United Nations University
Centre for Policy Research

Hate Speech Case Study

2024 PBF Thematic Review: Synergies between Human Rights and Peacebuilding in PBF-supported Programming

Lauren McGowan, Erica Gaston, Adam Day, and Luisa Kern

This case study is an excerpt from a larger 2024 PBF Thematic Review examining synergies between human rights and peacebuilding. The overall Review examined a select sample of PBF programming – 92 projects implemented in 45 countries and territories – that were supported between 2017 and 2022, with a view to collecting best practices and lessons learned, and contributing to better understanding of how human rights and peacebuilding tools and strategies may complement each other in advancing peace and preventing conflict. This thematic case study examined a sub-sample of those 92 projects, those which had activities or objectives related to countering hate speech, disinformation or misinformation. This case study appears on pages 47 to 58 of the full report.

The Peacebuilding Fund

The Peacebuilding Fund (PBF) was established in 2006 by the Secretary-General at the request of the General Assembly as the primary financial instrument of the UN to sustain peace in countries at risk of or affected by violent conflict. The PBF provides funds to UN entities, governments, regional organizations, multilateral banks, national multi-donor trust funds, and civil society organizations. From 2006 to 2023, the PBF has allocated nearly $2 billion to 72 recipient countries.

Since 2006, PBSO has commissioned Thematic Reviews to examine past practices and promising innovations in peacebuilding, and to reflect on the performance of the PBF in designated areas. The Review that this case study was part of was commissioned by PBSO in partnership with OHCHR and the Government of Switzerland. Research was led by United Nations University Centre for Policy Research (UNU-CPR), and conducted between January and October 2023. Full methodology details are provided in the full Thematic Review.

Rising hate speech, disinformation, and misinformation have already demonstrated the potential to disrupt key peacebuilding and transition processes, contribute to electoral violence, exacerbate intercommunal conflicts, and create negative consequences across the rights spectrum. In June 2023, the UN Security Council recognized that hate speech can contribute to "driving the outbreak, escalation and recurrence of conflict" and undermine peacebuilding efforts.[1]

For all these reasons, programming efforts to counter hate speech has gained increasing attention in the human rights and peacebuilding field. While this is still an emerging area of work, the findings from this case study of 12 ongoing and recent projects suggests that programming around hate speech can be very important for early warning and preventive action, particularly in electoral contexts. It also can be a crucial counterpart to other efforts to encourage social cohesion as a means of conflict prevention.

However, an important suggestion from the analysis of these projects is that **programming to counter hate speech and disinformation is at its strongest where it**

The full Thematic Review is available at: https://unu.edu/cpr/report/2024-pbf-thematic-review-synergies-between-human-rights-and-peacebuilding-pbf-supported.

UNU CPR
PEACEBUILDING FUND
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Federal Department of Foreign Affairs FDFA
UNITED NATIONS HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

gives equal attention to human rights risks and strategies and conflict prevention aims. This is what would enable counter-hate speech programming to contribute not only to immediate violence prevention but also to addressing the root causes of hate speech and violence.

Background: Conceptualizing Hate Speech and Its Impacts on Human Rights and Peacebuilding

There is no international legal definition of hate speech.[2] Nonetheless, the following definitions, based on UN guidance, help illustrate the distinctions between the terms hate speech, disinformation, and misinformation, which in practice are sometimes conflated:[3]

- Hate speech – any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, on the basis of their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor.
- Disinformation – information that is not only inaccurate but is also intended to deceive and is spread in order to inflict harm.
- Misinformation – the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods.[4]

Anything can be the subject of disinformation and misinformation, but only a person or a group can be the subject of hate speech. While recognized as distinct phenomena, policy and programming documents often discuss hate speech and disinformation (and to a lesser extent misinformation) collectively, recognizing interactive effects between them.[5] **This case study predominantly focuses on programming to counter hate speech, because this was the focus of 11 of the 12 projects** (all but the project implemented in the Central African Republic (CAR), PBF/CAF/H-1). For this reason, the analysis and findings generally relate to and refer solely to hate speech. The terms disinformation and misinformation will be introduced when relevant.

Hate speech and disinformation have gained increasing attention given their association with outbreaks of violence and human rights violations. Hate speech has been seen as a "precursor" to atrocity crimes, including in Rwanda, Bosnia and Herzegovina, and Cambodia.[6] In Myanmar, hate speech on Facebook helped fuel atrocities against the Rohingya in 2017.[7] In addition, hate speech and disinformation have helped trigger violence and destabilized other peacebuilding efforts during electoral periods or other transition moments.[8] In Côte d'Ivoire, ethnically based hate speech magnified political divisions and contributed to violence before the 2020 election.[9]

Hate speech and disinformation also have significant implications for the exercise of individual or collective rights. They can undermine or create barriers to the right to participate in political and public life, or in economic and social spheres.[10] Hate speech and disinformation can also deter or inhibit particular groups' or individuals' ability to fully access certain rights and can undermine inclusion more broadly.[11] Linkages are often drawn between shrinking civic space and prevalence of hate speech – and vice versa. For example, in Guatemala, hate speech against HRDs contributed to trends in the reduction of civic space between 2019 and 2022.[12]

> Hate speech can contribute to "driving the outbreak, escalation and recurrence of conflict" and undermine peacebuilding. – UN Security Council, resolution 2686 (2023).

Hate speech and disinformation can also have broader societal effects by eroding trust and social cohesion, which can undermine democratic institutions. A number of public bodies, including the UN General Assembly, have recognized that disinformation can undermine credibility and trust in electoral processes and impede people's ability to make informed decisions.[13]

Projects Related to Hate Speech and Disinformation

In response to the global trend of rising hate speech, the Secretary-General launched the United Nations Strategy and Plan of Action on Hate Speech in 2019.[14] He appointed the UN Office on Prevention of Genocide and Responsibility to Protect to be focal point for its implementation and established a UN Working Group comprised of 16 UN entities.[15]

Between 2017 and 2022, the PBF invested $58.2 million in 24 projects that include a countering hate speech component.[16] This case study features 12 projects spanning 15 countries and territories in Africa, Asia, Europe, and Latin America, including one regional project (Western Balkans: PBF/IRF-475-476-477-478-479).[17]

Table 4: Projects in the Hate Speech Case Study			
Project Code/ Duration	**Countries and Territories**	**Title***	**Implementing Agency**
PBF/CAF/H-1 (2019-2021)	CAR	Communication and awareness for social cohesion	UNFPA, UN Women, SFCG
PBF/CIV/D1 (2020-2021)	Côte d'Ivoire	Young people as drivers of hate speech prevention	UNICEF, UNDP, UNESCO
PBF/IRF-307 (2019-2021)	Guatemala	Creating new avenues of resilience to sustain peace: Kaqchiquel, Q'eqchi' and mestizo women pathfinders for peace at the center	UN Women, ILO, UNODC
PBF/IRF-453 (2022-2023)	Kenya	Enhancing Early Warning and Prevention to Counter Hate Speech and Incitement Ahead of the 2022 Elections in Kenya	UNDP, OHCHR
PBF/IRF-482 (2022-2024)	Liberia	Promoting Peaceful Electoral Environment and Community Security in Liberia	IOM, OHCHR, UNDP
PBF/IRF-481 (2022-2024)	Moldova	Building sustainable and inclusive peace, strengthening trust and social cohesion in Moldova	OHCHR, UN Women, UNDP
PBF/IRF-338 (2019-2021)	Myanmar	Empowering young men and women to advocate for peace and challenge hate speech in Myanmar	Christian Aid Ireland
PBF/IRF-367 (2020-2023)	Myanmar	Preventing hate speech and promoting peaceful society through media and information literacy	UNESCO, UNDP
PBF/SLE/B-11 (2022-2024)	Sierra Leone	Promote the creation of an enabling environment for [...] peaceful elections and the strengthening of social cohesion in Sierra Leone	UNICEF, UNDP
PBF/IRF-427 (2021-2022)	Sri Lanka	Countering hate speech through education and advocacy for improving social cohesion in Sri Lanka	UNICEF, UNDP
PBF/GMB/D-2 (2020-2022)	The Gambia	Young women and men as stakeholders in ensuring peaceful democratic processes and advocates for the prevention of violence and hate speech	UNFPA, UNDP, UNESCO
PBF/IRF-475-476-477-478-489 (2022-2024)	Western Balkans	Strengthening the role of youth in promoting increased mutual understanding, constructive narrative, respect for diversity, and trust in the region	UNFPA, UN Women, UNDP, UNESCO

* Titles in Spanish and French were translated by author.

