Countering online disinformation and misinformation as security threats

Input paper for the High-Level Advisory Panel on Global Public Goods

==

What's new? Social media has increasingly served as a valuable method of organising and sharing information in conflict. But mitigating the harm caused by disinformation, hate speech and incitement to violence – always a challenge – is particularly difficult during political crises and armed conflict.

Why does it matter? Harmful content spread online has inflamed tensions across a number of countries facing major crises. This is especially dangerous in conflict zones, particularly where quick fact-checking is difficult, hate speech is highly contextual, and platforms face accusations of bias.

What should be done? In addition to better resourcing and implementing technical solutions, platforms should strengthen partnerships with one another and with regional and multilateral organisations to coordinate policies. Both online programs, like counter-messaging, and offline actions, like resourcing local media groups to fact-check on the ground, will be crucial in mitigating harm. The UN response to the COVID-19 "infodemic" provides a blueprint for a global, multi-stakeholder approach to countering disinformation.

Why is this relevant to the Panel? High-quality information is a public good, as UNESCO has argued, as it "helps to advance collective aspirations and [. . .] forms the key building block for knowledge." Disinformation and misinformation are "public bads," as they can negatively affect large numbers of people, whether in the context of conflicts, pandemics or other dangerous environments. In an era in which (i) online information is available across borders; and (ii) violence, disease and other threats cannot be contained within or by states, countering disinformation and misinformation is a ***global*** public good.

The UN had success countering false and misleading information over COVID-19, and the Panel should consider how to scale up and institutionalize efforts to replicate this success with regards to conflicts.

==

Online disinformation, hate speech and incitement can subvert peace processes, heighten tensions, and lead to real-world violence in conflict zones.[1] But the tension between protecting speech and preventing offline harm is difficult to navigate. As a result, most attempts to safeguard the digital commons have thus far been ad-hoc or issue-specific. Social media companies continue to play the most central role in combating disinformation, incitement and hate speech. Many have voluntarily agreed to certain global standards of behaviour, but platforms are responsible for how they implement these policies and evaluating their own failures.[2]

Social media companies face both technical and political challenges to mitigating harm. On the technical front, platforms are faced with rooting out disinformation, hate speech, or violence incitement, while protecting users' freedom of expression more broadly. Fact-checking at this scale is enormously difficult. Platforms must operate across a plethora of languages and dialects. Content moderation typically relies on a mix of artificial intelligence and human moderators. Leaked documents released by Facebook whistleblower Frances Haugen highlighted that Meta is severely lacking in language capacity in both facets across a number of conflict zones, including Ethiopia and Afghanistan. One leaked study estimated that Meta took action globally on as little as 3 to 5 per cent of hate speech and less than 1 per cent of violence incitement, for example.

Platforms do have a number of technical fixes to slow the spread of harmful content. They can remove organisations from the platform entirely if deemed sufficiently dangerous. For example, Facebook was able to rapidly remove Myanmar's military junta from the platform after the February 2021 coup, largely because it had built up its language capacity and moderation team in the wake of the

[1] Both misinformation and disinformation refer to false or misleading content, but disinformation involves intent to deceive.
[2] See, for example, the policies of Twitter and Facebook.

Rohingya atrocities and was able to reach a quick judgment about the risks of continuing to post its content. Adding "friction" to the sharing of material – asking users to read an article before reposting it, for example, or limiting reshares on a viral post until content moderators can review it – can reduce the spread of disinformation. Platforms can adjust their algorithms to ensure that polarizing content is less visible on users' feeds.

But often the challenges that platforms face are political. Companies decide what constitutes hate speech, which is often highly contextual. Content moderation may benefit certain parties to a conflict to the detriment of others. Unverified reporting can play an important role in providing information about on-the-ground events in conflict zones, but it can also be used to incite violence. Platforms decide when world leaders' posts are taken down, or even when their accounts are disabled. And they determine what counts as a "dangerous organization", to be de-platformed entirely.

Ideally social media companies should not make such deeply political decisions about the types of content and actors allowed on platforms alone. Facebook's removal of Myanmar's military after the coup was a relatively straightforward decision: it was actively spreading disinformation about elections, and had previously played the leading role in fabricating and spreading anti-Rohingya content. But in other cases, deciding who can and cannot use the platform, and what they can say, is significantly more complicated. The Arakan Army, an ethno-nationalist armed group in Myanmar's Rakhine state, was removed from Facebook after the platform designated it as a "terrorist organization". Some reports have suggested that the decision to classify the group as such was rushed in the wake of attacks targeting government security forces, because the company was concerned about its reputation. Government critics in Cameroon have expressed concerns that their posts are censored more than the government's. Facebook's Oversight Board requested an independent review to look into accusations of online bias against Palestinians during the 2021 Israel-Palestine conflict. In Ethiopia, platforms are asked to moderate content in a context of deep polarization in which the question of who is victim and who is aggressor is often highly context specific.

