ARTICLE 19

unesco
In partnership with UNESCO

Funded by the European Union

Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society

June 2022

ARTICLE 19

T: +44 20 7324 2500

F: +44 20 7490 0566

E: info@article19.org

W: www.article19.org

Tw: @article19org

Fb: facebook.com/article19org

This publication was produced with the financial support of the **European Union** and **UNESCO**. 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 or the European Union concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries.

The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions expressed therein, which are not necessarily those of UNESCO or the European Union and do not commit the Organizations.

Contents

Executive summary

This report presents a summary analysis of research on current practices of content moderation in Bosnia and Herzegovina, Indonesia, and Kenya, with a specific focus on 'harmful content' such as 'hate speech' and disinformation. The methodology combined desk research with qualitative interviews with key informants (representatives from local stakeholders). Findings from the country reports indicate that social media platforms, rather than serving as spaces for democratic debate and participatory citizenship, have increased ethnic-driven disinformation and politically motivated hatred, and reinforced the exclusion of marginalised groups. Given the importance of social media platforms, in countries where such tensions have in the past caused real-life violence, addressing the weaknesses of content moderation practices is of the utmost importance to ensure sustainable peace and enduring democracies.

The country reports identified a disconnect between tech giants' practices of content moderation and the local communities in which content was produced and distributed. Key findings included the following:

- Social media platforms lack understanding of specific cultural and societal settings and local languages. This disconnect is often linked to a perceived unwillingness or lack of resources allocated by social media platforms to specifically address certain country or regional contexts that they do not see as of strategic global importance.
- Current mechanisms set up by social media platforms to allow users to appeal content moderation decisions are insufficiently effective.
- There is insufficient transparency in content moderation practices, especially in relation to the lack of disaggregation of data on a per-country basis.
- Even when they are engaged in collaboration with social media companies, local civil society actors often feel powerless and unable to make their voices heard in relation to content moderation issues.

Under the UN Guiding Principles on business and human rights, companies have obligations to respect human rights and to offer remedy. Social media companies should therefore ensure that decisions on content moderation are made with sufficient awareness and understanding of the linguistic, cultural, social, economic, and political dimensions of the relevant local or regional context.

Most respondents from local stakeholders in the three countries welcomed the idea of a local coalition on content moderation and freedom of expression. Such a coalition could serve as an effective means to engage with social media companies in order to contribute to the development of content moderation practices that meet the requirements of international standards on human rights and are informed by a detailed understanding of all dimensions of the local context.

Introduction

This publication has been produced as part of the United Nations Educational, Scientific and Cultural Organization's (UNESCO) project **Social Media 4 Peace** funded by the European Union (EU).

The report forms part of ARTICLE 19's contribution to **Social Media 4 Peace**, the instrument contributing to Peace and Stability. The overall ambition of this three-year endeavour is to support the resilience of conflict-prone and polarised societies to the impact of the massive online circulation of content that spreads disinformation and incites violence and hatred, while protecting freedom of expression and enhancing the promotion of peace by maximising the potential of digital technologies, notably social media, to promote initiatives and narratives that create incentives for peace rather than violence. The project's geographical focus is on Bosnia and Herzegovina, Indonesia, and Kenya.[1] ARTICLE 19's contribution focuses on concerns raised with regard to current practices of content moderation on dominant social media platforms in the three target countries.[2]

The project is concerned with the circulation of categories of content such as incitement to hatred and disinformation. These are often grouped under the umbrella of 'harmful content', a term that is frequently used in today's policy conversations about the regulation of online content. As it focuses on the potential harms that can derive from its online dissemination, 'harmful content' is a very broad concept that ends up encompassing content that may or may not be protected by freedom of expression. By its nature, the notion of 'harmful content' is a very subjective and slippery concept that does not lend itself to legislative use because legitimate speech would inevitably be caught up in this net. Nonetheless, the massive online dissemination of items of content that may individually 'fly under the radar' – in other words, social media posts that, considered individually, may remain below the threshold that would trigger legal sanctions or the application of social media platforms' content rules – can have a dramatic impact on societies. The case of Myanmar provides a very clear illustration of the actual harms that can result from 'hate speech': a UN investigation found that the spread of 'hate speech' on

Facebook played a determining role in the genocide against the Rohingya population. More generally, content such as 'hate speech' and disinformation raises serious concerns because its massive circulation on social media platforms, especially in the specific contexts of conflict-prone, complex, and diverse societies, is considered to be one cause of severe societal harms (such as real-world violence, increased polarisation between social groups, or the undermining of trust in democratic institutions and electoral processes) or an aggravating risk factor of these societal harms.

Through a range of measures that include taking down a problematic post, reducing its visibility and distribution, labelling it as unverified information, and demonetising or suspending the account of its author, social media platforms have taken some actions to mitigate the risks linked to problematic content, thus contributing to supporting peace rather than increasing division and violence in society.[3] However, in order to do this successfully, companies need to base their content moderation decisions on a robust understanding of the local context in which content is circulated. Under the pressure of civil society organisations and the public, social media companies have in recent years increased their efforts and initiatives to deal with the problems of content moderation. In view of the information brought to light by whistleblower Frances Haugen, whether or not they are willing and able to invest the necessary resources to effectively achieve meaningful results remains to be seen.[4] One recent investigation, for instance, questions the attitude of Facebook towards 'hate speech' and disinformation in the context of the ongoing violent conflict in Ethiopia.

In the initial stage of the project, ARTICLE 19, with support from three research consultants, conducted research into how local stakeholders in Bosnia and Herzegovina, Indonesia, and Kenya perceive content moderation on social media.[5] For each country, through a combination of desk research and qualitative interviews with key informants, we sought to understand what the main categories of problematic content were, and how local actors analysed and reacted to the dissemination of such content. We were particularly interested in understanding if and how civil society was able to engage with

social media companies in order to share its knowledge of the national context to influence content moderation processes.

The project works from the premise that while social media companies rely on global content rules to govern content on their platforms, any particular case of content moderation can only be resolved on the basis of a detailed analysis of the linguistic, political, social, cultural, and historical circumstances in which it arises. Individuals and organisations that are part of the society targeted by certain pieces of online content are also particularly well positioned to reflect on the meaning and potential impact of such content. But are they able to take part in conversations about the regulation of content by giant entities that operate on the global level and are based in the Global North?

As the country reports show, even though individuals and associations intensively use social media on a daily basis, they often feel entirely powerless in relation to tech giants. In the interviews, some reported that they had never realised that engaging with social media platforms in order to seek resolution of a content moderation problem was even an option available to them. Stakeholders that are involved in various forms of engagement with social media companies, including as Trusted Flaggers, also report their frustration at not being listened to by platforms when they try to highlight problems and risks linked to the circulation of online content. The country reports show a dramatic imbalance of power and a stark disconnect between global companies and local civil society actors.

Working with local actors to support their efforts to have an effective voice is a key concern of ARTICLE 19's work to defend and promote the right to freedom of expression. The observation of a gap between social media companies and national civil society actors was the reason why we suggested that Social Media Councils (SMCs) should be created at the national level (provided such initiatives would not put individuals at risk and would not be subject to the risk of capture by government or other power-holders). The Social Media Council is a model for a participatory, transparent, and voluntary multi-stakeholder mechanism that would ensure oversight of content moderation on social media on the basis of international standards on freedom of expression and other fundamental rights. At the national level, SMCs would enable local civil society actors and

other stakeholders to take part, alongside social media companies, in designing content moderation practices and decisions that are based on international human rights and properly informed by the local context. The concept of SMCs has been endorsed by the former and current UN Special Rapporteurs on freedom of expression and broadly discussed in policy and academic circles.[6]

Along with monitoring legal developments in the regulation of social media platforms, ARTICLE 19 has been developing a model for a multi-stakeholder voluntary-compliance mechanism, known as a Social Media Council (SMC). The SMC provides a transparent and independent forum to address content moderation issues on social media platforms on the basis of international human rights standards.