As noted, 11 out of the 12 projects were focused on countering or responding to hate speech, with little to no emphasis on disinformation or misinformation.[18] The project in the CAR (PBF/CAF/H-1) is the one project that is focused on countering disinformation and misinformation and disinformation. The project stemmed from concern that disinformation about the country's peace agreement had the potential to undermine public participation and engagement in realizing its components, thereby undermining prospects for advancing peace.[19]

Of the projects examined, most were ongoing and four were only approved within a few months of the beginning of this Thematic Review.[20] Nonetheless, the 12 projects examined reveal several important issue areas in this emerging space: hate speech related to electoral violence; to youth vulnerability and inclusion; to ethnic, religious, or political fault lines and discrimination; and to gender-based hate speech. Each of these areas is explored below, together with a box with additional information on emerging practice related to technology and social media engagement in such programming.

Hate Speech in the Context of Electoral Violence

Hate speech has been linked to electoral violence in a number of countries and is one of the most significant areas of emerging counter-hate speech and peacebuilding work. Five projects were centred around detecting or countering hate speech in the context of forthcoming elections: in Côte d'Ivoire, Kenya, Liberia, Myanmar, and Sierra Leone.[21]

The project in Kenya, PBF/IRF-453 (implemented by UNDP and OHCHR), is a key example of emerging trends in programming on hate speech and elections. Hate speech was used as a tool of incitement in the contested 2007–2008 elections in Kenya, contributing to post-election violence that left more than 1,000 people dead, and hundreds of thousands displaced.[22] The Kenya National Commission on Human Rights also documented increased levels of hate speech, incitement, and ethnic profiling leading up to the 2017 elections, which resulted in dozens of casualties and hundreds of cases of sexual violence.[23]

In the wake of both elections, independent monitors and the UN Country Team recommended stronger policies and national legislation related to hate speech, and the establishment of platforms that would detect and respond to hate speech as a tool for early warning and conflict prevention.[24] In 2021, the UN Kenya Country Office developed a Plan of Action for countering hate speech and incitement in relation to the 2022 elections.[25] The project contributes to this larger strategy by supporting national institutions to improve their early warning and response capacities with regard to hate speech. The programming has a significant focus on artificial intelligence (AI) based analysis and detection, made available to and enhanced by national and subnational response networks.

Both the Kenya project and another project in Sierra Leone (PBF/SLE/B-11, implemented by United Nations Children's Fund (UNICEF) and UNDP) illustrate a prominent theme within counter-hate speech programming: enhancing national early warning and response systems, with a view to reducing violence around elections. The election-related project in Liberia, PBF/IRF-482 (implemented by IOM, OHCHR, and UNDP), also sets up early warning activities, though the focus is principally around working with youth and women at the grass-roots level.

While early warning was prominent in the projects related to elections, it is not the only strategy for electoral violence prevention. The project in Myanmar, PBF/IRF-367 (implemented by UNESCO and UNDP), is geared towards creating an inclusive media ecosystem for the electoral period and establishing a multi-stakeholder platform to lead long-term inclusion efforts. Meanwhile, the project in Côte d'Ivoire, PBF/CIV/D1 (implemented by UNICEF, UNDP, and UNESCO), focuses on engaging youth groups and leaders in identifying hate speech around the elections and offering more positive narratives on social spaces.

A final theme surrounds gender-based hate speech during electoral periods. In the context of Kenyan elections, gender-based hate speech was prominent, with particular repercussions for women's participation in elections, the instigation of sexual violence, and "online gender-based violence".[26] Implementing partners from the project in The Gambia, PBF/GMB/D-2 (implemented by United Nations Population Fund (UNFPA), UNDP, and UNESCO), Kenya, and Côte d'Ivoire, as well as other experts interviewed, stressed that during elections, hate speech is often directed towards women seeking public office, which can result in lower levels of participation, both as candidates and in the voting process.[27]

From 2020 to 2021, in Côte d'Ivoire, more than 2,673 pieces of false information on social media networks were reported by newly trained young bloggers.

A big focus of many of the counter-hate speech projects that took place in electoral contexts was not only to monitor and remove hate speech that might contribute to inciting violence, but also to combine this with outreach to communities to enable other forms of early warning, as captured in the photo above, from the project Kenya (PBF/IRF-453). Photo provided by UNDP Kenya.

At the time of research, three of the five projects were still ongoing, offering some limitations on the ability to extract overall findings on this stream of work. Nonetheless, the projects in Kenya and in Côte d'Ivoire (both of which had closed) suggested some positive short- and long-term results. The final evaluation for the project in Kenya (PBF/IRF-453) reported that by enhancing monitoring capacities and through engagement with social media companies, over 800 cases of "hate speech, incitement, and mis/disinformation" were identified and addressed.[28] It also noted that the project's use of technology and AI sparked interest from both MONUSCO and the UN Multidimensional Integrated Stabilization Mission in Mali (MINUSMA), who asked for further information to inform similar work in the Democratic Republic of Congo (DRC) and Mali, respectively.[29] From 2020 to 2021, in Côte d'Ivoire, more than 2,673 pieces of false information on social media networks were reported by newly trained young bloggers through the use of technological tools which contributed to an overall reduction of inflammatory discourse on social media, according to the final evaluation.[30] It also noted that the technological tools used created a community for young people through which they will continue to monitor hate speech and thus have an impact on broader early warning and conflict prevention efforts.[31]

Overall, experts and practitioners saw counter-hate speech programming as an important emerging area within electoral contexts, given the substantial implications for human rights and peacebuilding.[32] Counter-hate speech programming in these contexts was thus identified as being in need of further investment.

Youth and Hate Speech

Another theme across the projects examined, and more broadly in the field, relates to youth vulnerability to the effects of hate speech. The projects in Côte d'Ivoire, The Gambia, the Western Balkans (implemented by UNFPA, UN Women, UNDP, and UNESCO), and one of the projects in Myanmar (PBF/IRF-338) (implemented by Christian Aid Ireland) highlighted how young people's exclusion from governance and decision-making processes is a trigger of conflict and a driver of hate speech.[33] The project in the CAR (PBF/CAF/H-1) (implemented by Search for Common Ground, UNFPA, and UN Women) also links the spread of "rumours and false information" regarding the 2019 peace agreement to the exclusion of young people from the peace process.[34]

In addition, both because youth may be more likely to receive their information from online or social media spaces, and because of socioeconomic disadvantages or marginalization

particular to young people, **youth may be more vulnerable to hate speech and incitement than other groups.**[35] For example, the ProDoc for the regional project in the Western Balkans (PBF/IRF-475-476-477-478-479) noted that hate speech is the "most common form of violence or discrimination" faced by youth in the region.[36]

Several of the projects also identified a troubling intersectionality: **youth who are vulnerable based on certain identities or characteristics (e.g. ethnic, political, or sectarian) may be at greater risk of being affected by hate speech than other age categories.** For example, the ProDoc for PBF/GMB/D-2 argued that youth have been the most impacted by ethnic and/or religiously motivated hate speech connected to the reform process in The Gambia.[37] The project in Myanmar, PBF/IRF-338, also paid attention to this intersectional aspect, considering how youth were particularly affected by hate speech in the context of rising intercommunal and religious conflict between Buddhist and Muslim communities.[38]

Several projects were guided by the idea that, while youth may be more vulnerable to ethnically charged hate speech, they may also have the most potential to address it and become peacebuilders. The ProDoc for one of the projects in Myanmar (PBF/IRF-338) notes that youth have been at the forefront of campaigns to counter hate speech, despite being on the periphery of public decision-making.[39] Youth are thereby framed as potential "change-agents", and engaging young religious leaders to monitor and respond to hate speech is presented as a way to address religious dimensions of the conflict in Myanmar.[40] Other projects take a similar approach. For example, the project in The Gambia (PBF/GMB/D-2) responds to a rise in hate speech since 2016 by combining mechanisms that might empower youth to counter hate speech with efforts to address the root causes of conflict by increasing youth participation and inclusion in governance and decision-making.[41]

