FACEBOOK RESPONSE

Sri Lanka Human Rights Impact Assessment

Overview

Beginning in late February 2018, Sri Lanka faced a wave of religiously-motivated communal violence. Triggered by a violent cross-communal road rage incident, anti-Muslim riots broke out in the eastern town of Ampara. Unrest soon spread to the central Kandy District.

In the first week of March 2018, mobs drawn predominantly from the Sinhala-Buddhist ethnic group began attacking Muslims and burning mosques, Muslim-owned shops, and homes. The government declared a state of emergency and deployed armed forces to assist local police. The riots resulted in two deaths and ten injuries, and caused extensive cultural and property damage.

This was not the first significant outbreak of communal violence in Sri Lanka. A similar incident played out in 2014, five years after the end of the country's civil war. But crucial reporting by media and civil society organisations, as well as valuable feedback we received from Sri Lankan authorities, suggested that, this time, content shared on social media platforms—and Facebook and WhatsApp in particular—may have helped stoke the violence.

Initial violence in Ampara, for instance, was reportedly linked to a viral video falsely claiming a Muslim restaurant owner admitting he had targeted Sinhala men by mixing "sterilization pills" into restaurant food. The rumor led others to share a flood of hateful, inflammatory content.

At the time, Facebook lacked significant country-specific staff or product interventions. Following the violence in the spring of 2018, and recognizing important concerns about the role of our products, we began a major effort to invest in understanding and responding to the positive and negative impacts of Facebook in Sri Lanka. Our goal was to help ensure Sri Lankans could continue to express themselves freely and safely on Facebook.

Indeed, our work on Sri Lanka marked the formal start of a **company-wide effort dedicated to systematically understanding the intersection of our products and offline conflict globally**, and to building product, policy, and people-based mitigations for countries experiencing or at risk of conflict.

By the end of 2018, Facebook had made substantial changes: among other efforts, we hired dozens more Sinhala and Tamil-speaking content moderators to review and remove content that violates our Community Standards; we hired dedicated Sri Lankan policy and program managers to work full-time with local stakeholders; and we deployed proactive hate speech detection technology in Sinhala to help us more quickly and effectively identify potentially violating content. We have continued to build on this work and deepen our investment to help ensure our products contribute positively to Sri Lanka.

About This HRIA

We commissioned an independent human rights impact assessment (HRIA) of Facebook's role in Sri Lanka to identify and mitigate human rights risks, and to inform our future strategy.

The HRIA was conducted by **Article One**, a specialized human rights and ethics consulting firm. Article One's methodology was based on the **UN Guiding Principles on Business and Human Rights** (UN Guiding Principles), and included interviews with 29 organizations, as well as observation of Facebook-led focus groups and individual interviews with 150 Sri Lankans who frequently and actively used Facebook.

In-country research took place in August and September 2018. Subsequent unrest, as well as the efforts required to mitigate the adverse impacts identified, delayed the HRIA's disclosure.

While much has since changed, both for Facebook and Sri Lanka, we are sharing the HRIA's executive summary now, along with this response, to help local and international stakeholders better contextualize Facebook's impacts and investments in Sri Lanka. It is also part of Facebook's broader commitment to meaningful transparency about our human rights due diligence, and our product integrity work.

What's an HRIA?

A human rights impact assessment (HRIA) is a detailed form of human rights due diligence. An HRIA allows Facebook to identify human rights risks; know its positive and negative human rights impacts; and strengthen positive impacts as well as mitigate human rights harms.

Facebook has committed to do human rights due diligence, including HRIAs, as a member of the Global Network Initiative, which we joined in 2013. The UN Guiding Principles, applicable globally, also state businesses should conduct human rights due diligence, and HRIAs should include meaningful consultation with potentially affected groups and other relevant stakeholders (2.A.18(b)).

Why This Facebook HRIA Response?

There's no standard format for disclosure. Facebook has chosen the unusually transparent step of disclosing HRIA executive summaries, with recommendations. This response is intended as a summary of the process, the findings, and a guide to what we have done and will do to follow up.