The key objectives of the SMC are to:

- Review individual content moderation decisions made by social media platforms on the basis of international standards on freedom of expression and other fundamental rights. The right of appeal gives the SMC more credibility in the eyes of the public and gives individual users an opportunity to be heard on matters that directly impact on them;
- Provide general guidance on content moderation that is informed by international standards on freedom of expression and other fundamental rights. While there is a growing consensus on the relevance of international human rights law to content moderation, this is still an emerging field with many open questions;
- Act as a forum where all stakeholders can discuss and adopt recommendations (or the interpretation thereof). This participatory methodology promotes collective adoption and interpretation of guidelines and can help embed international standards in content moderation practices; and
- Use a voluntary-compliance approach to the oversight of content moderation where social media platforms and all other stakeholders sign up to a model that does not create legal obligations and where they voluntarily implement the SMC's decisions

and recommendations. The SMC will be a self-regulatory mechanism where representatives of the various stakeholders come together to regulate practices in the sector.

The current situations in the countries covered by the project may not directly lend themselves to the creation of SMCs which, even in the ideal theoretical environment of a stable democracy, would inevitably involve a lengthy and complex process. At the same time, tensions in these countries call for a quicker response – all the more so when elections are on the horizon. This is why ARTICLE 19 suggested that, possibly as a preliminary step towards the development of an SMC, a local coalition on freedom of expression and content moderation could play an effective role in bridging the gap between the realities for local actors and companies that operate on a global scale. Establishing a local coalition would involve a leaner process than an SMC, and could be facilitated and supported within a shorter time frame. Basing its work on international standards on freedom of expression and other fundamental rights, a multi-stakeholder coalition could provide valuable input to inform content moderation practices, notably through its knowledge and understanding of local languages and circumstances. As a one-stop shop representing a critical mass of local stakeholders, it could engage in a sustainable dialogue with social media platforms, and contribute to addressing flaws in content moderation and improving the protection of fundamental rights online. Finally, the coalition could provide training and support on freedom of expression and content moderation to local civil society actors that are impacted by content moderation. In short, the coalition would contribute to the development of content moderation practices that uphold international standards on freedom of expression while giving all due consideration to the local context.

In order to assess the feasibility of a local multi-stakeholder coalition on freedom of expression and content moderation, we conducted analyses of the respective stakeholder landscapes in Bosnia and Herzegovina, Indonesia, and Kenya. These mapping exercises looked at local stakeholders' strengths and needs in order to make recommendations on

the most effective approach to the development of a coalition. Through this research, at the initial stage of the **Social Media 4 Peace** project, we also consulted a broad range of stakeholders in each of the three countries. Using the views of local stakeholders, we wanted to test the idea that a coalition on freedom of expression and content moderation could play a role in bridging the gap between the realities for local actors and companies that operate on a global scale. Most of the interviewees responded positively to the idea, and their contributions have enabled the formulation of recommendations on how to facilitate a coalition in the specific contexts of the **Social Media 4 Peace** countries. In order to guarantee the effective ownership of the coalition by its members, the development process will necessarily include a validation exercise that ensures potential members have the opportunity to discuss the findings of the research. This will contribute to building consensus among participants on definitions of shared values, a common vision, and a clear goal for the local coalition.

The research conducted in the initial year of the **Social Media 4 Peace** project combined a policy and literature review conducted through desk research with qualitative interviews with key informants from civil society, the private sector, public actors, and social media companies.[7] The desk research identified issues linked to the circulation of problematic content on social media. The content moderation issues identified were then discussed during the interviews, which aimed to obtain an understanding of local groups' experiences and challenges when dealing with platforms to address such issues. The idea of a local coalition on content moderation and freedom of expression was also discussed with the interviewees, who provided their views on the overall idea of a coalition, as well as its potential structures, members, roles, and dynamics. In each country report, the introduction highlights the diversity and complexity of the country. The first chapter describes the landscape of social media platforms, and explores the dynamics of social media use and content moderation practices in the country. The second chapter provides an analysis of the different stakeholder groups that deal with or are impacted by content moderation practices. The third chapter puts forward recommendations, based on interviews, for facilitating the formation and operation of a civil society coalition on

content moderation and freedom of expression to establish an effective dialogue between social media platforms and local civil society actors.

In this summary report, the first chapter looks at the current state of content moderation practices and presents recommendations to address the flaws identified on the basis of international standards on human rights. The second chapter then delves deeper into the idea of a local coalition on freedom of expression and content moderation. On the basis of the country reports' findings, it presents recommendations on the possible role and development of such a coalition. This report focuses on the relationships between civil society stakeholders and social media platforms.

Content moderation and its discontent

Content moderation on social media has become the subject of heated debate in a wide range of contexts. On one hand, in response to public and government pressure, platforms have invested in the development of their systems, launched collaborations with Trusted Flaggers or fact-checkers, and continuously reviewed their content policies. Companies have also developed new features that give users more control over what they see, and some have started to undertake human rights impact assessments. On the other hand, civil society organisations, academics, and whistleblowers have denounced the ineffectiveness of platforms' measures to deal with 'harmful content', the lack of transparency, the inconsistent application of content rules, and the absence of available and effective remedies for users to appeal content moderation decisions. In ARTICLE 19's view, while social media companies have indeed implemented a range of increasingly sophisticated measures to address 'problematic' content, they still have a long way to go. This report looks specifically at the need for social media companies to inform their content moderation processes with an understanding of the local context and the role that local civil society actors could play in that respect.

To be sure, legislative and regulatory frameworks also play a very significant role in the governance of content.[8] In that respect, it is important to keep in mind that many areas of 'problematic content' investigated in this report are hard to define. Disinformation, for instance, is a complex issue, first of all because it is often difficult to draw a line between facts and opinions. Reporting accurately on current events is a similarly complex endeavour. This is why any legal prohibition of 'fake news' is likely to lend itself to abuse: it could easily serve as an instrument to silence critical media voices, and it opens the door to devastating consequences for the public sphere. In addition, some academic studies have pointed out that the extent to which disinformation is indeed 'harmful' to democracy is highly speculative,[9] which reinforces the conclusion that legislating against false information is a dangerous path, and one that would inevitably violate international standards on freedom of expression. In reality, there is no easy, quick-fix solution to societal harms. On the contrary, a combination of diverse expertise and viewpoints would

usually be required to design appropriate and careful responses. Nonetheless, the massive circulation of 'harmful content' such as 'hate speech' and disinformation has been shown to exacerbate real-world violence and divisions within societies. It is also known that this type of content is prone to going viral, which may be explained both as a consequence of natural human biases towards extraordinary or provocative information and as a result of the design of the platform. For instance, when Facebook decides to allocate more weight to 'angry' reactions (the anger emoji below posts on Facebook) than other reactions, the company has chosen to make the type of content that triggers negative reactions more visible in users' feeds.[10]

While platforms' current initiatives in relation to 'hate speech' and disinformation should not be disregarded out of hand, it is legitimate to expect social media platforms to do more to address these problems. In our view, when it comes to designing, implementing, or assessing content moderation measures, it is particularly important that the various stakeholders in this debate, including governments, companies, and civil society, engage in dialogue to ensure that responses to problematic content are compatible with the protection of freedom of expression and other fundamental rights. In this respect, new forms of multi-stakeholder or independent self-regulation models, such as SMCs, could be part of the solution.

This chapter examines the current state of content moderation practices, giving particular consideration to the findings from the research conducted in Bosnia and Herzegovina, Indonesia, and Kenya in 2021 in relation to the live issues of content moderation that were identified in each country, and paying special attention to the gap between global content rules and their application in the specific circumstances of the three countries. It then presents recommendations based on international standards. ARTICLE 19 considers that social media companies are in principle free to restrict content on the basis of freedom of contract, but that they should nonetheless respect human rights, including the rights to freedom of expression, privacy, and due process.