The evaluations for the Côte d'Ivoire (PBF/CIV/D1), The Gambia (PBF/GMB/D-2), and Myanmar (PBF/IRF-338) projects shed some light on how well this theory of youth as "change-agents" played out. All three evaluations noted anecdotal evidence of youth taking a more proactive role in countering hate speech and peacebuilding-related dynamics that were at issue in the project.[42] For example, youth participants in The Gambia project appeared more prepared and active in online fact-checking and some participants gave examples of their greater mediation efforts within their communities.[43] The evaluations also pointed to other positive changes in the environment, including evidence of greater youth resilience to hate speech and improved social cohesion in the Côte d'Ivoire project;[44] and evidence that hate speech had gone down and interreligious solidarity had increased in Myanmar.[45]

There was insufficient evidence to draw a causal link between these macrochanges and youth engagement as a result of the project activities, particularly in Myanmar, where these changes may have been equally affected by the change in political dynamics following the coup.[46] Overall, the evaluations tended to see the project activities as having supported the youth movements in question, possibly contributing to changes over time. However, the evaluations also raised the point that the short duration of programming may limit the degree to which these effects endure and result in any sustained changes in youth behaviour.[47]

Hate Speech in the Context of Political, Ethnic, or Religious Divisions

Hate speech related to or used to exacerbate ethnic, religious, or political strife is a cross-cutting theme.[48] Within several of the election-related projects, it was the fact that hate speech played into and exacerbated ethnic (e.g. Kenya) or communal and religious divisions (e.g. Myanmar) that created the "nexus" between violence and elections.[49] In the project in The Gambia, which related to both ongoing elections and other reform processes, the issue addressed was the "rising tide of ethnic and religious based hate rhetoric" that sharpened divides and undermined peacebuilding.[50] The project in the Republic of Moldova, PBF/IRF-481 (implemented by OHCHR, UN Women, and UNDP), was developed in response to the "considerable spike in hate speech" since the outbreak of conflict in Ukraine, which had triggered underlying ethnic, linguistic, and political divisions within Moldovan society.[51]

Programming in this area has sought to promote counter-narratives; open avenues for social, economic, or political inclusion; and to use counter-hate speech programming as a complement to other activities aimed at strengthening social cohesion.[52] The project in Sri Lanka, PBF/IRF-427 (implemented by UNICEF and UNDP), is illustrative of this approach. In 2018 and 2019, hate speech and disinformation on social media "fanned existing ethno-religious tensions" and fuelled communal violence in Sri Lanka.[53] Social divisions, and the potential for hate speech to ignite them, has only ratcheted up with the economic crisis arising from COVID-19.[54] In response, the Sri Lanka project adopted a multidimensional approach, working through both offline and online platforms to support information awareness and (primarily CSO) monitoring of hate speech and social cohesion indicators, positioning

A cross-cutting theme of projects in the counter-hate speech case study was to engage youth in dialogue about the effects of hate speech and misinformation, as in this exchange between youth councils in the regional Western Balkans project. Photo provided by UNDP/UNFPA Albania.

media and CSOs to promote positive counter-narratives and develop evidence-based advocacy, and generally support "safer and more inclusive spaces" for speech (online and offline).[55]

The project in the Republic of Moldova, PBF/IRF-481, combined capacity-building on responding to hate speech (for both duty-bearers and CSOs) with broader strategies for encouraging tolerance and creating space for dialogue about the political faultlines that were driving hate speech. Many of those who bore the brunt of the spike in hate speech since the Ukraine conflict broke out were Ukrainian refugees and other associated minority groups. In response, the project included sensitization on non-discrimination and the risks of hate speech with school personnel in areas with larger refugee populations, and also engaged Ukrainian refugees in community "deep-listening" exercises, where they were able to share their life experiences with their host communities.

The Western Balkans project, PBF/IRF-475-476-477-478-479, had similar activities and programming components, but took a regional approach. For example, there was emphasis on hosting regional dialogues, exchanges, and cultural events to support youth and activists in promoting positive counter-narratives, understanding, and tolerance of diversity.

All of the projects highlighted in this category of work were ongoing at the time of research; however, interviews with implementing partners and experts involved suggested some important emerging lessons. Many of these projects focused on short-term responses that could be accomplished within the scope of a project – for example, monitoring hate speech, supporting counter-narratives, and encouraging dialogues. It is difficult to measure the impact of activities such as supporting counter-narratives; however, several of those interviewed expressed doubt about whether they were really effective in countering or mitigating issues that stemmed from deeper-seated political or communal fault lines.[56]

Others observed that because hate speech programming tends to focus more on immediate monitoring and violence prevention, it can be poorly positioned to address deeper issues that may be driving hate speech linked to ethnic, political, or religious divisions. **Projects that focus only on the hate speech itself, rather than the deeper grievances**

and fault lines driving it, run the risk of focusing too much on the symptoms, rather than the underlying cause. Countering hate speech in such situations may require a more long-term and root-cause approach. While some of the projects examined appeared aware of this deficit within counter-hate speech programming and tried to correct for it,[57] the majority of projects examined still focused more on short-term interventions.

Gender-based Hate Speech

Gender-based hate speech can be conceptualized as part of the continuum of GBV and, like most forms of GBV,[58] has been on the rise since the COVID-19 pandemic.[59] For example, a report by UN Women showed that from March to June 2020 the rate of online hate speech targeting women in South and South-East Asia increased by 168 per cent compared with the same period in 2019.[60] A report by UNESCO showed that disinformation and online violence, including hate speech targeting women journalists, has also increased since the onset of the pandemic.[61] In response to these concerning trends, in June 2023, the UN Office of the Special Adviser on the Prevention of Genocide launched The Plan of Action for Women in Communities to Counter Hate Speech and Prevent Incitement to Violence that Could Lead to Atrocity Crimes.[62] **A need to do more to address gender-based hate speech was the most commonly cited challenge by hate speech experts interviewed.**

The project in Guatemala, PBF/IRF-307 (implemented by UN Women, International Labour Organization, and UNODC) is an interesting example of how considering gender-based hate speech can contribute to the broader programming objectives of addressing inequal political access for women and violence against women. At its core, the project is about the protection and participation of women in political and social spaces, particularly Indigenous and mestizo HRDs. However, the recognition that this must also extend to protection from online platforms was a novel approach. Those who worked on this project noted that this was the first time they had incorporated interventions to address gender-based hate speech within a PBF-supported project, and that they developed this approach specifically because of feedback from women stakeholders that online hate speech was a barrier to entering political spaces dominated by men.[63]

Other relevant practices among the projects were the operationalization of a "Women's Situation Room" in Liberia (PBF/IRF-482) and the Sri Lanka (PBF/IRF-427) project's model for capturing gender-disaggregated data when monitoring hate speech, which enables more targeted responses for gender-based hate speech.[64] The former enabled women to be directly involved in early warning and efforts to mitigate violence against women, including hate speech against female candidates and politicians.[65]

Two of the projects work to address masculinities in relation to hate speech.[66] The project in the Western Balkans (PBF/IRF-475-476-477-478-479) sought to address gender-based and homophobic hate speech by tackling gender norms and "toxic and militarized masculinities in the region".[67] The Guatemala (PBF/IRF-307) project worked on transforming communities' understanding of masculinities and acclimating or sensitizing men to women's leadership capacities with a view to turning them into potential allies.[68]

That two of the 12 projects considered masculinities is notable **– globally, not enough attention has been paid to working with men and boys on perceptions of masculinity, as a way to address gender-based or homophobic hate speech.**[69] The findings from this work were not yet available, but may offer important lessons for this area in future.

Last, it is worth noting that although three projects explicitly mention LGBTQI+ communities and their vulnerability to hate speech, none of the projects were focused on exploring means to counter this.[70] Thus it would be difficult to extrapolate lessons learned in this sub-area for this Thematic Review. Experts highlighted this as an important area, provided sufficient "do no harm" considerations are accounted for.[71]

Technological Tools and Social Media Partnerships

The use of technological tools, or partnerships with social media companies and use of their monitoring tools, is an important part of this emerging field of work. All of the projects examined relied to some extent on analytical or data-driven technological tools to monitor and counter hate speech. Several of the projects relied on global tools developed for use in counter-hate speech programming, including iVerify, a UNDP-developed fact-checking tool for identifying hate speech, disinformation, and misinformation;[72] and the UNICEF U-Report, which aims to engage youth in discussions and dialogue on hate speech and related conflict triggers.[73] The former was used in the project in Liberia (PBF/IRF-482) and the latter in the project in Côte d'Ivoire (PBF/CIV/D1). Both were used in the project in Sierra Leone (PBF/SLE/B-11).

