On February 4, 2022, Meta referred a case to the Board concerning content posted on Facebook in November 2021, during ongoing armed conflict between the Tigray People’s Liberation Front (TPLF) and the Federal Democratic Republic of Ethiopia.

The content was posted in Amharic by the Tigray Communication Affairs Bureau page, which states that it is the official page of the Tigray Regional State (Ethiopia) Communication Affairs Bureau. The post discusses the losses suffered by the Federal National Defense Forces under the leadership of Prime Minister Abiy Ahmed. It goes on to say that the armed forces must surrender to the TPLF if they hope to save their lives, and if they refuse, they will die. The post also encourages the national army to “turn its gun” against the Prime Minister’s group in order to make amends with the people it has harmed (for more on the conflict see the Board’s prior decision in Case 2021-014-FB-UA). The page is set to public, meaning it can be viewed by any Facebook user, was verified and was previously subject to cross-check, but not at the time the content was posted and reviewed. Cross-check is a system that Meta claims helps ensure accurate enforcement through additional levels of human review (for more on cross-check see the Board’s announcement of the policy advisory opinion that is currently in-process). The page has about 260,000 followers.

The content was reported by 10 users for violating the Violence and Incitement, Dangerous Individuals and Organizations, and Hate Speech policies. In addition to the user reports, Meta’s automated systems also identified the content as potentially violating and lined it up for review. During Meta’s initial review, the company determined that the content was not violating and left it on the platform. Following another review initiated through the company’s crisis response system, Meta determined the content violated its Community Standard on Violence and Incitement and removed it.

Under its Violence and Incitement policy, Meta states that it will remove any content that “incites or facilitates serious violence.” The policy prohibits “threats that could lead to death (and other forms of high-severity violence) … targeting people or places.” The policy also states that for “coded statements” or “veiled or implicit” threats, the company will look to other signals to determine whether there is a threat of harm. These signals include whether the content was “shared in a retaliatory context” or if it “[references] historical or fictional incidents of violence,” among others.
In its referral of the case to the Board, Meta states that the decision regarding the content was difficult because it involves removing “official government speech that could be considered newsworthy,” but noted that it may pose a risk of inciting violence during an ongoing conflict. In its analysis, Meta also told the Board that the company took into account the documented atrocities committed during this conflict by all parties involved.

Following the referral of this case to the Board, the user was given the option to submit a statement to the Board. The Board has not received a statement from the user.

The Board would appreciate public comments that address:

- How Meta enforces its Violence and Incitement policy in conflict situations, including whether its actions are consistent across different conflicts.
- Whether credible threats of violence made between parties during an armed conflict should be treated differently under Meta’s policies and under what circumstances.
- When content that violates Meta’s policies should be allowed under the “newsworthiness” allowance in conflict situations, noting that that allowance was not applied in this case because, according to the company, the allowance does not apply to content that presents a risk of contributing to physical harm.
- Whether Meta should allow content that violates its Violence and Incitement Community Standard if the actions threatened, incited, or instigated are permitted under international humanitarian law (also known as the law of armed conflict).
- Whether and how Meta’s cross-check program should work during an armed conflict.
- Content moderation challenges specific to Ethiopia and languages spoken in the country, particularly during times of heightened tension or conflict.
- Evidence or analysis of statements from armed groups or the military in Ethiopia on social media that have incited or instigated violence, including any violations or abuses of international law.
- The information environment in Ethiopia during the conflict, including access to the internet and independent sources of reporting, and how this should influence Meta’s approach to moderating content from parties to the conflict.

In its decisions, the Board can issue policy recommendations to Meta. While recommendations are not binding, Meta must respond to them within 60 days. As such, the Board welcomes public comments proposing recommendations that are relevant to this case.
The Oversight Board is committed to bringing diverse perspectives from third parties into the case review process. To that end, the Oversight Board has established a public comment process.

Public comments respond to case descriptions based on the information provided to the Board by users and Facebook as part of the appeals process. These case descriptions are posted before panels begin deliberation to provide time for public comment. As such, case descriptions reflect neither the Board’s assessment of a case, nor the full array of policy issues that a panel might consider to be implicated by each case.

To protect the privacy and security of commenters, comments are only viewed by the Oversight Board and as detailed in the Operational Privacy Notice. All commenters included in this appendix gave consent to the Oversight Board to publish their comments. For commenters who did not consent to attribute their comments publicly, names have been redacted. To withdraw your comment, please email contact@osbadmin.com.