Observations on the current state of content moderation

Content moderation issues have generally been studied and documented by civil society and academic endeavours, and the results of such investigations have supported and informed the development of the recommendations presented below. While being critical of the flaws it identifies in content moderation practices, this report also takes into consideration the fact that social media platforms have provided individuals with unprecedented opportunities to exercise their right to freedom of expression, to share information and ideas, and to organise in communities and progressive social movements. In Bosnia and Herzegovina, for example, some positive initiatives have emerged to counter gender-based discrimination and encourage inter-ethnic relationships. The civil-society-led initiative Sve su to vještice ('It's all witches') offers humorous perspectives on the female experience as a form of media activism (communicating feminist and empowerment messages via humour and satire). The initiative is meant to represent a response to, and to voice disagreement with, the pervasive violence and scrutiny that women experience. Mainly active on Facebook, It's all witches has received growing interest from all genders (18% of users interacting with the Facebook page are men) and strengthened dialogue and solidarity among women on topics related to gender discrimination or violence. Another example of the use of social media platforms for good is the Balkan Discourse Platform led by the peace-building organisation Post-Conflict Research Center (PCRC) in Bosnia and Herzegovina. The platform engages with young activists in the country to collect local and community stories of courage, love, and heroism during the time of the armed conflict ('Ordinary heroes') or stories related to queer and inter-ethnic relationships in Bosnia and Herzegovina ('The love tales') to challenge existing stereotypes around ethnic divides, and increase dialogue and understanding among different ethnic groups in the country. In exchange, PCRC provides basic journalism and photography training to activists and gives them the freedom to choose the stories they would like to cover.

Impact on peace and stability: Disinformation and incitement to hatred

While 'fake news' and 'hate speech' are separate categories of 'harmful content' (both of which present definitional issues), the country reports found that disinformation and discriminatory speech are often interlinked[11] and that this type of content is often produced for political purposes. In Kenya, for instance, 'hate speech' and propaganda were observed to migrate from SMS and radio to the online world of blogs and social media platforms during the 2013 and 2017 elections. (Kenya, Bosnia and Herzegovina, and Indonesia will have general elections in August 2022, October 2022, and February 2024 respectively.) The spread of disinformation is used to exploit existing social, racial, and religious divisions, and such efforts are intensified during election periods. In Indonesia, a study by election watchdog Perludem conducted in collaboration with Facebook showed that disinformation about electoral procedures, aimed at delegitimising democratic processes, could seriously undermine the right to vote.[12]

It is worth noting that pieces of content that do not individually amount to a violation of social media companies' content rules may become problematic through repetition and amplification. The report on Indonesia pays particular attention to such 'grey-area' content and notes that people who seek to spread manipulative speech have become expert at carefully crafting messages that claim to promote democracy, freedom of expression, or the 'Unity in Diversity' motto of Indonesia, but in effect belittle opposition groups and contribute to increasing polarisation in society. Interviewees from Kenya observed that users had learnt to 'weaponise' social media to propagate problematic content while avoiding reactions by platforms. Such users' tactics included setting up groups on social media specifically for sharing problematic content; creating and coordinating troll armies or so-called 'keyboard warriors' to run smear campaigns, and threaten and intimidate certain individuals; manipulating content reporting tools and alleging copyright infringement; and circumventing detection measures by, for example, using multiple accounts and bots.

However, societal harms are not caused by the mere circulation of online speech: a study by PeaceTech Lab on youth and radicalisation in Mombasa noted that social media alone

did not advance violent extremism, but was a component in an interlocking network that could drive individuals towards a path of radicalisation for violence. Research also shows that the reasons why people share disinformation are complex. In that regard, one of the relevant factors in the Kenyan context is the common use of satire in political communication. A 2021 comparative study notes that 'though the boundaries between satire used for political ends and malicious or misleading information may be nebulous, the long social history of such practices in Africa makes this an important factor to consider. Given the entrenched role of satirical and humorous content in informal networks of media use in Africa, and the progressive uses to which these types of intentionally false—albeit not misleading—content have been put, media users on the continent might be less resistant to sharing information that they know is untrue.'[13]

Respondents in the Kenya research also highlighted that automated content moderation systems were unable to detect problematic content in local languages or to detect the nuances within these local languages. An abusive word in one language might not be abusive in another, because in some cases the same words had different meanings in different languages, and the meaning of certain words can change depending on the issues being discussed or evolve over time. As a result, 'hate speech' or disinformation posted in local languages could remain undetected by automated content moderation systems.

Research has underlined the importance of localising the understanding of what amounts to incitement to violence. In that sense, a recent UNESCO discussion paper on the challenges of moderating 'hate speech' on social media recommends empowering 'stakeholders and notably local communities to monitor and detect hate speech on social media tailored to their context and languages'.[14] Similarly, research financed by Facebook concluded that consultation with local groups was necessary in order to fully grasp the complexity of the local context and the degree of harm experienced by the groups targeted by such content.

Impact on marginalised voices

In the three country reports, the impact of problematic content on marginalised groups is clear.[15] The Kenya report notes that prominent female personalities in the media or political sphere are not only attacked for their opinions, but also based on their gender, sexuality, and appearance. A recent Kenya ICT Action Network (KICTANet) study concluded that online harassment of women had not gained attention from policymakers and the public, with outcries against harassment being dismissed as toxic feminist rants. Interviewees from Kenya also observed that LGBTQI+ communities were targeted by cyberbullying. The report on Bosnia and Herzegovina noted that groups including immigrants, Roma, and LGBTQI+ communities are regularly targeted on social media platforms, often for political purposes.

Accessibility of content rules in relevant languages

Platforms' content rules in local languages are not fully accessible. The research on Bosnia and Herzegovina notes that the community standards and terms of service, although they are key contractual documents between the platforms and their users, are not available in full in local languages, and the parts that are available appear to be poorly translated. In Kenya, only Facebook's community guidelines are available in Kiswahili while YouTube and Twitter rules are not – not to mention the numerous other languages spoken in the country. In Indonesia, translations of up-to-date versions of platforms' content rules do not appear to be consistently available.

Platforms' internal remedies

In each of the three countries covered by this project, interviewees mentioned instances of platforms taking down legitimate content. In such cases, it is acutely important that users have access to an effective way to engage with platforms in order to address these questionable removals.

For instance, in Bosnia and Herzegovina, a media outlet from Sarajevo was prevented from publishing content related to the International Criminal Tribunal for the former Yugoslavia

(ICTY) war crimes judgment in the case of Ratko Mladić, a notorious war criminal, charged with genocide in Srebrenica, because of a 'misinterpretation'. Facebook's algorithms came to the conclusion that the article sought to promote 'criminal organisations', which resulted in this label being applied to the ICTY, and the article was not published. The irony is that while the Facebook algorithm wrongly classified the ICTY as criminal, outlets like Despotovina.info – known for its glorification of Mladić – and SAFF – a Muslim religious magazine known for its anti-LGBTQI+ articles and Bosniak ethno-nationalist political agenda – were able to continue their activities on social media unhindered.

In all three countries, interviewees described remedies offered by the companies as not easily accessible and not necessarily effective. They said they often struggled to navigate platforms' internal complaint mechanisms. Respondents in Kenya observed that while platforms publicised the use of their services for entertainment, they did not place similar emphasis on reporting mechanisms. One interviewee in Bosnia and Herzegovina noted that reporting content issues in English led to a faster reaction from the platform than making the complaint in a local language. In Indonesia, a respondent reported that even public authorities found platforms' internal reporting mechanisms to be very complicated to use.[16]

According to Kenyan respondents, individual users who were able to contact representatives of social media platforms directly had more success in obtaining responses to their complaints. However, there is little transparency about who the representatives of social media platforms in the country are, and it may require a lengthy and intensive process of networking to find a personal connection to someone who works for a social media platform in order to get a response to one's complaint. Consequently, users who are unable to reach platforms' staff through personal or professional connections find themselves disadvantaged in their attempts to discuss questionable removals of content with social media platforms.