In other projects, implementing partners engaged directly with global social media companies and used their social media tools to help monitor hate speech or disinformation.[74] For example, the projects in Kenya (PBF/IRF-453), Myanmar (PBF/IRF-338), and Sri Lanka (PBF/IRF-427) use a tool owned by Meta called CrowdTangle.[75] CrowdTangle is designed to follow, analyse, and report on content across Facebook, Instagram, and Reddit.[76] The project in Sri Lanka also uses Meta's Trusted Partners programme[77] and YouTube's Trusted Flaggers programme.[78] These programmes give priority to CSOs to monitor and report content that may violate the companies' policies.

There were two major concerns raised with the growing use of technological tools. First, many were concerned about the risks of these tools being used to take down lawful as well as unlawful speech, inadvertently restricting free speech rights. This issue connects to broader concerns with counter-hate speech programming and is discussed further in the concluding section.

Second was the issue of sustainability. A number of the projects developed their own bespoke, project-specific technological tools – for example a national online fact-checking platform was used in the project in The Gambia (PBF/GMB/D-2).[79] In some cases, there may be no alternative to bespoke tools, given specific project needs. For example, for the hate speech project in Myanmar (PBF/IRF-338) it was necessary to develop a Natural Language Processing algorithm that could identify hate speech in Burmese.[80] However, in general, developing bespoke tools for two-year projects was viewed as raising larger sustainability and compatibility concerns. By contrast, projects that connected with larger technological platforms or used existing global tools had a greater chance of their activities and benefits being taken up by other actors and continued after the project life cycle. For example, the project in Côte d'Ivoire (PBF/CIV/D1) utilized U-Report, which had been operating in the country since 2017. As of July 2023, there were 4,116,371 U-Reporters (users) in the country.[81]

Despite these concerns, there was overall a positive view on the innovative use of technological tools, and of the way that these tools enabled partners to expand peacebuilding work – such as dialogues, civil society networking, and positive peace messaging – into the virtual arena.

Findings: What Did We Learn?

Counter-hate speech programming shows promise in contributing to early warning and conflict prevention, particularly in electoral contexts. Projects like those in Kenya and Côte d'Ivoire appeared to have some success in monitoring, detecting, and generating responses to hate speech in ways that contributed to national or community early warning systems and overall conflict prevention efforts in those contexts.[82] The most prominent critique of the existing early warning efforts was that they would be even more impactful if better integrated and connected with national, regional, or even international prevention mechanisms – a constructive critique that suggests a need to reinforce the work and better link it in future.

Although less mature than the counter-hate speech work in electoral contexts, the findings also showed strong promise of continuing to explore counter-hate speech programming as it relates to youth vulnerability, in the context of prevention of violence or tensions driven by political ethnic or communal fault lines, and as it relates to gender-based hate speech.

The positive track record of being able to monitor, track, and develop countermeasures to online hate speech may also offer learning for other peacebuilding, suggesting the need for broader reflection on "peacebuilding in a digital era". Many of the most innovative and impactful approaches – including partnering with social media companies, use of online tracking tools, and engagement with youth influencers – could also be used in other peacebuilding programming more generally. The use of online data and trend tracking seems especially important in providing early warning signals around escalation and risks, which could be helpful if incorporated into other peacebuilding interventions.

The positive track record of counter-hate-speech programming offers points for broader reflection on "peacebuilding in a digital era."

While promising overall, the results suggest that counter-hate speech programming could be even more impactful with a stronger human rights focus. On the larger inquiry of this Thematic Review – the integration of human rights and peacebuilding – the counter-hate speech projects tilted strongly towards the conflict prevention side of the spectrum.[83] Even in projects that demonstrated attention to rights dimensions, those involved tended to describe them as primarily conflict prevention projects.[84] The strong conflict prevention tilt can, in some cases, lead to the neglect of human rights objectives and considerations. This may leave some human rights risks unaddressed within the project design, or simply limit the degree to which these projects realize their full potential.

Some of the key findings for further innovation, growth, and learning in this field include:

- **Reinforce attention to root causes of hate speech**

Hate speech does not exist in a vacuum; it is often a symptom of deeper-rooted issues within a society, including challenges in accessing and exercising individual or collective rights. Inattention to those underlying rights dimensions may inhibit impact. For example, election-related projects tended to focus on the immediate concerns about violence in the election cycle, rather than on the long chain of rights restrictions and grievances that led to spikes in hate speech at electoral moments.[85] They might contribute to some quick wins in the immediate election cycles, but practitioners often argued that it would be more valuable to be able to prevent these issues from re-surfacing in future elections by addressing the underlying root causes.

Additional examples of the deficits of this more short-term focus manifested in the youth-related counter-hate speech projects. While all the projects in this category recognized that there is a link between young people's exclusion from political processes, hate speech, and conflict, the project activities tended to focus on short-term interventions (for example, monitoring hate speech online, awareness-raising, or positive messaging campaigns) rather than addressing the underlying grievances that are causing youth to spread hate speech. Taking a human rights-centred approach and focusing on the root causes of youth vulnerability and disenfranchisement might offer more opportunities for impact.

One positive example of practice was the counter-hate speech project examined in Sri Lanka (PBF/IRF-427). The project document observed that **a shortfall of past programming was that it attempted to address rising hate speech among youth solely through short-term means like supporting youth engagement in generating counter-narratives.**[86] It was therefore proposed that such strategies be combined with a greater focus on addressing the root causes of divisions and building resilience among stakeholders.[87]

This could be a model for other work. The ideal would be to combine existing approaches focusing on countering polarizing rhetoric in the public space, with longer-term programmes to address the underlying issues of exclusion and grievance – in essence, fusing some of the approaches seen in existing counter-hate speech work with some of the other strategies and theories of change seen in other projects that attempt to address root causes.

This observation is not limited to the projects examined. A general observation of those working in the field was that too often counter-hate speech projects tend to emphasize immediate conflict and violence prevention aspects, and neglect the underlying root causes.[88] The United Nations Plan of Action calls for the UN system to address the root causes and drivers of hate speech, so what is being identified here is not a failing in policy but a need to reinforce this overall policy in practice.[89]

Some also suggested that exploring the intersection between work countering hate speech and work related to expanding or protecting civic space might offer further avenues for identifying and addressing root causes.

- **Ensure adequate human rights safeguards in technological tools**

Future programming should more systematically build human rights standards and safeguards into counter-hate speech programming, particularly when using or developing technological tools. While this field shows tremendous promise, there is still a significant risk that removal of what is perceived as harmful content could inadvertently restrict freedom of expression. This is a general issue in the field, and one that merits greater attention from those continuing to invest in counter-hate speech projects (including but not limited to the PBF). Particular concerns were raised about emerging (and sometimes untested) AI-based detection tools, bespoke tools that may not be fully vetted, and tools developed by private companies with different definitions and standards than the UN system.[90]

There are a number of UN guidelines and system standards designed to ensure human rights safeguards. For example, the UN Strategy and Plan of Action on Hate Speech, the International Covenant on Civil and Political Rights, and the UN Rabat Plan of Action provide guidelines for delineating between lawful and unlawful expression.[91] However, implementing partners trying to develop a rapid response to an emerging situation may not always make the right determination, or may develop bespoke platforms that respond to certain conflict dynamics, but pay insufficient attention to other rights risks. Interviewees also observed that despite these standards, in practice, it can be challenging to determine when lawful expression crosses a line into unlawful expression.[92]

While some implementers sought out human rights guidance, this still appeared to be an ad hoc practice, depending on time, and availability of resources.[93] Siloing also plays a role: because many of the implementing partners viewed these projects as primarily focused on conflict prevention rather than human rights, they may not have considered building in consultation with human rights experts. **The level of scrutiny applied to technological tools remains uneven. Experts worried that not enough is being done to guard against human rights risks that might stem from using these technologies.** Overall, the diversity of tools, many of which were developed by private companies, and each screening or monitoring speech according to different definitions, creates a somewhat chaotic environment in terms of application of human rights standards.[94]

When designing future counter-hate speech programming, implementing partners may want to consider more systematic processes for evaluating the risks that technological tools pose to the right to freedom of expression or other human rights. PBSO can contribute to this by prompting implementing partners to consider human rights safeguards when it receives proposals related to counter-hate speech tools. It may also be worthwhile to encourage discussion of human rights risks, and appropriate risk mitigation and technical safeguards, in the risk mitigation section of the ProDoc, and in any follow-on monitoring and reporting.

