that adequate resources are provided for specialized organizations dedicated to monitoring and countering hate speech, disinformation and gender-based violence, particularly those best equipped to take local contexts into account and provide them with support. [59]
- // Support the development of affordable, accessible and user-friendly tools and methodologies that can be used to monitor and detect hate speech across multilingual, multicultural contexts within a timeframe that allows for counteraction. [60]
- // Provide funding and resources for the development of educational programmes that foster resilience to hate speech, informed by current hate speech trends and responding to related challenges. This requires close collaboration between social media companies, research institutes and education stakeholders. [61]
- // Support the training of new fact-checkers in local languages.

[59] Ibid.

[60] Ibid.

[61] Ibid.



5.6 For Multi-stakeholder initiatives

- // Companies and CSOs should engage in long-term and frequent dialogues seeking to align their visions and achieve joint understanding of how to deal with hate speech, disinformation, and other kinds of harmful content. This dialogue can include:
 - Creating a Code of Conduct and other self- or co-regulation tools based on the instruments of the international system for the protection of human rights.
 - Defining classifiers for hate speech [62] based on local contexts.
 - Promoting training for fact-checkers and capacity building for civil society monitoring and research.
 - Building transdisciplinary capacity to respond to hate speech in societies, including actions aimed at protecting victims.
 - Engaging in legal reform debates related to content moderation.

- // In the future, establish Pro-Social Media Council Working Groups, with the following goals [63]:
 - Draft and adopt a constitution for local Social Media Councils as bodies to ensure oversight over content moderation and to act as individual complaint system mechanisms.
 - Define the essential principles for the operation of the complaint mechanism and the establishment of general guidelines (statutes, bylaws, etc.)

[59] Ibid.

[60] Ibid.

[61] Ibid.

[62] A type of machine learning algorithm used to classify automatically a data input.

[63] Based on the report brought by ARTICLE 19 about the Irish implementation of the pilot experience of Social Media Councils.

- **Ensure funding for the first year(s) of operating the Social Media Council.**
- **Organize a steering committee, once the constitution is adopted, in the spirit of a start-up that can lead the Social Media Council to its full operational capacity.**
- **Further explore the application of international human rights law to content moderation by preparing the adoption of a Code of Human Rights Principles for Content Moderation.**



References

ARTICLE 19, "Content Moderation and Local Stakeholders in Bosnia and Herzegovina," June 2022

ARTICLE 19, "Content Moderation and Local Stakeholders in Indonesia," June 2022

ARTICLE 19, "Content Moderation and Local Stakeholders in Kenya," June 2022

ARTICLE 19, "Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society," June 2022

Amnesty International Indonesia et al., 2021

Authority of Kenya, Fourth Quarter Sector Statistics Report for the Financial Year 2020/21, 2021

Build Up, "Mapping of Legal Framework and Responses by Actors to Address Harmful Content Online in Kenya," August 2022

Center for Digital Society, Universitas Gadjah Mada, "Regulating Harmful Content in Indonesia: Legal Frameworks, Trends, and Concerns," 2022

Council of Europe, Media Habits of Adults in BiH, 2021

Council of the European Union, Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law

CSIS Indonesia, Fire in the husk: The phenomenon of hate speech in Indonesia

Digital Report Kenya: 2021

Escola de Cultura de Pau, Alert21!, "Report on conflicts, human rights and peacebuilding," Barcelona, ECP, 2021

European Commission, A multi-dimensional approach to disinformation, Report of the independent High level Group on fake news and online disinformation, Luxembourg, Publications Office of the EU, 2018

Google Transparency Report, Government requests to remove content

H. Bailey and P. N. Howard, Country Case Studies Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation, 2020

I. Gagliardone, D. Gal, T. Alves and G. Martinez, Countering Online Hate Speech, Paris, UNESCO, 2015

International Convention on the Elimination of All Forms of Racial Discrimination, United Nations, Treaty Series, vol. 660, p. 195.