The three country reports also mention instances where internal reporting mechanisms have been abused. In Bosnia and Herzegovina, for instance, a media outlet that had published an extensive report on the financial malversations of an Islamist group observed that in response this group had published a how-to video inviting its followers to use the platform's mechanisms to report the investigation as content that should be taken down. Fortunately, the manoeuvre was not successful and the investigation report remained available on social media (but so did the how-to video posted by the Islamist group). However, the media had no effective way to alert the platform to the false complaints: their attempts to reach out to the platform's European offices remained unanswered.

Indonesian interviewees observed that platforms tend to react quickly and positively when faced with mass reporting, allowing internal remedies to be used as a tactic to silence legitimate voices. For example, in June 2021, Instagram reacted to user reports by taking down content posted by at least two accounts run by activists advocating anti-corruption efforts. These activists received a notification that their posts contained incitement to violence and thus infringed community guidelines. Civil society groups highlighted that the technique of reporting en masse against these activists was part of a counterattack on the strengthening of the Corruption Eradication Commission (KPK).

Indonesia: Algorithmic content moderation decisions on gender-based violence

The following two examples from Indonesia show how automatic content moderation may lead to wrong takedown decisions.

A representative from the Society Participation Division of the National Commission on Violence Against Women (Komnas Perempuan) complained that the anti-violence education material[17] that the organisation live streamed to YouTube was taken down by the platform. She explained to the researcher: "The live streaming was two hours long, from 10:00 to 12:00. However, in the first hour, YouTube accidentally cut off the live streaming and deleted the content of the first hour. Then we continued our live streaming again by changing the title of the streaming event using letters mixed with numbers (note: from violence into v10l3nc3 or k3k3r454n in the Indonesian language). After the live event was done, we renamed the title using the correct spelling."

She said that they tried to contact YouTube, but even this established public authority found the reporting and appeal mechanisms challenging and one-sided. They did not know the precise reason behind the takedown decision, but they presumed it was because the video used the word 'violence' in the Indonesian language ('kekerasan'). Consequently, the automated content moderation system identified the video as promoting violence.

On the other hand, the Head of External Communication of Arus Pelangi, a civil society organisation that promotes the protection of LGBTQI+ rights, mentioned during the interview their observations about content posted on Instagram by an organisation that campaigns against the protection of these minority groups: those posts stayed on the platform, presumably because the uploader typed 'violence' as 'v10l3nc3' ('k3k3r454n' in the Indonesian language) in the content.

External oversight

Some interviewees complained about the absence of oversight mechanisms that would force platforms to be accountable for the decisions they make about how content is moderated on social media. To date, Meta is the only social media company that has established an external oversight complaint mechanism – known as the Oversight Board. However, the question of this mechanism's long-term effectiveness remains open.

The Indonesia report notes that less than 8% of submissions to the Oversight Board come from the Asia-Pacific region. The Kenya report observes that the Oversight Board includes a Kenyan member, Maina Kiai, who, on his appointment, welcomed cooperation with local stakeholders and encouraged the public to raise issues for investigation by the Board. Out of the 524,000 cases submitted between October 2020 and June 2021, only 2% were from sub-Saharan Africa, and the Board admits that it does not believe that this 'represents the actual distribution of Facebook content issues around the globe'.[18]

Transparency and availability of data

While the level of transparency reporting by social media platforms has increased over the years, the country reports highlight the limitations of the information about content moderation that social media platforms release. In particular, social media companies do not publish data disaggregated by country on the reasons for and volume of content removal, on other forms of content moderation, or on the number of complaints received and the outcome of such procedures. It therefore remains difficult to assess the presence and circulation of 'harmful content' in each country or to evaluate the application of content rules at country level.

Similarly, little information is available on how algorithmic and human content moderation are organised. There is little to no transparency on how companies allocate moderation tasks per country, the number of languages moderators are conversant with, the specific issues they respond to, or where they are located. And companies do not provide much information on the role of automated systems in moderation.

Relations with Trusted Partners and other stakeholders

While Meta has opened offices in Kenya and Indonesia, the company has no presence in Bosnia and Herzegovina. Google and TikTok also have local offices in Kenya. Twitter has no local office in any of the three countries. (The Kenya report notes that Twitter has a single public policy official serving sub-Saharan Africa, who is based in Ireland.) Generally, it remains complicated for local stakeholders to reach out to social media platforms as there is little or no publicity on the existence of a country contact point.[19]

Social media companies have worked with local stakeholders and supported local initiatives in various ways. However, there is generally little information available on the conditions and sustainability of such engagement. For instance, in Kenya, Meta has collaborated with different organisations on areas such as child online protection, digital safety, Internet governance, 'hate speech', and disinformation. The Facebook Journalism Project and Reuters launched a free e-learning programme to train journalists in digital newsgathering, news verification and reporting, wellness, and resilience training while reporting. Some organisations in Kenya reported long-standing collaborations with Meta, such as the Watoto Watch Network for their annual Safer Internet Day, KICTANet for the Kenya Internet Governance Forum (KIGF), and PesaCheck, which is contracted to fact-check content on the company's platforms.[20] The report also notes a lack of transparency around Meta's Trusted Flagger programmes, including how to join, information on Trusted Flaggers per country and the categories of content they respond to, the actions taken by the platforms as a result of reports, and the content removed based on reports.

In Indonesia, local organisations are part of social media companies' initiatives, such as YouTube's Trusted Flagger programme, Facebook's Trusted Partner initiative, and Tik Tok's Child Safety Partner scheme. Twitter's Global Trust and Safety Council also invited Indonesian civil society organisations to join. The Indonesia report observes that giving civil society groups a special line of communication can enable them to contribute to resolving content moderation issues – as was the case in attacks on online social movements in Wadas village in the context of a conflict between the police and residents who reject the Bener Dam construction and mining plans affecting the village in Bener

District, Purworejo Regency in Central Java. SAFEnet received reports from Wadas residents and youth activists who sided with the residents that their Twitter accounts were being suspended due to mass flags. When they appealed these decisions, SAFEnet was able to support them by corresponding with Twitter to explain that those affected were credible and legitimate activists. Within a few days, Twitter reactivated the accounts and even verified the account of Wadas Melawan/Wadas Fights Back with a blue tick.

The Indonesia report also notes that while a Meta representative claimed in a public discussion that the company has 12 Trusted Partners in the country, the list of Trusted Partners is not publicly available and consequently the researcher was only able to identify and reach out to some of them. This confirms the findings of academic research into the transparency and inclusiveness issues of trust-related partners in social media.

In Bosnia and Herzegovina – with the exception of Raskrinkvanje, an official third-party checker for Facebook – there is minimal information about the company's relationship with local stakeholders. Interviewees have described an absence of interest in their country among social media companies. The dynamics of content moderation are therefore mostly unidirectional, with companies implementing their global content rules with little to no coordination with local actors. This disconnect is surely one of the reasons why the country has been left with a proliferation of online 'hate speech' and disinformation, coupled with a lack of (visible) content that promotes peace, tolerance, and public participation.

Bosnia and Herzegovina: The competing roles of the Press Council and a fact-checking organisation in ensuring media compliance with ethical standards in journalism

Raskrinkvanje (in the local language, the name means 'debunking') is the major fact-checking body in Bosnia and Herzegovina and an official member of the International Fact-Checking Network (IFCN). It is also part of the well-established local civil society organisation Zašto ne? ('Why not?'). Once Raskrinkvanje has flagged news content as disinformation, the consequences for the media outlet that published the news item fall outside its direct control, as they are directly managed by Facebook.

Media outlets are subject to Facebook's system of penalties, which it may impose once content has been flagged by fact-checkers. There was increasing dissatisfaction among national media organisations with Raskrinkvanje, which was accused of adopting non-transparent and arbitrary methods in its analysis, and this initial dissatisfaction soon escalated to harassment of fact-checkers. The industry position is that media companies should only have to respond to the Press Council about any complaint related to their adherence to ethical standards in journalism; they therefore question Raskrinkvanje's legitimacy in this process.