Having the ability to reach back for expertise on the appropriate standards to apply, or other human rights risk considerations will also be crucial. While a number of actors within the UN system are available to provide such advice (for example, experts within OHCHR), there is not sufficient capacity for this limited number of experts to provide support on all potential peacebuilding projects. Implementing agencies interested in investing more in counter-hate speech programming may also wish to nurture greater in-house human rights expertise. This might encourage more regular consideration of human rights risks within counter-hate speech programming.

Last, any of the UN entities involved in communities of practice related to hate speech could facilitate mainstreaming of tools with appropriate human rights safeguards by identifying those that have proven strong at balancing human rights considerations in various contexts. For its part, PBSO might also help feed learning on human rights-centred tools and approaches into its own community of practice forums, or in others that it participates in, for example, within the UN Working Group on Hate Speech.

- **Expand counter-hate speech and disinformation programming to respond to growing demand or fill gaps**

The findings suggest that PBF would be well suited to supporting greater investments in countering hate speech and disinformation, given its existing lines of work and its willingness to explore innovative approaches, which are key in this emerging field.

As suggested by this sample, PBF-supported projects have already been incorporating counter-hate speech programming during electoral contexts, but an even greater focus on this may be needed. **Hate speech tends to spike quickly during elections and can quickly generate**

volatility. Yet attention to hate speech (much less disinformation and misinformation) is far from a universal practice within electoral assistance work. For example, there are three other projects in the overall sample of 92 in this Thematic Review that aim to support elections and prevent violence but do not place a considerable focus on monitoring and countering hate speech.[95] This may be a valuable space for further investment, either by PBF or other actors. Any future support would need to navigate the limitations on UN electoral assistance, which is only provided on the basis of a request by the host government, and due consideration for the particular sensitivities within each electoral context.[96]

An additional area for expanded attention would be that of gender-based hate speech. **The most commonly cited issue raised by experts in this field was a need for greater investment in countering gender-based hate speech.** In electoral contexts, hate speech is often directed towards women (either voters or candidates). Other projects in this sample suggested that hate speech can be seen as part of a continuum of GBV and marginalization, but also that counter-hate speech programming can be an important part of the toolkit for women's empowerment, and expanding and protecting the (virtual) civic space that allows women to realize their rights. PBF's existing strong portfolio on GEWE, and special funding mechanisms for work on this, like the GPI, could allow it to be a powerful innovator in this space. In addition, while PBSO cannot choose which proposals are brought to it, it might be able to encourage specific funding proposals in this area within the GYPI or encourage consideration of gender dimensions in any counter-hate speech programming that is proposed.

- **Nurture greater attention to intersectionality and particular vulnerabilities or susceptibilities to hate speech and its effects**

Across existing counter-hate speech work, experts identified inattention to intersectionality as an issue, and argued for more tailored programming – according to the different needs and capacities of the target group based on their intersecting identities.[97] One interviewee observed that, **"intersectionality is oftentimes overlooked or simply ignored," which can be problematic because it can lead to root causes being neglected or the full effects of "compounded discrimination" missed.**[98]

This tendency was also evident in the projects examined, particularly in those related to hate speech and youth vulnerability. For example, in the project in Sri Lanka, the project analysis identified that it was predominantly male youth who were responsible for spreading hate speech, but the project activities and strategy were not then tailored to the particular vulnerabilities or needs of the male youth in question. A lack of sufficient tailoring in youth programming was also raised in the independent evaluation of Myanmar (PBF/IRF-338) project – in that case, inattention to specific components relevant to and likely to encourage participation of female youth.[99]

- **Need for greater emphasis on sustainability and interconnectivity issues in counter-hate speech programming**

The counter-hate speech projects showed promise, but too often the benefits began and ended with the project cycle. Projects that develop bespoke technological tools invest significant resources in detection and dialogue platforms that have very little chance of outliving the project. In Myanmar, for example, despite talks to incorporate a Natural Language Processing tool into OHCHR's early warning system in the country, the tool has not yet been taken up by UN actors; the evaluation states that the "system had not yet reached a stage where integration was viable".[100] Another project in The Gambia developed its own specific website, FactCheck Gambia, which was led by a journalism institute in the country. While the institute was able to cover some operational costs, the current administrator noted that additional funding will be required to help the website continue to operate.[101] **Future programming proposals that seek to develop new tools should consider how programming will be sustained after the project lifecycle.**

Several experts suggested that the most sustainable linkage may be the degree to which projects are able to strengthen and empower national mechanisms and civil society groups (especially in situations or areas with limited UN resources).[102] This is something already seen in many of the projects but is to be further encouraged in future programming.

The counter-hate speech projects showed promise, but too often the benefits began and ended with the project cycle.

Another issue raised by practitioners and experts was that the data collected in counter-hate speech projects is often not sufficiently synchronized with other national, regional, or international systems and programming. Hate speech monitoring that contributes to early warning is only as

strong as its linkage with other actors or mechanisms that can take preventive action. Much of the information gathered during programming on countering hate speech provides important early warnings of conflict, yet findings tend to stay within the project itself. While there is a need to ensure sensitivity in information-sharing, there may nonetheless be ways to explore how the data captured in such projects can contribute to conflict prevention or rapid response through other parts of the UN system. This might include higher level policy planning and prevention platforms, rapid response centres like the United Nations Operations and Crisis Centre,[103] or other bespoke regional or national crisis or early warning platforms.

- **Improvements in cross-pillar, regional, and inter-agency platforms are needed to fully realize the early warning benefits of counter-hate speech programming**

While some of the sustainability issues might be better addressed at a project level, part of the reason that the early warning data was not fully utilized was due to larger challenges in the UN system's capacity and readiness for coordinated preventive responses. In many countries, during crisis moments (such as during elections) a range of UN actors have collaborated with Resident Coordinators and other members of the UN Country Team to set up collaborative platforms to monitor and collate early warning data.[104] In some cases, hate speech monitoring has been synchronized in with these larger early warning or crisis management platforms. However, these platforms remain ad hoc, and those involved in them note that lack of specific resourcing for these coordination platforms can result in them not being fully operationalized or used.[105] They have sometimes failed through lack of equal follow-through by all agencies involved or lack of resources to enable further action.

Some suggested that the PBF could be a resource to help support preventive responses (as requested). However, while possible in some cases, this would depend on the assessment of needs by the Resident Coordinator in coordination with the relevant Member States, and also the pace of response required. PBF programming is relatively nimble and flexible, but is still not designed to be a quick-reaction fund.

More broadly, what this suggests is that part of addressing the connectivity issues with counter-hate speech programming may involve addressing the downstream platforms and mechanisms that might make use of such early warning data. **A key learning from this Thematic Review is that human rights monitoring and data can indeed be used as a form of early warning, but there is still a gap in how well positioned the UN system is to take up and respond to those early warning signs.**

- **Invest in greater learning and data collection for this emerging area**

Interviewees suggested that because this is an emerging field, greater investment in learning on emerging strategies would be fruitful. This could include greater use of population surveys and other qualitative tools, and tracking of longitudinal data, to enable measurement of changes in public opinion and polarization over time. Agencies or organizations developing work in this field might consider systematically building such learning components into future work.

Improvements in data tracking and analysis could also directly feed into better project design and implementation. Overall, it appeared that projects are not yet systematically capturing disaggregated data on the identities of perpetrators and victims of hate speech. For example, only two projects provide data on the sex of perpetrators (in both cases, males were identified as the primary spreaders of hate speech).[106] Without monitoring this and capturing disaggregated data, efforts to counter hate speech may not be targeted at the populations most responsible for spreading hate speech or the populations most vulnerable to it. **A more systematic practice of capturing gender-disaggregated data in this emerging field would be particularly important.**

Past PBF-funded counter-hate speech programming has generated important lessons learned, which are already being taken up outside of these projects. For example, a practitioner who worked on the Côte d'Ivoire project noted that the United Nations Office for West Africa and the Sahel (UNOWAS) relied on learning form the project to inform UNOWAS approaches and activities in advance of the October 2023 elections in Côte d'Ivoire, and that UNOWAS might also apply some of the learning to its electoral work in other parts of West Africa and the Sahel.[107] PBSO might also consider additional ways to facilitate better knowledge-sharing and transfer of best practices in this space, including in the UN Working Group on Hate Speech.