Good human rights due diligence is not a compliance exercise. We are actively seeking to learn from this HRIA to inform our work in Sri Lanka, to mitigate risks in other conflict-affected settings, and to serve users around the world. This HRIA, along with accompanying processes, has helped change Facebook policies, operations, and products. The impact has been timely and real.

Sri Lanka HRIA Findings

The HRIA found that Facebook has had important, positive human rights impacts in Sri Lanka, as well as contributed to salient human rights risks.

The research found Facebook was:

- An important platform for Sri Lankan human rights defenders, journalists, and disadvantaged groups to express their views, seek and share information and build communities;
- A powerful tool for activism, promoting civic and political engagement and enabling activists to shine a light on uncomfortable truths;
- A safe space in which LGBTQ+ individuals could communicate and organize around LGBTQ+ rights;

- Important to and improved emergency responses to natural disasters; and
- Significant in the ways it contributed to the Sri Lankan economy, including increased opportunities for historically disadvantaged groups.

The Executive Summary notes: "[d]espite frequent public criticism of Facebook, the majority of stakeholders engaged defended the platform and argued its value to the country—if managed appropriately—was significant." (p. 24)

The HRIA also noted significant salient human rights risks. Relevant findings included:

- Misuse of the platform to spread hate speech and rumors against religious minorities, especially Muslims, that may have inflamed tensions and escalated violence to people and destruction of property. Stakeholders stated they had warned Facebook earlier of such misuse, but that the company had been largely unresponsive;
- Harassment and surveillance of human rights defenders by other users;
- Gender-based hate speech and online sexual harassment against women, including cyberviolence and non-consensual sharing of images (including intimate images);
- Child sexual exploitation related to the posting of sexually explicit photos and comments about children, as well as online grooming of children;
- Despite the fact that Facebook generally provided valuable safe space for LGBTQ+ individuals, there were adverse impacts associated with harassment and bullying campaigns against LGBTQ+ Facebook users, some of whom had been involuntarily "outed" on the platform; and
- While many of these harms were directly caused by users, not the company, the report found Facebook's platform may have contributed through: limited formal due diligence; inadequate enforcement of its Community Standards; and limited or untimely escalation channels for victims and/or advocates.

In addition, stakeholder feedback indicated lack of cultural context, lack of linguistic expertise and overreliance on user feedback had contributed to a situation in which Facebook could not appropriately assess the risks to rights holders, including risks to victims of cyberviolence.

Facebook Response to Findings and Recommendations

The Sri Lanka HRIA was initiated in early 2018, at a time when Facebook began to significantly strengthen its understanding of, and response to, global human rights and conflict-related risks. Facebook thus prioritized responding to the salient risks surfaced by the HRIA before disclosing its response.

This HRIA, and the discussions that informed it, has helped catalyze real changes to Facebook policy, operations, and products. Its recommendations have triggered change to Facebook's approach not just to Sri Lanka, but globally—particularly in its product work and decision-making related to countries at risk of conflict. We note, for example, that while many of the recommendations in the HRIA and our responses are specific to Sri Lanka, many also have implications for other contexts, and are reflected across multiple HRIAs that Facebook has commissioned. This reflects the universal nature of human rights, the global reach of our products, and the intersectionality of the impacts identified.

Specifically, we have set up a dedicated, multi-disciplinary team with subject matter experts focused on countries at risk of conflict, invested significant engineering resources to better understand and address Facebook's potential contribution to offline conflict and created new strategic response and policy positions dedicated to these issues.

A. WHAT WE'VE IMPLEMENTED

Facebook has made significant changes as a result of the Sri Lanka HRIA and related processes. These are grouped according to the seven recommendation areas in the HRIA report.

Improve Corporate-Level Accountability

We have:

- Continued to comply with our commitments to privacy and freedom of expression as a member of the Global Network Initiative (GNI). Earlier this year, GNI completed its biennial assessment of Facebook,

finding that we "strengthened its systematic review of both privacy and freedom of expression" and "is making good-faith efforts to implement the GNI Principles with improvement over time.";

- Formalized an approach to help us determine which countries require high priority conflict prevention interventions, such as product changes, UX research, or other risk mitigation steps;
- Incorporated human rights principles into the Community Standards Values in September 2019;
- Created and recruited a new senior role to lead company work on human rights as it relates to policies, products and partnerships, and is expanding related roles and resources;
- Built clear processes for expert human rights input into product and content policy development;
- Significantly improved mechanisms for user control over their privacy and information, with new features like Privacy Checkup, Privacy Shortcuts, Off-Facebook Activity, and others; and
- Explored opportunities for remediation, including through recognition and apology.