The goals of the Press Council and Raskrinkvanje are similar: reliability of information and sustainability of independent journalism. There is, however, some jurisdictional conflict between the organisations resulting from the increased importance of Facebook in the distribution of media content. The question now is whether the platform is ready to contribute to a resolution of the growing tension between these two national actors.

Resources allocated to content moderation

The country reports highlight a disconnect between social media companies and local communities, media organisations, and individual users. For instance, the Bosnia and Herzegovina report describes the lack of common ground between platforms and their

local users as a 'terra nullius'. Interviewees in Bosnia have remarked that "Social media companies do not even know that we exist, that their policies are affecting us" or that "We are on the margins of the global social media processes and we are not recognised as important actors."

Hinting at discrimination in the levels of resources allocated to content moderation in the Global North and Global South, respondents in Kenya also observed that the lack or limitation of investment in moderating content in local languages could be interpreted to mean that the needs of non-English speakers are not a key priority for these companies.

The difference in the levels of attention Meta pays to different geographical zones is well illustrated by a simple number: according to information brought to light by whistleblower Frances Haugen, 87% of the company's spending on misinformation is devoted to the English language although only 9% of the platform's users speak English. The 'Facebook papers' have shed light on the fact that the platform allocates very different levels of resources to content moderation in different countries. In the United States and a limited number of countries that are considered to be at high risk of political violence or social instability, 'Facebook offers an enhanced suite of services designed to protect the public discourse: translating the service and its community standards into the official languages; building AI classifiers to detect hate speech and misinformation in those languages; and staffing teams to analyze viral content and respond quickly to hoaxes and incitement to violence on a 24/7 basis'.[21] By contrast, 'other countries, such as Ethiopia, may not even have the company's community standards translated into all of its official languages. Machine learning classifiers to detect hate speech and other harms are not available. Fact-checking partners don't exist.'[22]

In terms of the consequences of under-resourced content moderation, the situation in Ethiopia is particularly revealing as a recent investigation has brought to light 'a litany of failures' in Facebook's content moderation. In the context of a conflict which has involved atrocities that have been described as ethnic cleansing, posts that incite violence or make false allegations in order to fuel hatred between ethnic groups have been allowed to circulate freely. According to this research, Facebook 'is said to have frequently ignored

requests for support from fact checkers based in the country and some civil society organisations say they have not met with the company in 18 months'.[23] While content flagged as disinformation by Facebook's partner fact-checkers (who are mostly based outside Ethiopia) led to quick action by the platform, no action was taken when a local fact-checking organisation reported content. The situation led the Oversight Board, in December 2021, to ask Facebook to conduct 'an independent human rights due diligence assessment on how Facebook and Instagram have been used to spread hate speech and unverified rumours that heighten the risk of violence in Ethiopia'. Quoting a previous decision related to Myanmar, the Board declared that 'in situations of armed conflict in particular, the risk of hateful, dehumanising expressions accumulating and spreading on a platform, leading to offline action affecting the right to security of person and potentially life, is especially pronounced. Cumulative impact can amount to causation through a "gradual build-up of effect", as happened in the Rwandan genocide.'[24]

The example of Ethiopia supports similar conclusions to those of our country reports: if local civil society actors were able to bring their experience and knowledge of the local context to content moderation processes, the risks linked to 'hate speech' and disinformation could be mitigated and real-world violence could even be averted. In response to criticism, Facebook declared that it has invested in safety and security measures to identify and swiftly remove content that violates its content rules. However, it is interesting to note that local Ethiopian civil society organisations have reported being ignored: 'Facebook organised a meeting with several groups in June 2020, to discuss how the platform could best regulate content before scheduled elections. As of November, two of the organisations involved said they had heard nothing about any subsequent meetings.'[25]

In addition to the unequal allocation of platforms' resources to content moderation in different geographical zones, it should also be noted that working conditions within the companies that implement content moderation on social media platforms are harsh, even appalling, as illustrated by a recent article in *TIME* about Kenya-based content moderation firm Sama. The author notes that 'despite their importance to Facebook, the workers in

this Nairobi office are among the lowest-paid workers for the platform anywhere in the world, with some of them taking home as little as $1.50 per hour, a TIME investigation found. The testimonies of Sama employees reveal a workplace culture characterized by mental trauma, intimidation, and alleged suppression of the right to unionize. The revelations raise serious questions about whether Facebook—which periodically sends its own employees to Nairobi to monitor Sama's operations—is exploiting the very people upon whom it is depending to ensure its platform is safe in Ethiopia and across the continent.'[26]

Recommendations on human rights and content moderation

This section begins by summarising the progressive development of the notion that social media companies are duty-bearers of human rights obligations and the requirements that come from such obligations. It then presents key recommendations for social media companies to address the flaws in content moderation practices identified in the previous sections.

The human rights obligations of social media companies

The Guiding Principles on business and human rights: Implementing the United Nations 'Protect, Respect and Remedy' framework (the Guiding Principles) provide a starting point for articulating the role of the private sector in protecting human rights on the Internet.[27] They recognise the responsibility of business enterprises to respect human rights, independent of State obligations or the implementation of those obligations, and recommend several measures that companies should adopt. These include incorporating human rights safeguards by design to mitigate adverse impacts, building leverage and acting collectively to strengthen their power vis-a-vis government authorities, and making remedies available where adverse human rights impacts are created.[28]

Freedom of expression mandate holders have addressed the role of social media platforms in promoting freedom of expression and recommended that they respect and promote the Guiding Principles in this regard. They have recommended, inter alia, that

companies should establish clear and unambiguous terms of service in line with international human rights norms and principles, produce transparency reports, and provide effective remedies for affected users in cases of violations. In 2018, the UN Special Rapporteur for freedom of expression also made it clear that companies should open themselves up to public accountability, suggesting that this could be done through SMCs.

The human rights mandate holders have also highlighted the responsibilities of companies in relation to specific content, such as 'violent extremism', 'fake news', online gender-based harassment and abuse, and 'hate speech'. They have called on companies to ensure that their users can easily access and understand company policies and practices, how they are enforced, and how they respect minimum due-process guidelines, and to ensure that these policies are directly tied to international human rights law. In his August 2018 report, the UN Special Rapporteur on freedom of expression recommended that social media companies should explicitly state where and how AI technologies are used on their platforms, services, and applications; publish data on content removals, case studies, and education on commercial and political profiling; and give individual users access to remedies for the adverse human rights impacts of AI systems.

In his April 2018 report, the UN Special Rapporteur on freedom of expression recommended that '(c)ompanies should incorporate directly into their terms of service and "community standards" relevant principles of human rights law that ensure content-related actions will be guided by the same standards of legality, necessity and legitimacy that bind State regulation of expression'.[29] As demonstrated by the scholarly debates,[30] the application of international human rights law to content moderation leaves many questions open beyond the principle.[31] But, as one scholar argues, '(o)ne of the key benefits of IHRL [international human rights law] in this context is that it can provide a common vocabulary for content moderation debates so that even as rules are contested and "the participants in these debates plainly disagree about which policies promote the public good[,] ... there is value to putting them in conversation with one another." (...) But the important caveat is that for argumentative practice to be successful, participants must

actually be *in conversation*. Creating legitimacy and accountability through argumentative practice requires an institutional structure that facilitates exactly this kind of argument and contestation.'[32]

Civil society has also developed recommendations on the requirements that social media companies should respect international human rights standards. For instance, the Manila Principles on Intermediary Liability elaborate the types of measures that companies should take to respect human rights. In particular, companies' content-restriction practices must comply with the tests of necessity and proportionality under human rights law and should provide users with complaints mechanisms to challenge companies' decisions. The Ranking Digital Rights project assesses the major Internet companies for their compliance with digital rights indicators, which include inter alia availability of terms of service; reasons for content, account, or service restriction; notification to users; a process for responding to third-party requests; and various transparency activities.