Endnotes

1 See United Nations Security Council, S/RES/2686, (2023).

2 UN Office of the Special Adviser on the Prevention of Genocide, Strategy and Plan of Action on Hate Speech, May 2019, https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf. According to several experts, the lack of consistency over the definitions of these terms has contributed to a gap in understanding of how these issues relate to one another. For further discussion, see, e.g., Philip N. Howard, Lisa-Maria Neudert, Nayana Prakash, and Steven Vosloo, "Rapid analysis, digital misinformation/ disinformation and children," UNICEF Office of Global Insight and Policy, August 2021, https://www.unicef.org/globalinsight/media/2096/file/UNICEF-Global-Insight-Digital-Mis-Disinformation-and-Children-2021.pdf, p. 8. However, international human rights law defines and prohibits incitement speech, which is the most serious form of hate speech. Language on this can be found in the aforementioned Strategy and Plan of Action.

3 This definition of hate speech is according to the UN Office of the Special Adviser on the Prevention of Genocide, Strategy and Plan of Action on Hate Speech, May 2019, https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf. The PBF defines hate speech in accordance with the Plan of Action. The definitions for disinformation and misinformation is according to the United Nations Secretary General, "Our Common Agenda Policy Brief 8: Information integrity on digital platforms," June 2023, https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf, p. 5.

4 Centre for Law and Democracy, "UN Special Rapporteur for Freedom of Expression, Submission on an Annual Thematic Report on Disinformation," March 2021, https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/disinformation/2-Civil-society-organisations/UN-SR-on-FOE-CLD-Submission-Disinformation-Mar21-final.pdf.

5 This recurred in many of the ProDocs. Some policy analyses have also offered that disinformation can function as a driver of hate speech or that they have "symbiotic" effects on each other. Kevin Deveaux, Tim Baker, Mary O'Hagan, and David Ennis, "Stepping forward: Parliaments in the fight against hate speech," *United Nations Development Programme Development Futures Series*, January 2023, https://www.undp.org/publications/dfs-stepping-forward-parliaments-fight-against-hate-speech, p. 3.

6 United Nations, "Secretary-General launches United Nations strategy and plan of action against hate speech, designating Special Adviser on Genocide Prevention as focal point," 18 June 2019, https://press.un.org/en/2019/pi2264.doc.htm.

7 For a further discussion, see UNHRC A/HRC/39/64, (2018); Amnesty International, *The Social Atrocity, Meta and the Right to Remedy for the Rohingya*, (London: Amnesty International, 2022), https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.

8 United Nations General Assembly A/77/287 (2022).

9 United Nations, "Press briefing notes on Côte d'Ivoire," 27 October 2020, https://www.ohchr.org/en/press-briefing-notes/2020/10/press-briefing-notes-cote-divoire.

10 Kevin Deveaux, Tim Baker, Mary O'Hagan and David Ennis, "Stepping forward: Parliaments in the fight against hate speech," *United Nations Development Programme Development Futures Series*, January 2023, https://www.undp.org/publications/dfs-stepping-forward-parliaments-fight-against-hate-speech.

11 United Nations General Assembly A/77/287, (2022). For further discussion, see Jeremy Waldron, *The Harm in Hate Speech*. (Cambridge: Harvard University Press, 2012), https://doi.org/10.4159/harvard.9780674065086.

12 UNHRC A/HRC/49/20, (2022). For further discussion, see: Muna Abbas, Elaf Al-Wohaibi, Jonathan Donovan, Emma Hale, Tatyana Marugg and Jonathan Sykes, "Threats: Mitigating the risk of violence from online hate speech against human rights defenders," *American Bar Association Center for Human Rights*, May 2019, https://www.americanbar.org/groups/human_rights/reports/invisiblethreats-online-hate-speech/; Richard Ashby Wilson and Molly K. Land, "Persecution of human rights defenders on social media: what to do about it, Guatemala illustrates the risks in advance of June 16 Presidential Elections," *Just Security*, 6 June 2019, https://www.justsecurity.org/64422/persecution-of-human-rights-defenders-on-social-media-what-to-do-about-it/.

13 United Nations General Assembly A/77/287, (2022). For further discussion, see, for example; Carme Colomina, Héctor Sánchez Margalef, and Richard Youngs, "Study – The impact of disinformation on democratic processes and human rights in the world," European Parliament Directorate-General for External Policies, Policy Department, April 2021, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf.

14 UN Office of the Special Adviser on the Prevention of Genocide, strategy and plan of action on hate speech, May 2019, https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf.

15 Ibid, p. 1. The UN Office on Prevention of Genocide and Responsibility to Protect supports UN Country Teams (with RCOs as the main entry point) and peace and special political missions to develop context specific action plans to counter hate speech, based on the global Strategy. The Office has also provided input to several project proposals to the PBF related to countering hate speech. The UN Working Group on Hate Speech acts as a forum for exchange of best practices and learning. The 16 UN entities in the Working Group includes DPPA, of which PBSO is a part.

16 PBSO and Department of Political and Peacebuilding Affairs (DPPA), "PBF Tip Sheet on hate speech prevention programming," June 2023, https://www.un.org/peacebuilding/sites/www.un.org.peacebuilding/files/documents/pbf_tip_sheet_on_hate_speech_final_rev_12_june_2023.pdf, p.2.

17 As of the time of publication, the websites hosting documentation of the projects for the regional Western Balkans initiative (PBF/IRF-475-476-477-478-479) were not available due to a larger system issue. The links are still provided throughout this report, in expectation that these website links will be reactivated in the future.

18 Some of the election-related projects were nominally designed to respond to both hate speech and disinformation, but in practice the focus was on hate speech. For example, the context analysis for project PBF/IRF-453 (Kenya) considered hate speech and disinformation, but the majority of activities, the theory of change, and other elements of the project were almost solely focused on hate speech. Implementing partners interviewed also identified hate speech as the focus. The project PBF/IRF-481 (Republic of Moldova) also placed some degree of emphasis on countering misinformation in its ProDoc, but these did not appear to be prominent in most of the activities, and were largely overshadowed by the focus on hate speech.

19 See ProDoc PBF/CAF/H-1 (CAR), pp. 14–15. In other projects as well, the rationale for engaging on disinformation was to improve political participation. For example, the rationale voiced in the project PBF/GMB/D-2 (The Gambia) was that addressing disinformation, misinformation and hate speech would allow young people to make informed decisions on policies – generating greater participation and inclusion in governance processes.

20 PBF/IRF-482 (Liberia), PBF/IRF-481 (Republic of Moldova) (PBF/

SLE/B-11, (Sierra Leone), and PBF/IRF-475-476-477-478-479 (Western Balkans) were all approved in 2022.

21 The projects were: PBF/CIV/D1 (Côte d'Ivoire), PBF/IRF-453 (Kenya), PBF/IRF-482 (Liberia), PBF/IRF-367 (Myanmar), and PBF/SLE/B-11 (Sierra Leone). In addition, informants noted that the project in PBF/GMB/D-2 (The Gambia) was designed, in part, to respond to the concerning trend of the increase in hate speech given that the next election was to be held in 2021 (during the lifecycle of this project).

22 See, e.g.; Maina Kiai "Speech, power and violence: Hate speech and the political crisis in Kenya," *National Holocaust Memorial Museum, Washington, D.C.*, 6 August 2010, https://www.ushmm.org/genocide-prevention/blog/kenya-votes-yes; "Report from OHCHR Fact-finding Mission to Kenya, 6–28 February 2008," *OHCHR*, 28 February 2008, https://reliefweb.int/report/kenya/report-ohchr-fact-finding-mission-kenya-06-28-feb-2008.

23 ProDoc PBF/IRF-453 (Kenya), p. 8.

24 See, e.g., Maina Kiai "Speech, power and violence: Hate speech and the political crisis in Kenya," *National Holocaust Memorial Museum, Washington, D.C.*, 6 August 2010, https://www.ushmm.org/genocide-prevention/blog/kenya-votes-yes; ibid.

25 As noted in the ProDoc, the Plan of Action was developed with the support of the Office of the Special Adviser for the Prevention of Genocide and takes into account relevant guidance on gender-based hate speech. ProDoc PBF/IRF-453 (Kenya), p 9.

26 For example, the ProDoc for the Kenya project highlighted "hate speech and incitement against women candidates, voters and journalists based on their gender including online and offline attacks, trolling and harassment" as a particular issue. Examples of "online gender-based violence", included "doxing, trolling, cyberstalking, instigation to violence, blackmail, trolling, hate speech, humiliation, discrimination, defamation, identity theft and hacking, and sexual objectification". ProDoc PBF/IRF-453 (Kenya), pp. 8–9.