Evolve Community Standards

We have:

- Significantly increased the number of content reviewers, in various locations, including full time staff and scaled support, to provide 24/7 content moderation coverage in Sinhala and Tamil specifically for Sri Lanka. All content moderators and outsourced operations are audited;
- Created and hired new staff positions dedicated to Sri Lanka policy and programs, enabling significantly better consultation with local CSOs and experts on our Community Standards and other policies;
- Established and intensified a regular cadence of engagement with key stakeholders on the ground, including listening sessions, product research, and co-design workshops with CSOs and human rights defenders
- Created a new policy to remove verified misinformation and unverifiable rumors that may contribute to the risk of imminent offline physical harm. This policy is highly relevant to conflict-affected and other challenging settings;[1]

1. For a detailed overview of our global stakeholder engagement process and how it contributes to policy making, see https://www.facebook.com/communitystandards/stakeholder_engagement. Minutes of relevant policy meetings at https://about.fb.com/news/2018/11/content-standards-forum-minutes/

- Updated existing policies to protect vulnerable users, including the protection of users whose "outing" might increase risks of offline harm (e.g. involuntary outing of veiled women, LGBTQ+ individuals or human rights activists);
- Expanded our **bullying policies** to increase protections provided to all individuals, including public figures such as human rights defenders and journalists (we have, for example, updated our policies to explicitly prohibit female-gendered cursing and attacks on the basis of derogatory terms related to sexual activity). The updates we've made to our bullying and harassment policies also cover dehumanizing speech, which is referenced among Article One's recommendations on hate speech; and
- **Expanded our policies against voter interference** to prohibit misrepresentations about how to vote, and statements about whether a vote will be counted, which enabled us to remove election-related misinformation ahead of Sri Lanka's 2019 presidential election.

Invest in Changes to Platform Architecture

We have:

- Conducted an audit of harmful content in Sri Lanka, and blocked additional search terms, including those related to sexual content;
- Developed proactive hate speech detection technology in Sinhala to help us more quickly and effectively identify potentially violating content, reducing Facebook's dependence on user reporting;
- Increased friction for sharing problematic content across the platform. We have made product interventions on WhatsApp to limit the spread of disinformation—clearly labelling forwarded messages, capping forward limits for all messages to five, and lowering the limit for highly forwarded messages to just one, with the latter cutting the virality of highly forwarded messages by 70%;
- Applied friction to the sharing of misinformation by applying interstitial warnings when users attempt to share content that has been debunked by a third-party fact checker;
- Introduced new content demotions, targeting frequently reshared messages to reduce incentives on users to reshare a message beyond a certain threshold. At times we have also demoted all content of users who have a demonstrated recent pattern of

violating our Community Standards. These demotions seek to respect the guidance on permissible limits to freedom of expression under Article 19 of the ICCPR;

- Invested in new technologies to proactively detect child nudity and child exploitative content when it's uploaded. This investment enables us to enhance our reporting to law enforcement. It also means we can identify accounts seeking to engage in inappropriate interactions with children more quickly, so we can remove them and mitigate risk of harm. We have open sourced this technology to encourage use by others;
- Deployed new tools to fight non-consensual sharing of intimate images. We're using machine learning and artificial intelligence to proactively detect near nude images or videos that are shared without permission on Facebook and Instagram. This enables us to limit harms by finding this content before anyone reports it;
- Raised user awareness of mitigation approaches through on the ground research trips, as well as co-design workshops with civil society, that allowed us to validate and refine product intervention strategies; and
- Raised user awareness of mitigation approaches with the global NCII victim-support hub in our Safety Center, Not Without My Consent, available in 50 languages.