The Santa Clara Principles, first developed in 2018 by a group of civil society organisations and endorsed by a variety of Internet companies, including Apple, Facebook, Google, and Twitter, outline minimum standards to respect freedom of expression in content moderation, calling for Internet platforms to provide adequate transparency and accountability about their efforts to moderate user-generated content or accounts that violate their rules. In 2020–2021, these Principles were further expanded through the work of a larger civil society group[33] (Santa Clara Principles 2.0) to provide more operational guidance to Internet companies in relation to content moderation. They include a call to other stakeholders (such as governments or other State actors) to ensure the conditions for full transparency and due process are achieved. Foundational principles include the need for clear and transparent content moderation processes that respect human rights and due-process guarantees; outline understandable rules and policies on content moderation; acknowledge and operationalise cultural competence and understanding of local languages and societal contexts in content moderation practices; demonstrate awareness of the risks that may arise from State involvement in content moderation; and promote integrity and explicability in companies' content moderation systems. The Santa

Clara Principles 2.0 also reiterate the importance of transparency of data on content moderation, the need to give users' notice when their content is removed, and the need for an appeal mechanism to be available to users whose content is removed.

Key recommendations for social media companies

1. The principle of 'cultural competence' set forth in the Santa Clara Principles is of particular importance for this project. Interviewees in all three countries reported issues with companies' capacity to understand and take into consideration the complexity of the local circumstances in which content moderation cases arise.

This principle requires 'that those making moderation and appeal decisions understand the language, culture, and political and social context of the posts they are moderating. Companies should ensure that their rules and policies, and their enforcement, take into consideration the diversity of cultures and contexts in which their platforms and services are available and used.'[34]

This means that:

- All content rules as well as all reporting and appeals processes should be available to the users in the language in which they engage with the platform;
- Decisions on content moderation should be made by people familiar with the relevant language;
- Decisions on content moderation should be made with sufficient awareness and understanding of the linguistic, cultural, social, economic, and political dimensions of the relevant local or regional context;
- Content moderation processes should not disadvantage users on the basis of language, country, or region;
- Companies should demonstrate, through data published as part of their transparency reports, their cultural competence relevant to the users they serve, such as information on the languages and geographical distribution of their content moderators.

2. Companies should ensure that their content rules are sufficiently clear and accessible, and in line with international standards on freedom of expression and privacy. This includes ensuring that sanctions for non-compliance with their terms of service are proportionate.

3. In addition, social media companies should put in place internal complaint mechanisms, including for the wrongful removal of content or other restrictions on their users' freedom of expression. In particular, individuals should be given detailed notice of a complaint and the opportunity to respond prior to content removal. Internal appeal mechanisms should be clear and easy to find on company websites.

4. Companies should publish comprehensive transparency reports, including detailed information on their decision-making processes; the tools they use to moderate content, such as algorithms and Trusted Flagger schemes; and content removal requests received and actioned on the basis of their terms of service. They should also provide additional information on appeals processes, including the number of appeals received and their outcomes. Such information should be disaggregated on a per-country basis.

5. Companies should collaborate with other stakeholders to develop new, independent self-regulatory mechanisms, such as an SMC, modelled on effective self-regulation archetypes in the journalism field.

6. Finally, the question of local representation may raise specific concerns. ARTICLE 19 has observed that local laws sometimes require social media companies to establish a local presence in the country. This can become a matter of concern, particularly in countries where governments have poor records on the protection of freedom of expression, because national establishments may exert pressure on companies to remove content that is legitimate under international human rights law. Social media companies should make themselves easily and transparently accessible to local stakeholders through online means that enable local actors to engage with them effectively.

A local coalition on freedom of expression and content moderation

This chapter starts by presenting the idea of a local coalition on freedom of expression and content moderation as a way to bridge the gap between local stakeholders and social media companies, and to ensure that human rights and the relevant local context are appropriately integrated into content moderation decisions. As explained in the Introduction, this idea was raised with local stakeholders during the interviews conducted as part of this project, and their responses have contributed to suggestions about how to approach the development of a coalition in each of the three countries. Therefore, on the basis of recommendations from the three country reports, this chapter goes on to discuss an effective approach to facilitating the creation and development of such a coalition.

The role of a local coalition on freedom of expression and content moderation

In the context of the **Social Media 4 Peace** project, the notion that a local coalition on freedom of expression and content moderation could play a positive and effective role in enabling local civil society organisations and other local stakeholders to engage with social media companies in an effective way is based on the SMC. This is a model for a multi-stakeholder, voluntary-compliance mechanism for the oversight of content moderation on social media. In the course of developing the SMC model, opinions have differed on what its precise mission or form should be, but there has been broad agreement that such multi-stakeholder, transparent mechanisms could 'put the societal back into social media. They could establish fair, reliable, transparent and non-arbitrary standards for content moderation. At a time when decisions by social media companies increasingly structure our speech, councils could offer a comparatively swift method to coordinate and address pressing problems of democratic accountability. Creating a democratic, equitable and accountable system of platform governance will take time. Councils can be part of the solution.'[35] A local coalition on freedom of expression and content moderation pursues the same goal. In countries where the conditions are not necessarily met for the long-term efforts required to set up an SMC, and possibly as a

preliminary step towards the creation of such an institution, a local coalition would serve as a means to foster transparent and sustainable engagement between social media companies and local actors. This would contribute to ensuring that human rights and the various dimensions of the local context could be integrated into content moderation practices.

While the local coalition on freedom of expression and content moderation is, in the context of the **Social Media 4 Peace** project, inspired by the SMC model, other work in this area has similarly highlighted the need to establish forms of collaboration between social media companies and local civil society organisations.

For instance, a recent report on harmful content by Search for Common Ground, an organisation that focuses on peace-building, recommends the development of partnerships with in-country organisations 'with deep understanding of conflict dynamics to help identify and transform cultural and social barriers to content reporting'.[36] The report encourages social media companies to engage with local civil society by assigning a specific individual as a contact and holding regular meetings. Such regular contacts can facilitate the exchange of intelligence and the co-design of mitigating interventions that respond to harmful content.

A 2021 UNESCO discussion paper on the challenges of addressing 'hate speech' on social media recommends the creation of multi-stakeholder coalitions, principally in order to 'empower stakeholders and notably local communities to monitor and detect hate speech on social media tailored to their context and languages'.[37] The paper also encourages the facilitation of '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'.

The research in the three project countries has shown that local civil society organisations, even when they are engaged in partnerships with social media platforms (e.g., through the status of Trusted Flaggers), generally feel a serious imbalance of power between themselves and the giant companies. The local civil society organisations are aware of the

need for content moderation decisions to be informed by a solid knowledge of the multiple dimensions of the local context, and that knowledge of this context is something that local actors can bring to the table in conversations with tech giants. These conversations could extend to defining appropriate content moderation measures that can mitigate the risk of harmful real-world consequences while protecting freedom of expression. A coalition, by bringing together the multiplicity and diversity of civil society, can form a critical mass of local voices that could provide a national interlocutor for social media companies. Such a coalition would benefit both individual stakeholders, by increasing their capacity to be heard and take part in conversations, and social media companies, by allowing them to collect aggregated input through a single channel.

In the course of the research, the idea of a local coalition was discussed with a broad range of stakeholders in Bosnia and Herzegovina, Indonesia, and Kenya. During interviews, most respondents welcomed this suggestion. In Indonesia, for instance, in a discussion on how to handle 'grey-area' speech (posts or messages that do not in themselves amount to a violation of content rules but nonetheless may lead to disastrous consequences if allowed to circulate at a massive scale), a platform representative acknowledged that the application of content rules could be complemented by dialogue with locally relevant multi-stakeholder expert groups in order to guide content moderation processes.

Facilitating the creation and development of a local coalition on freedom of expression and content moderation

A successful coalition

Based on an OSCE practical guide to coalitions, the conditions for the success of a coalition are:

- '**A clear vision and mission**: Having an explicit vision that is created and shared by the whole coalition is critical to success.
- **Action planning**: If the coalition is to enact changes, it must draw up and implement action plans to realize its vision.