27 Interview with UN implementing agency, MS Teams, 14 February 2023 (Interview #21); interview with UN implementing agency, MS Teams, 28 February 2023 (Interview #24); interview with UN implementing agency, MS Teams, 2 March 2023 (Interview #25); interview with UN implementing agency, MS Teams, 2 March 2023 (Interview #27); interview with UN official, MS Teams, 22 May 2023 (Interview #55).

28 Final Evaluation "Enhancing early warning & prevention to counter hate speech and incitement ahead of the 2022 elections in Kenya," (PBF/IRF-453), pp. 11, 25, [hereinafter Evaluation PBF/IRF-453 (Kenya)].

29 Ibid, p. 26.

30 Final Evaluation Côte d'Ivoire (PBF/CIV/D1), p. 21, (translated by author).

31 Ibid, p. 40.

32 ProDoc PBF/CIV/D1 (Côte d'Ivoire), p. 10; ProDoc PBF/IRF-453 (Kenya), p. 8; ProDoc PBF/IRF-482 (Liberia), p. 7; ProDoc PBF/IRF-367 (Myanmar), p. 8; ProDoc PBF/SLE/B-11 (Sierra Leone), p. 7.

33 See, e.g.; ProDoc PBF/GMB/D-2 (The Gambia), p. 6; ProDoc PBF/IRF-338 (Myanmar), p. 5; ProDoc PBF/IRF-475-476-477-478-479 (Western Balkans) p. 8; ProDoc PBF/CIV/D1 (Côte d'Ivoire), pp. 10–11.

34 ProDoc PBF/CAF/H-1 (CAR), p. 11 (translated by author).

35 See, e.g., ProDoc PBF/IRF-453 (Kenya); ProDoc PBF/SLE/B-11 (Sierra Leone); ProDoc PBF/IRF-427 (Sri Lanka); ProDoc PBF/IRF-475-476-477-478-479 (Western Balkans).

36 ProDoc PBF/IRF-475-476-477-478-479 (Western Balkans), pp. 7–8.

37 ProDoc PBF/GMB/D-2 (The Gambia), p. 6. The reform process itself is also seen by some as non-transparent and exclusionary toward young people. A similar issue of youth more vulnerable to ethnically or group-motivated hate speech was also observed in the ProDoc PBF/CIV/D1 (Côte d'Ivoire).

38 ProDoc PBF/IRF-338 (Myanmar), pp. 5–6.

39 Ibid, p. 5.

40 Evaluation for PBF/IRF-338 (Myanmar), p.3.

41 In addition, the ProDoc for Côte d'Ivoire references the learnings from a previous PBF-supported project to posit that young people can contribute to the prevention of hate speech. ProDoc PBF/CIV/D1 (Côte d'Ivoire), p. 13 (citing learning from the previous PBF-supported project in Côte d'Ivoire: PBF/CIV/A-4).

42 Evaluations for PBF/CIV/D1 (Côte d'Ivoire), p. 18; evaluation for PBF/IRF-338 (Myanmar), p. 23; evaluation for PBF/GMB/D-2 (The Gambia), p. 20.

43 Evaluation for PBF/GMB/D-2 (The Gambia), p. 20. The evaluation of the project in the Côte d'Ivoire found that the project had "contributed to a more effective role for young people, to a reduction in the level of conflict, to an improvement in social cohesion" and that youth in the area were "demanding" ways to continue the initiatives. Evaluation for PBF/CIV/D1 (Côte d'Ivoire), p. 6.

44 Evaluation for PBF/CIV/D1 (Côte d'Ivoire), p. 6.

45 Evaluation for PBF/IRF-338 (Myanmar), p. 23.

46 Ibid, p.23.

47 Evaluation for PBF/CIV/D1 (Côte d'Ivoire), p.7; Evaluation for PBF/GMB/D-2 (The Gambia), p.20; for PBF/IRF-338 (Myanmar), p.3.

48 For a further discussion, see, for example, "Preventing hate speech, incitement, and discrimination, lessons on promoting tolerance and respect for diversity in the Asia Pacific," *Global Action Against Mass Atrocity Crimes*, August 2021, https://gaamac.org/wp-content/uploads/2022/07/APSG-REPORT_FINAL.pdf.

49 ProDoc PBF/IRF-453 (Kenya), pp. 6–9; ProDoc PBF/IRF-367 (Myanmar), pp. 7–10.

50 ProDoc PBF/GMB/D-2 (The Gambia), p. 6.

51 ProDoc PBF/IRF-481 (Republic of Moldova), p. 10.

52 For examples of programming supporting social cohesion, see PBF/IRF-481 (Republic of Moldova), PBF/IRF-427 (Sri Lanka), and the PBF/IRF-475-476-477-478-479 (Western Balkans).

53 ProDoc PBF/IRF-427 (Sri Lanka), p. 6.

54 Ibid, pp. 2, 6. Also noted was the proliferation of and greater reliance on digital technologies in this period.

55 Ibid, pp. 2, 42–48.

56 See, e.g., interview with expert on hate speech and countering violent extremism programming, by MS Teams, 6 April 2023 (Interview #36).

57 ProDoc PBF/IRF-427 (Sri Lanka), p. 13, (noting findings from a literature review of 60 programmes).

58 "Gender-based violence women and girls at risk", *UN Women*, last accessed 30 November 2023, https://www.unwomen.org/en/hq-complex-page/covid-19-rebuilding-for-resilience/gender-based-violence.

59 Interview with UN official, MS Teams, 7 February 2023 (Interview #18); "Take five: Why we should take online violence against women and girls seriously during and beyond COVID-19," *UN Women*, 21 July 2021, https://www.unwomen.org/en/news/stories/2020/7/take-five-cecilia-mwende-maundu-online-violence.

60 "Eliminating online hate speech to secure women's political participation," *UN Women*, April 2021, https://asiapacific.unwomen.org/sites/default/files/Field%20Office%20ESEAsia/Docs/Publications/2021/04/ap-WPP_online-hate-speech_brief.pdf. For further reading, see; "Women, Peace and Cybersecurity," *UN Women – Asia and the Pacific*, last accessed 30 November 2023, https://asiapacific.unwomen.org/en/what-we-do/peace-and-security/cybersecurity.

61 Julie Posetti, Nabeelah Shabbir, Diana Maynard, Kalina Bontcheva, and Nermine Aboulez, "The Chilling: Global trends in online violence against women journalists," *UNESCO*, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000377223.

62 Also known as the Napoli Women in Communities Plan of Action, see; "Initiative to enhance crucial role of women in countering hate speech launched," United Nations, 12 June 2023, https://news.un.org/en/story/2023/06/1137587.