To reflect the wide range of views on cases, the Oversight Board has included all comments received except those clearly irrelevant, abusive or disrespectful of the human and fundamental rights of any person or group of persons and therefore violating the Terms for Public Comment. Inclusion of a comment in this appendix is not an endorsement by the Oversight Board of the views expressed in the comment. The Oversight Board is committed to transparency and this appendix is meant to accurately reflect the input we received.
### Regional Breakdown

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia Pacific &amp; Oceania</td>
<td>1</td>
</tr>
<tr>
<td>Central &amp; South Asia</td>
<td>0</td>
</tr>
<tr>
<td>Europe</td>
<td>3</td>
</tr>
<tr>
<td>Latin America &amp; Caribbean</td>
<td>0</td>
</tr>
<tr>
<td>Middle East and North Africa</td>
<td>0</td>
</tr>
<tr>
<td>Sub-Saharan Africa</td>
<td>1</td>
</tr>
<tr>
<td>United States &amp; Canada</td>
<td>2</td>
</tr>
</tbody>
</table>
Short summary provided by the commenter

While this case is not about Ukraine, Meta’s Ukraine policies should matter to this case. Meta’s approach to Ukrainian conflict-content, seemingly inspired by aspects of the laws of armed conflict, seems a better fit for the realities of armed conflict than its Community Standards. However, Meta’s justification for this approach – national self-defense – makes it difficult to apply to other conflicts. Relying on such a narrow justification will lead to inconsistent treatment of different armed conflicts, particularly those between state and nonstate actors. Instead of backing just causes by granting them exceptions, Meta should strive to apply standards inspired by the laws of armed conflict to the content of all combatants.

Full Comment

2022-006-FB-MR Public Comment Will Duffield, Policy Analyst, Cato Institute In this case, the Oversight Board reviews Meta’s decision to remove a post by the Tigray People’s Liberation Front, combatants in the ongoing Tigray War in Ethiopia. The post warns Federal National Defense Forces soldiers to surrender or face death, and encourages them to turn their guns on Ethiopia’s president. The board asks whether “Meta should allow content that violates its Violence and Incitement Community Standard if the actions threatened, incited, or instigated are permitted under . . . the law of armed conflict.” The board also asks “whether [Meta’s] actions are consistent across different conflicts.” Since February, Meta has made several exceptions to its Community Standards in response to Russia’s invasion of Ukraine, in some cases allowing content that would otherwise violate its Violence and Incitement policy. Meta recently requested a review of its special policies for the Russo-Ukrainian War, before withdrawing the request. Thus, while this case is not about Ukraine, Meta’s Ukraine policies should matter to this case. Meta’s approach to Ukrainian conflict-content, seemingly inspired by aspects of the laws of armed conflict, seems a better fit for the realities of armed conflict than its Community Standards.
However, Meta’s justification for this approach – national self-defense – makes it difficult to apply to other conflicts. This is perhaps illustrated by Meta’s failure to implement similar policies in Ethiopia. Relying on such a narrow justification will lead to inconsistent treatment of different armed conflicts, particularly those between state and nonstate actors. Instead of backing just causes by granting them exceptions, Meta should strive to apply standards inspired by the laws of armed conflict to the content of all combatants. Toward Platform Laws of Wars At the State of the Net conference days after Russia began its invasion of Ukraine, I discussed the need for platforms to establish their own “laws of war”, or alternative community standards tailored to the unique demands of wartime communication. Community standards designed for peacetime often produce unwanted outcomes when applied to conflict-content. In the face of this reality, Meta was quick to make exceptions to its existing policies for Ukraine. Meta’s decision to support Ukraine outright is laudable, but ultimately unsustainable. It will be more difficult to identify the “right side” in future conflicts. If Meta makes a policy of supporting just causes, it may end up turning a blind eye to ambiguous conflicts in which no just cause can be found. Instead of supporting one side over another, Meta should strive to enforce standards of just conduct in war. Social media platforms cannot hope to prevent war, or stop harm during wartime, but they may be able to curb unique wartime abuses. Preventing the Greater Harm Platform community standards are traditionally intended to prevent physical and emotional harm. During wartime, however, this goal becomes nonsensical. In war, combatants on both sides aspire to do harm to their enemies. Violence is justified as a way of preventing further or more lasting harms, such as subjugation by a foreign power. In light of this wartime prerogative, platform prohibitions on calling for harm or organizing harmful activities are an ill-fitting default that benefits less media reliant combatants. Crowdfunding platform Patreon’s removal of fundraiser for the Ukrainian military illustrates how content moderation’s focus on harm can lead to perverse outcomes during wartime. Ukrainian NGO “Come Back Alive” uses donations to purchase protective equipment and ammunition for Ukrainian soldiers. In the early days of the war, Patreon removed Come Back Alive’s page because the platform prohibits fundraising “for anything that facilitates harmful or illegal activities.” While the extent to which the categories “harmful” and “illegal” overlap will always be debated, they move apart during wartime, when law provides for, if not encourages or mandates, doing harm. Indeed, to the extent that Patreon avoided facilitating whatever harm Ukrainians might have done with Patreon funded ammunition, Russians may have been able to do more harm to Ukrainian combatants, civilians, and infrastructure. Although this was not the intent of Patreon’s policy, it is the most likely effect. The Patreon example shows that both platform rules and the conceptions of harm that they must be reconsidered in wartime. A First Step On March 10th, Meta took a first step toward wartime community standards. Reuters reported that Meta had relaxed its prohibitions on calls for violence and incitement when such speech was directed at the Russian military or leaders. Although an initial headline implied that the policy changes covered incitement directed at Russians in general, a quoted Meta email made clear that it only applied in the context of Russia’s invasion. "We are issuing a spirit-of-the-policy allowance
to allow T1 violent speech that would otherwise be removed under the Hate Speech policy when: (a) targeting Russian soldiers, EXCEPT prisoners of war, or (b) targeting Russians where it's clear that the context is the Russian invasion of Ukraine (e.g., content mentions the invasion, self-defense, etc.)” This is a reasonable change that accommodates predictable nuances in the use of language during wartime. However, it is one-sided. It applies only to speech by Ukrainians about Russians, and not vice-versa. While the Russian government has since restricted access to Instagram in Russia, limiting the extent to which such an allowance would actually be used, this restriction was ostensibly a response to Meta’s relaxed incitement policies. This is not to say that such a one-sided policy cannot be justified, indeed, Meta President of Global Affairs Nick Clegg grounds the policy in self-defense, and, to an extent, western publ
Case Review No 2021-014-FB-UA Comment by Dr. Samson Esayas