- **Developing and supporting leadership**: Successful coalitions understand that different voices can foster trust and legitimacy among different beneficiaries. They work to identify leaders across all coalition partners, continually develop leadership within the coalition and recognize that sharing leadership strengthens the coalition's ability to achieve its goals.
- **Documentation and ongoing feedback**: The coalition must track its activities and outcomes and provide regular feedback to all coalition members.
- **Technical assistance and support**: The most successful coalitions recognize when they need help and seek advice from consultants, outside facilitators and peers conducting similar work.
- **Securing resources**: Coalitions require some resources to be successful. (…) These resources may be secured through fundraising or by collecting in-kind contributions from coalition partners.
- **Making outcomes matter**: In successful coalitions, results matter most. It is critical to never lose sight of how the coalition's work will lead to changes in line with its vision.'[38]

While the observations detailed in each country report about facilitating the creation and development of a local coalition respond to these criteria in specific ways, the following general recommendations highlight the key areas of focus for any pilot coalition.

General recommendations

Vision and mission

In order to play the role assigned to it, a local coalition on freedom of expression and content moderation should be built around:

- The recognition of international standards on freedom of expression and other fundamental rights;
- A shared understanding of content moderation issues, in particular the negative impact on society of the disconnect between content moderation processes and the local circumstances in which specific cases have to be decided;

- A shared understanding of the theory of change for the coalition, namely that enabling local civil society organisations to engage transparently and sustainably with social media companies will contribute to content moderation practices that comply with international standards on freedom of expression and are informed by a clear understanding of all aspects of the local context.

Membership

- Coalition membership should be open to civil society organisations, academic institutions, journalists' associations, media and content producers, and private sector organisations that are interested in and committed to working on freedom of expression and content moderation.
- Special attention should be paid to existing coalitions and networks in order to avoid competition and identify possible complementarity.

Inclusivity

- The coalition should be inclusive, notably in its capacity to include and represent marginalised groups and the whole diversity of the society.

Capacity building

- In order to enable all stakeholders to take part in the coalition on an equal footing, training needs should be assessed at a very early stage of the formation of the coalition. As a minimum, training should cover: (a) international standards on freedom of expression and other fundamental rights; and (b) content moderation.
- Training needs may extend to topics such as coalition building and the operation of coalitions in countries where multi-stakeholder collaborations are not common.
- In the longer term, an additional role for the coalition could be to develop, promote, and implement digital literacy training for the general population, including marginalised communities.

Local ownership

- The coalition should be collectively owned, designed, and controlled by its members. The initial stages of its development should provide opportunities for potential members to fully discuss the findings of this research and drafts of all documents establishing the coalition.

Leadership and governance

- The leadership should represent the coalition without conflict of interest with any member organisation; it should be accountable to members.
- The governance, structure, and all procedures should be clearly established in a memorandum of understanding that all members agree on.
- The coalition should establish annual (SMART) objectives and work plans.
- The coalition should establish a clear external communication plan.
- Two-way internal communication processes should be established to ensure that the leadership can update members on developments and receive feedback from members.
- The coalition should have a detailed MEL (monitoring, evaluation, and learning) plan to ensure that it assesses its progress, measures the impact of its actions, learns lessons from its experiences, ensures the continued involvement of members, and adapts to changes in circumstances. Its annual reports should be made public.

Engagement with social media companies

While the presence of powerful global actors in the coalition might generate risks of capture or the fear thereof, it is nonetheless of the utmost importance that the coalition engages with social media companies in a transparent and sustainable manner. To that end, the coalition should:

- Campaign for social media companies to make available permanent online points of contact for the country; and

- Hold regular meetings with social media companies, and agree on a common annual work plan at the first meeting.

Funding

- There should be appropriate funding to ensure that the coalition is able to implement its work plan.
- Funding should be sustainable over the duration of the coalition.
- The coalition should start with a limited number of goals; secondary outcomes could be pursued at a later stage in the life of the coalition.

Role of public authorities

- The country reports discuss the possibility of involving certain State institutions – independent public authorities, for instance – in the establishment of the coalition. While there is a need to communicate or even coordinate with public authorities in their respective areas of competence, any risks to the independence, effectiveness, and credibility of the coalition must be averted. The coalition should only include relevant public authorities with the status of observers at specific meetings.

Research and monitoring

- As a way to develop capacity and knowledge, and to further build a common understanding of content moderation issues, the coalition could take part in research initiatives and programmes on content monitoring and content moderation.

Reach

At a later stage in its development, the coalition could:

- Engage with Facebook's Oversight Board with a view to bringing relevant national cases before the Board; and
- Develop interactions and relations between local stakeholders and regional and international actors on issues related to platform governance and content moderation.

Legal and regulatory framework

- At a later stage, and in coordination with other national and regional actors, including social media companies, the coalition could engage in advocacy initiatives to ensure that the legal and regulatory framework applicable to online content regulation complies with the requirements of international standards on freedom of expression and other fundamental rights.

Elections

- As electoral periods are characterised by an increase in the production and dissemination of 'hate speech' and disinformation on social media, the coalition could engage with social media companies and public authorities in initiatives that seek to ensure that content moderation processes serve the organisation and implementation of fair, inclusive, and just elections, and the protection and promotion of individuals' right to vote.

Social Media Council

- The coalition could in the longer term facilitate the creation of an SMC that will promote the development and enforcement of content moderation practices that uphold international standards on freedom of expression while being informed by a robust understanding of the local context.

Conclusion

Looking at the current state of content moderation practices in Bosnia and Herzegovina, Indonesia, and Kenya with a specific focus on 'harmful content' ('hate speech' and disinformation), this project has highlighted the existence of a disconnect between social media companies operating at the global level and local stakeholders. When these companies fail to take into consideration the various (linguistic, political, social, cultural, and economic) dimensions of local contexts, content moderation processes can have dramatic impacts on the societies affected, such as increasing polarisation and the risk of violence.

In that respect, in order to comply with their obligation to respect and protect human rights, social media companies should, as advocated by the Santa Clara Principles, ensure that those making moderation and appeal decisions understand the language, culture, and political and social context of the posts and messages that they are moderating. Social media companies should ensure that their rules and policies, and their enforcement, take into consideration the diversity of cultures and contexts in which their platforms and services are available and used.

This implies that:

- All content rules as well as all reporting and appeals processes should be available to users in the language in which they engage with the platform;
- Decisions on content moderation should be made by people familiar with the relevant language;
- Decisions on content moderation should be made with sufficient awareness and understanding of the linguistic, cultural, social, economic, and political dimensions of the relevant local or regional context;
- Content moderation processes should not disadvantage users on the basis of language, country, or region; and

- Companies should demonstrate, through data published as part of their transparency reports, their cultural competence relevant to the users they serve, such as information on the languages and geographical distribution of their content moderators.

As a means to implement these recommendations, local coalitions on freedom of expression and content moderation would provide a platform through which local civil society actors could engage in a sustainable manner with social media companies. Such collaborations would contribute to ensuring that content moderation processes are compatible with human rights and based on genuine understanding of the local context.

In the course of the research, most interviewees responded positively to the idea that a local coalition on freedom of expression and content moderation could play an effective role in bridging the gap between the local society and social media companies. Their contributions have helped to formulate recommendations on how to facilitate the creation and development of such coalitions in Bosnia and Herzegovina, Indonesia, and Kenya.

Endnotes

[1] In its second year (2022), the **Social Media 4 Peace** project is expected to extend to a fourth country: Colombia.

[2] In certain sections, the report may appear to devote more attention to Meta than other social media companies. This is due to the fact that Meta owns the social media platforms that are the most important in the three project countries. In that sense, an interviewee observed that in Bosnia and Herzegovina, "for many people, the Internet is Facebook." While this is not necessarily the case in Indonesia and Kenya, the combined importance of Facebook and Instagram make Meta the dominant social media company in these countries too. One should also note that, in a number of instances, interviewees preferred that we refer to "a social media platform" without naming the specific company they were talking about.