Indonesian Supreme Court Decision No. 183 K/Pid/2010

Joint Decree of the Minister of Communications and Informatics, the Attorney General and the Chief of the Indonesian National Police No. 229 of 2021; No. 154 of 2021; No. KB/2/VI/2021

Kemp, S., 18 February 2020, Digital 2020: Indonesia report; and APJII internet survey report 2019 – 2020, November 2020

Komnas Perempuan, Records of Violence against Women in 2020, 5 March 2021

Media Center Sarajevo, "Regulation of Harmful Content Online in Bosnia and Herzegovina: Between Freedom of Expression and Harms to Democracy," 2022

Melissa Fleming, Medium, "Ukraine is a moment of reckoning for social media," 2022

META Transparency Center, Indonesia, Country-specific information on content we restricted based on local law, Jan-June 2022

National Cohesion and Integration Commission Kenya, "Guidelines on Prevention of Dissemination of Undesirable Bulk and Premium Rate Political Messages and Political Social Media Content Via Electronic Networks," 2017

Note to Correspondents: Statement by Alice Wairimu Nderitu, Special Adviser on the Prevention of Genocide, on the introduction of amendments to the Criminal Code of Bosnia and Herzegovina, 23 July 2021

Office of the UN High Commissioner for Human Rights, OHCHR, "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework," New York and Geneva, 2011

Pérez Ana Laura, "The Hate Speech Policy of Major Platforms during the Covid-19 Pandemic," Paris/Montevideo, UNESCO, 2022

SAFEnet, "Research on increasing doxing attacks and their protection challenges in Indonesia," (In Indonesian "Riset peningkatan serangan doxing dan tantangan perlindungannya di Indonesia"), 2 December 2020

SAFEnet, "The Rampant Case of Online Gender-based Violence, The role of law enforcement officers needs to be improved," joint press release, 10 March 2021

S. Bradshaw and P. N. Howard, The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation, Oxford Internet Institute/University of Oxford, Computational Propaganda Research Project, 26 September 2019

Soroush Vosoughi et al, Science359,1146-1151, "The spread of true and false news online," 2018

Srebrenica Genocide Denial Report, 2021

T. Diela and F. Potkin, "We're not Chinese officers": Indonesia fights anti-China disinformation, Reuters, 25 May 2019

The Santa Clara Principles On Transparency and Accountability in Content Moderation, 2.0

UNDP in Kenya, Uwiano Peace Platform project

UN Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/74/486, 2019

United Nations, Strategy and Plan of Action on Hate Speech, May 2019

UN Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/67/357, 2012

UN Committee on the Elimination of Racial Discrimination, General recommendation No. 35, CERD/C/GC/35, 2013

UNESCO World Trends in Freedom of Expression and Media Development, Global Report 2021/2022, 2022

UNESCO, "Journalism, 'fake news' and disinformation: Handbook for Journalism Education and Training," Paris 2017

UNESCO/UN Office on Genocide Prevention and the Responsibility to Protect, "Addressing Hate Speech on Social Media: Contemporary Challenges," UNESCO, CI/FEJ/2021/DP/01, Paris, 2021

UNESCO, Video: The Rabat Plan of Action on the Prohibition of Incitement to Hatred, 2021

UNESCO, "Letting the Sun Shine In: Transparency and Accountability in the Digital Age," Paris, 2021

UN Human Rights Council, Annual Report of the United Nations High Commissioner for Human Rights, Rabat Plan of Action, A/HRC/22/17/Add.4, 2013

UN Human Rights Council, Annual report of the United Nations High Commissioner for Human Rights and reports of the Office of the High Commissioner and the Secretary-General, A/HRC/50/56, 2022

UN Human Rights Committee, International Covenant on Civil and Political Rights, General comment No. 34, CCPR/C/GC/34, 2011

UN, Human Rights Council, Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan, Disinformation and freedom of opinion and expression, A/HRC/47/25, 2021

UNHCR, Global Trends Report, Forced Displacement in 2022, June 2023

Social Media 4 Peace

Local lessons for global practices

This publication, developed under the UNESCO project “Social Media 4 Peace” funded by the European Union, overviews research conducted under the project focusing on Bosnia and Herzegovina, Kenya, and Indonesia. These include analyses of the regulatory frameworks governing harmful content online in these target countries, assessments of self-regulatory tools and content moderation policies of platforms, and the mapping of the local efforts by civil society.

The publication aims to inform global discussions on countering harmful content, especially in conflict-prone environments, by delving into the complexities of these countries' political, cultural, linguistic, and societal contexts. Its insights aim to serve as guideposts for stakeholders seeking to promote freedom of expression and a safer online environment.



unesco

United Nations
Educational, Scientific
and Cultural Organization