63 Interview with UN official, MS Teams, 3 May 2023 (Interview #41).
64 ProDoc PBF/IRF-427 (Sri Lanka), pp. 2, 12.
65 ProDoc PBF/IRF-482 (Liberia), p. 19.
66 PBF/IRF-307 (Guatemala); PBF/IRF-475-476-477-478-479 (Western Balkans).
67 ProDoc PBF/IRF-475-476-477-478-479 (Western Balkans), p. 9.
68 Evaluation for PBF/IRF-307 (Guatemala), p.48.
69 Interview with four UN policy officials, by MS Teams, 7 February 2023 (Interview #18).
70 See: ProDoc PBF/IRF-427 (Sri Lanka), p. 7; PBF/IRF-475-476-477-478-479 (Western Balkans), p. 30; PBF/IRF-481 (Republic of Moldova), p. 52.
71 Some experts highlighted that work on LGBTQI+ issues can be very sensitive in some countries and may put project partners or beneficiaries at risk. Thus, while some encourage more work in this area, "Do No Harm" considerations and risk evaluations are extremely important when considering new initiatives in this area. Interview with OHCHR experts, MS Teams, 27 October 2023 (Interview #113).
72 UNDP, "iVerify, Supporting actors around the world for the prevention and mitigation of disinformation, misinformation and hate speech," last accessed on 8 May 2024, https://www.undp.org/digital/iverify.
73 U-Report can gather feedback through polls, offer advice and services through live chats, help young people find information, and mobilize youth to take action. UNICEF, "U-Report, A mobile empowerment programme that connects young people all over the world to information that will change their lives and influence decisions," last accessed on 8 May 2024, https://www.unicef.org/innovation/U-Report.
74 ProDocs noting engagement with social media companies include: PBF/CIV/D1 (Côte d'Ivoire), PBF/IRF-481 (Republic of Moldova), PBF/IRF-367 (Myanmar), PBF/IRF-338 (Myanmar), PBF/SLE/B-11 (Sierra Leone), PBF/IRF-427 (Sri Lanka), and PBF/IRF-475-476-477-478-479 (Western Balkans). Interviews indicate that this happened with the project PBF/IRF-453 (Kenya) as well. At the global policy level, there are several notable collaborations between social media companies and the UN. For example, OHCHR has partnered with Meta to translate the Rabat Plan of Action into 30 languages to assist content moderators. Interview with UN official, MS Teams, 26 May 2023 (Interview #56). The PBF tip sheet on hate speech lists additional technological tools, including those that did not appear in the project examples, such as DPPA's Sparrow tool. See PBSO and DPPA, "PBF Tip Sheet on Hate Speech Prevention Programming," June 2023, https://www.un.org/peacebuilding/sites/www.un.org.peacebuilding/files/documents/pbf_tip_sheet_on_hate_speech_final_rev_12_june_2023.pdf. See also https://mysparrowreport.org/. Broader reflections and policy guidance on engaging with social media companies are available in: The UN Office on Genocide Prevention et al., "Countering and Addressing Online Hate Speech: A Guide for policy makers and practitioners," June 2023, https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf.
75 The ProDoc for PBF/IRF-338 (Myanmar) indicates that it will pull data through CrowdTangle, p. 28.
76 Meta CrowdTangle, "CrowdTangle About Us," https://help.crowdtangle.com/en/articles/4201940-about-us.
77 Meta, "Bringing local context to our global standards," last accessed 30 November 2023, https://transparency.fb.com/policies/improving/bringing-local-context.
78 YouTube, "About the YouTube Trusted Flaggers program," last accessed 30 November 2023, https://support.google.com/youtube/answer/7554338?hl=en.
79 FactCheck Gambia, "Methodology/How we work," last accessed 30 November 2023, https://factcheckgambia.org/methodology-how-we-work/.
80 ProDoc PBF/IRF-338 (Myanmar), p. 10. This was in collaboration with the organization Koe Koe Tech.
81 U-Report, "U-Reporters – Ivory Coast," 31 July 2023, https://cotedivoire.ureport.in/.
82 Final Evaluation for PBF/IRF-453 (Kenya), pp. 11, 25; Evaluation for PBF/CIV/D1 (Côte d'Ivoire), p. 40.
83 This is not to suggest that human rights considerations were ignored. For example, in PBF/IRF-453 (Kenya), which was overall focused on preventing electoral violence, there were also steps taken to ensure that detection of hate speech is incorporated into early warning in a way that protects human rights. For example, although those involved identified the project in Sri Lanka (PBF/IRF-427) as primarily focused on conflict prevention, the project deploys important human right safeguards (e.g. a human rights risk matrix) in the components focused on monitoring online content. ProDoc PBF/IRF-427 (Sri Lanka), p. 12; interview with UN official, MS Teams, 28 February 2023 (Interview #24).
84 Interview with UN official, MS Teams, 2 March 2023 (Interview #25). For example, those involved in the project in PBF/SLE/B-11 (Sierra Leone) emphasized that it is not a "pure human rights project" because the focus is on the prevention of violence, rather than the promotion of individual human rights. Interview with UN implementing agency, MS Teams, 15 March 2023 (Interview #29).
85 One expert in the field observed that too often the focus on countering violence during the election cycle obscures attention to the underlying root causes that are driving the spike of hate speech and violence in that electoral period. Interview with UN official, MS Teams, 8 June 2023 (Interview #59).
86 ProDoc PBF/IRF-427 (Sri Lanka), p. 13 (noting findings from a literature review of 60 programmes).
87 Ibid. pp. 8, 12–13.
88 Interview with OHCHR experts, MS Teams, 27 October 2023 (Interview #113); interview with expert on hate speech, by MS teams, 26 May 2023 (Interview #56).
89 UNHRC A/HRC/22/17/Add.4, (2013), p. 3.
90 Interview with OHCHR official, MS Teams, 26 May 2023 (Interview #56). For example, AI-based detection tools were featured or planned in the projects in Kenya and the Western Balkans.
91 The Rabat Plan of Action suggests a high threshold for defining restrictions on freedom of expression, incitement to hatred, and for the application of article 20 of the International Covenant on Civil and Political Rights. UN Office of the Special Adviser on the Prevention of Genocide, "Strategy and Plan of Action on Hate Speech," May 2019, https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf; United Nations General Assembly A/RES/2200A(XXI), (1966); UNHRC A/HRC/22/17/Add.4, (2013).
92 For further discussion of the complexities of making this determination, see: UNHRC A/HRC/47/25 (2015).
93 For example, the independent evaluation of one project in Myanmar noted collaboration with OHCHR to develop the technological tool used. Evaluation for PBF/IRF-338 (Myanmar), p. 18. See, also: ProDoc PBF/IRF-427 (Sri Lanka), p. 12. UNDP has developed a guidance note on risk-informed use of online data for preventing violent extremism and hate speech, including considerations for carrying out a risk assessment, ensuring due diligence of partnerships and utilizing online and AI tools. See; UNDP, "From Pilots Toward Policies: Utilizing Online Data for Preventing Violent Extremism and Addressing Hate Speech," 13 May 2022, https://www.undp.org/publications/pilots-toward-policies-utilizing-online-data-preventing-violent-extremism-and-addressing-hate-speech.
94 Three projects explicitly note that they are utilizing the Rabat Plan of Action as a framework for monitoring and countering hate speech; ProDoc PBF/IRF-427 (Sri Lanka), p. 9; ProDoc PBF/IRF-453 (Kenya), pp. 16, 21; ProDoc PBF/IRF-367 (Myanmar), pp. 11, 25); however,

experts interviewed noted that the standards in the Rabat Plan of Action would also not be the appropriate framework for all situations or programmatic uses.

95 See PBF/IRF-366 (Bolivia) and PBF/MLI/A-5 (Mali). PBF/CIV/C-2 (Côte d'Ivoire) also deals with elections, though the focus is on supporting victims of post-election violence.

96 There are a number of steps involved in considering UN electoral assistance, and which measures would be applicable to any UN electoral projects. A further discussion is available within: United Nations Focal Point for Electoral Assistance Matters, Principles and Types of UN Electoral Assistance, Ref. FP/01/2012, 3 March 2021, ¶¶2, 9.

97 Currently, only a few projects examine how an individual's race, ethnicity, religion, gender identity, gender expression, sex, age, sexual orientation, disability, economic status, vocation, education, or other identity markers impact an individual's likelihood of being a target or perpetrator of hate speech in a given context. Interview with GNWP expert, MS Teams, 15 May 2023 (Interview #52).

98 Interview with UN official, MS Teams, 26 May 2023 (Interview #56).

99 ProDoc PBF/IRF-338 (Myanmar), p. 29.

100 Project evaluation PBF/IRF-338 (Myanmar), p. 18.

101 Project evaluation PBF/GMB/D-2 (The Gambia), pp. 6, 8.

102 Interview with OHCHR experts, MS Teams, 27 October 2023 (Interview #113).

103 The UNOCC provides rapid situational awareness to UN decision makers and can be seen as the UN's crisis hub. Some also suggested potential use for such data in the Regional Monthly Reviews (see further discussion in *infra* note 410).

104 Informants noted a well-performing regional hate speech prevention/freedom of expression dashboard led by UNDP and supported by OHCHR in Bangkok, which also informed other UN actors about the region. Interview with OHCHR experts, MS Teams, 27 October 2023 (Interview #113).

105 Interview with four practitioners working on early warning and emergency response, MS Teams, 8 November 2023 (interview #153).

106 The two projects are PBF/IRF-481 (Republic of Moldova) and PBF/IRF-427 (Sri Lanka). The ProDoc for Republic of Moldova states that those who are spreading hate speech are primarily male (78 per cent) (p. 10). The ProDoc for Sri Lanka states that "young men appear to play a significant role in the spreading of hate speech with as much as 90% of hate speech circulating online stemming from users identifying as male and a majority being in 15–30 age demographics" (p. 7).

107 At the time there was a plan for UNOWAS to provide funding for hate speech-specific programming in the months leading up to elections in all areas under its mandate; however, as of the time of writing, it was not clear that this had been provided. Interview with UN implementing agency, MS Teams, 2 March 2023 (Interview #27).