Addressing these issues is enormously complex, and will require a mixture of both on and offline capacity building. One of the important changes Facebook made before the U.S. 2020 elections was to boost authoritative sources and local news, pushing them higher in users' feeds as a means of slowing the spread of disinformation. But identifying trusted sources and fact-checking can be significantly more difficult in conflict zones. For example, Facebook's Oversight Board ruled that due to a lack of corroboration Meta should remove a post that alleged civilians were assisting an Ethiopian rebel group in committing human rights violations. Meta disagreed with the Oversight Board, however, reasoning that removing unverified rumors that could lead to violence "would impose a journalistic publishing standard on people that could prevent them from raising awareness of atrocities." Verification of nonpartisan bloggers and independent news sites, in consultation with local NGOs and civil society, can help empower trusted sources, as Crisis Group recommended in Cameroon.

Given these challenges, regional actors have played a crucial role in countering harmful content online. Platform partnerships with outlets providing fact-checking both regionally and globally – like Rappler and the VERA Files in the Philippines, Fact Crescendo in Sri Lanka, Africa Check, and AFP Fact Check – have improved their ability to respond appropriately to disinformation. These efforts should be further funded and expanded. Meta collaborates with partner organizations to monitor potentially harmful trends, and has provided digital literacy training to reduce the spread of disinformation and encourage users to flag harmful content when they see it. Regional and local actors are also in many cases ideally suited to preventing or limiting the impact of disinformation. For example, the UN Support Mission in Libya (UNSMIL) recently established a set of principles governing social media use around the peace process in consultation with journalists, influencers, and civil society actors. The UN Multidimensional Integrated Stabilization Mission in the Central African Republic (MINUSCA) countered disinformation targeted at its staff through its own social media platforms, mass texts, press releases and radio messages.

More broadly, however, strengthening local and unbiased media offline is crucial to reducing the spread of harmful content online and ensuring the prevalence of high-quality information requires trusted news sources. Rumors are particularly hard to verify in conflict zones, where fact-checking organizations typically do not have a strong on-the-ground presence. In Ethiopia, disinformation and violence incitement often originates in partisan news sources before getting shared on social media. The ability of platforms to fact-check on-the-ground reports and boost trusted sources – and, correspondingly, 'demote' more biased sources –depends on supporting reliable local journalists and civil society actors who provide such reporting.

The case for international cooperation and the role of the UN

Cooperation – across platforms, with local actors, and with regional and international bodies – will be crucial. One major success is the "hash-sharing database" spearheaded by the Global Internet Forum to Counter Terrorism, which facilitates the automatic removal of certain known pieces of terrorist propaganda across different sites. Collaborating across companies and with multilateral and regional groups can strengthen content moderation. On the technical front, knowledge sharing can refine platform responses: for example, a Facebook employee warned that the "seriously scant" list of slurs in languages used in Afghanistan hindered hate speech detection. Politically, collaboration can provide more structured input into decisions, for example on what groups to include on banned organizations lists. This aligns with United Nations Strategy and Plan of Action on Hate Speech, which outlined ways in which the Secretariat could support Resident Coordinators in countering harmful content.

The UN response to COVID-19 "infodemic" provides a blueprint for a global, multi-stakeholder approach to countering disinformation. Partnerships with social media platforms – including Facebook, WhatsApp, and Viber – allowed direct communication with the public. WHO-established networks of technical and social media experts, as well as regional "Information Centres", facilitated rapid response to disinformation across a variety of languages. Innovation labs like UN Global Pulse harnessed AI to monitor and respond to misinformation. The Verified initiative "flooded" the online space with accurate information, through partnerships with civil society, social media influencers, and private companies (studies suggest several of these programs reduced the sharing of misinformation). Many of these programs can be directly mapped to crisis zones: cross-sector partnerships, collaboration with local civil society and influencers, improved monitoring of online content across regions, and proactive efforts to offer factual messaging can mitigate the impacts of misinformation.

Options for the Panel

The Panel could (i) highlight these opportunities for a multi-stakeholder approach to governing the protection of high-quality information online as a security concern; (ii) point to best practices on the part of the UN, regional players, the private sector and other actors; and (iii) make proposals for scaling up international efforts to counter disinformation and misinformation relating to conflicts, diseases and other threats.