Address Challenges to the Platform-Level Grievance Mechanism

We have:

- Launched an improved reporting/blocking feature in Facebook Messenger;
- Introduced appeals of content moderation decisions in 2018, and expanded them significantly since to include almost all policy areas/abuse types, seeking alignment with the Santa Clara Principles;
- Significantly improved content moderation support and resources, with increased staffing, enhanced wellness and resiliency resources and improved tools that allow content reviewers to customize some of the ways they do their work; and
- Launched an ambitious independent operational grievance mechanism, the Facebook Oversight Board, informed by a detailed human rights review.

Ongoing Due Diligence

We have:

- Expanded and intensified human rights due diligence at the country, product, and product intervention level, with follow up due diligence underway; and
- Increased capacity to identify and mitigate human rights risks in real time, as a result of the combined effect of multiple other workstreams;

Transparency

We have:

- Enhanced transparency and disclaimer requirements for electoral ads in advance of the Sri Lankan elections of November 2019. (At the time of writing, the full set of political advertising transparency tools was due to be available prior to June 2020);
- Published the internal guidelines we use to enforce our Community Standards such that our public-facing policies include the granular details of what is and is not allowed on Facebook;
- Invited external guests to the cross-functional policy development meeting, the Product Policy Forum, at which we discuss and debate changes to our Community Standards, ads policies and major News Feed ranking changes. We also make minutes from the meeting available to the public;
- Continued to make metrics available that track how well we're doing at enforcing our Community Standards in a bi-annual Community Standards Enforcement Report. As of November 2019, the report includes numbers on how much content people appealed; and
- Developed clear criteria to inform the partnerships we develop at a local level.

Use Leverage to Address Root Cause Challenges

We have:

- Significantly expanded partnerships with local CSOs in Sri Lanka who are able to report potentially violating content to us using direct escalation channels, and often provide input into our policy development process;

- Continued digital literacy training, which has already reached 20,000 local students, implemented in partnership with a local civil society organization;
- Launched and funded a grant program in conjunction with Splice Media to support local media in Sri Lanka, as well as elsewhere in Asia; and
- Launched a Third Party Fact Checking Program for Sri Lanka, partnering with Agence France-Presse (AFP) and Fact Crescendo, to combat misinformation.

B. WHAT'S IN PROCESS

Facebook recognizes human rights due diligence is an ongoing process, and we have not yet implemented all the HRIA's recommendations.

Work in progress includes:

- Building on our human rights accountability and governance strategy, which is evolving rapidly;
- Designing the first iteration of a human rights defenders program, based on the findings of the this and other HRIAs, and specific due diligence on the needs of human rights defenders; and
- Increasing the practical tools we can use to educate users about our Community Standards in diverse local languages.

Likewise, Facebook has significantly expanded its contact with CSOs and other local stakeholders, but there is much more to be done, including in follow up to this report

C. WHAT WE HAVEN'T IMPLEMENTED

It is also important to be transparent, and note that some recommendations may not currently be technically or operationally feasible. The recommendation to develop AI to predict when online hate speech may trigger offline violence, for example, is possibly beyond the bounds of current global technical and academic capacity, although related interventions that assist us prioritize resources may deserve exploration.

Similarly, we are unlikely to create a process involving ongoing open global calls for comment on changes to our Community Standards, given the scale involved, although we do conduct extensive stakeholder engagement aspart of our policy development process. We'll continue to identify potential solutions to human rights challenges, and will implement them as they're developed.

Final Note

This human rights impact assessment is an important step forward for Facebook, combined with the simultaneous disclosure of HRIAs on Indonesia and Cambodia. By detailing the ways in which we have sought to implement Article One's recommendations—and where and why we have not—we have attempted to demonstrate accountability for our human rights impact.

Important as these HRIAs are, we see them as the start of a much bigger process of engagement with our users and those whose rights are impacted by our platform.

Acknowledgments

We are deeply grateful to the dozens of Sri Lankan and international human rights defenders, civil society organisations, and others who provided input to this assessment. You have enabled us to implement the very important mitigations described in this response, benefiting people in Sri Lanka and around the world.

Going forward, we will be intensifying our commitment to understand our human rights impacts and risks. The tech sector and social media platforms are driven by a commitment to innovation and a constant desire to do more and do it better. We're trying to ensure we apply this ethos to our human rights responsibilities.