[3] For a discussion of content moderation remedies beyond the binary 'on-off' approach, see for instance Golman, E. 'Content moderation remedies', *Michigan Tech Law Rev*., 28(1) 1-59, 2021.

[4] See, for instance, Garfield, L. What you need to know about the Facebook papers, European Digital Rights, 17 November 2021.

[5] In the context of the **Social Media 4 Peace** project, UNESCO has separately commissioned research on the legal and regulatory framework applicable to content moderation in the project countries. The legal dimension of content moderation is not covered in ARTICLE 19's reports.

[6] See, for instance, Kettemann, M.C. and Fertmann, M. 'One council to rule them all: Making social media more democratic', *Encore, The Annual Magazine on Internet and Society Research*, 2021/2022, 10-18, 2021.

[7] In spite of the researchers' best efforts, it was not possible within the time frame of this research to arrange interviews with all the organisations identified initially.

[8] For ARTICLE 19's views on the development of legislative and regulatory approaches to the governance of content and platforms, see the following policy briefs:

- Side-stepping rights: Regulating speech by contract, 2018;
- Watching the watchmen: Content moderation, governance and freedom of expression, 2021; and
- Taming Big Tech, 2021.

[9] See, for instance, Centre for the Study of Media, Communication and Power, King's College London, Submission to: *Inquiry into fake news* – Culture Media and Sport Select Committee, 2017; Berstein, J. 'Bad news: Selling the story of disinformation', *Harper's Magazine,* September 2021.

[10] Merill, J.B. and Oremus, W. 'Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation', *The Washington Post*, 26 October 2021. The article notes that:

'Starting in 2017, Facebook's ranking algorithm treated emoji reactions as five times more valuable than 'likes,' internal documents reveal. The theory was simple: Posts that prompted lots of reaction emoji tended to keep users more engaged, and keeping users engaged was the key to Facebook's business. (...) The company's data scientists confirmed in 2019 that posts that sparked angry reaction emoji were disproportionately likely to include misinformation, toxicity and low-quality news.' (Facebook later stopped taking the angry emoji into account in its algorithms.)

[11] In a similar way, research conducted recently by ARTICLE 19 in Bangladesh showed that a majority of people believe that misinformation, particularly on social media, fuels communal violence in the country.

[12] Maharddhika and Salabi, N.A. Interference to vote rights: Phenomenon and responsibility [Indonesian], Perludem, 21 September 2021.

[13] Madrid-Morales, D. et al. 'Motivations for sharing misinformation: A comparative study in six sub-Saharan African countries', *International Journal of Communication*, 15(2021), 1200-1219, 2021.

[14] UNESCO, Addressing hate speech on social media: Contemporary challenges, 2021.

[15] For a global analysis of online violence against women journalists, see, for instance, UNESCO/International Center for Journalists, Online violence against women journalists, 2020; and ARTICLE 19, Equally safe: Towards a feminist approach to the safety of journalists, 2022.

[16] This is not uncommon. Public authorities that rely on social media to publicise their work may run into difficulties. For example, in February 2022, the Tribunal de Cuentas Federal de Brasil (Federal Court of Accounts of Brazil) had its YouTube account suspended, apparently due to a breach of the company's policies on intellectual property.

[17] Komnas Perempuan, Churches against sexual violence, [YouTube], [Indonesian], 3 July 2020.

[18] Oversight Board, Oversight Board transparency reports - Q4 2020, Q1 & Q2 2021, 2021.

[19] The question of local representation may raise specific concerns. ARTICLE 19 has observed that local laws sometimes require social media companies to establish a local presence in the country. This is a matter of concern, particularly in countries with governments that have a poor record on the protection of freedom of expression, because national establishments may exert pressure for the removal of content that would be considered legitimate under international human rights law. However, social media companies

could nonetheless make themselves easily accessible to local stakeholders through online means that would enable local actors to effectively engage with the companies.

[20] Interview, Allan Cheboi, October 2021.

[21] Newton, C. 'The tier list: How Facebook decides which countries need protection – Leaked documents reveal a huge, opaque system', *The Verge*, 25 October 2021.

[22] Ibid.

[23] Jackson, J. et al. Facebook accused by survivors of letting activists incite ethnic massacres with hate and misinformation in Ethiopia, The Bureau of Investigative Journalism, 20 February 2020.

[24] Oversight Board, Oversight Board upholds Meta's original decision: Case 2021-014-FB-UA, December 2021.

[25] Jackson, J. et al. 'Facebook 'lets vigilantes in Ethiopia incite ethnic killing'', *The Guardian*, 20 February 2022.

[26] Perrigo, B. 'Inside Facebook's African sweatshop', *TIME*, 14 February 2022.

[27] Ruggie, J. 'The Guiding Principles on business and human rights: Implementing the UN 'Protect, Respect and Remedy' framework, developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises', 7 April 2008, A/HRC/8/5A/HRC/17/31. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

[28] See the OHCHR B-Tech Project, which seeks to provide authoritative guidance on the application of the Guiding Principles in the technology space.

[29] UN General Assembly, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 6 April 2018, §45.

[30] Howell, J.P. 'The arrival of international human rights law in content moderation', *The Lawfare Podcast*, 27 May 2021; Aswad, E.M. 'The future of freedom of expression online', *17 Duke L. & Tech. Rev*., 26(2018), 2018; Land, M.K. 'Against privatized censorship: Proposals for responsible delegation', *Virginia Journal of International Law,* August 2019; Celeste, E. 'Digital constitutionalism: A new systematic theorization', *International Review of Law, Computers and Technology*, 33(1) 76-99, 2019; Land, M.K. 'Regulating private harms online: Content regulation under human rights law', in Jørgensen, R.F. (ed.) *Human Rights in the Age of Platforms*, 2019; Aswad, E.M. 'To protect freedom of expression, why not steal victory from the jaws of

defeat?', *77 Wash. & Lee L. Rev*., 609 2020, 2019; Sander, B. 'Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation', *Fordham Int. L. Jour.*, 43(4) 940-1006, 2020,; Benesch, S. 'But Facebook's not a country: How to interpret human rights law for social media companies', *Yale Journal on Regulation Bulletin*, 38(86), 2020; Karanicolas, M. 'Squaring the circle between freedom of expression and platform law', *Journal of Technology Law & Policy*, XX (2019-2020), 2020; Douek, E. 'The limits of international law in content moderation', *UCI J. Int'l Tran'l & Comp. L*, 6(1), 2021.

[31] One scholar notes that: 'One benefit of the online environment is that platforms can develop far more nuanced remedies than have traditionally been available to governments. Far beyond the false binary of choosing to leave content up or take it down, they can choose to label it, amplify it, suppress it, demonetize it, or engage in counter-messaging. IHRL encourages if not mandates platforms explore these options by requiring any measure restricting expression be necessary, in the sense that it is the least intrusive instrument. Again, however, to date it does not offer any guidance beyond this general command, but essentially all content moderation decisions will be within this grey zone. A jurisprudence can develop over time, but the platform lawyer charged with making these determinations today is left on their own.' Douek, E. 'The limits of international law in content moderation', *UCI J. Int'l Tran'l & Comp. L*, 6(1), 2021.

[32] Ibid.

[33] ARTICLE 19, Access Now, ACLU Foundation of Northern California, ACLU Foundation of Southern California, Brennan Center for Justice, Center for Democracy & Technology, Electronic Frontier Foundation, Global Partners Digital, InternetLab, National Coalition Against Censorship, New America's Open Technology Institute, Ranking Digital Rights, Red en Defensa de los Derechos Digitales, and WITNESS.

[34] The Santa Clara Principles on transparency and accountability in content moderation, no date.

[35] Tworek, H. Social media councils, Centre for International Governance Innovation, 28 October 2019.

[36] Institutional Learning Team, Handling harmful content online: Cross-national perspectives of users affected by conflict, Search for Common Ground, April 2021.

[37] UNESCO, Addressing hate speech on social media: Contemporary challenges, 2021.

[38] OSCE/ODIHR, Coalition building for tolerance and non-discrimination: A practical guide, 2018.