We welcome the Oversight Board's initiative to seek public comments. In this document, we briefly address some of the points that the Board sought feedback on. (1) How Meta enforces its Violence and Incitement policy in conflict situations, including whether its actions are consistent across different conflicts. While we believe that Meta does its best to enforce its policy across conflict situations, the swift measures taken by Meta following the Russian invasion of Ukraine lay bare the differential treatment between this conflict and conflicts in other regions, particularly Ethiopia and Myanmar. One of Meta’s measures in the Ukraine war is particularly noteworthy in light of its relevance to the present case. In the case at hand, a post calling on “the national army to “turn its gun” against the Prime Minister” was considered to violate Meta’s policy and was subsequently removed. In contrast, as reported by Reuters, Meta made temporarily changes to its policy that “will allow Facebook and Instagram users in some countries to call for violence against Russians and Russian soldiers in the context of the Ukraine invasion’. The difference is striking, allowing “users in some countries to call for violence against Russians and Russian soldiers”
but taking down a post calling “the national army to “turn its gun” against the Prime Minister” who is the commander-in-chief of the army in conflict. (2) Whether and how Meta's cross-check program should work during an armed conflict. We believe that engagement with local communities and representation of different voices is essential in addressing the multifaceted challenges of fact-checking programs. We recognize the positive actions that Facebook has taken thus far regarding the situation in Ethiopia, including the increased efforts to tackle hate and inflammatory language. However, this is an area that requires further attention. The British Newspaper, the Guardian, recently published a report noting the Meta's lack of engagement with communities in Ethiopia, primarily relying on Nairobi-based content moderators. Even in cases where Meta has taken initiatives to engage, we believe that Tigrayan voices have been neglected in discussions about tackling hate speech and violence incitement in Ethiopia. Given that Tigrayans have been the primary victims of hate speech and violence in the current conflict, it is of utmost importance to have their voices heard in all relevant content moderation decisions. (3) Content moderation challenges specific to Ethiopia and languages spoken in the country, particularly during times of heightened tension or conflict. Over the last few months, several international media outlets have published reports that put Facebook’s role in Ethiopia under the spotlight. Citing an internal document from Facebook, the WSJ reported in September last year about the difficulties Facebook encountered in removing harmful content targeting Tigrayans. According to the report, content inciting violence against Tigrayans “slipped through the cracks due to a lack of moderators who speak the native language.” More recently, the British Newspaper, the Guardian, published a similar report titled “Facebook ‘lets vigilantes in Ethiopia incite ethnic killing’”. Citing an investigation by the Bureau of Investigative Journalism (TBIJ) and the Observer, the Guardian report accuses Facebook for “inaction and indifference” in the face of “content inciting violence through hate and misinformation.” In one instance, a Facebook post calling for cleansing Tigrayans from Amhara region, in particular Gondar city, has remained on the platform for more than four months. Thus, the fact that Facebook’s systems are not able to detect content written in Ethiopian languages often plays either in favor of keeping the content in the platform or its knee-jerk removal. (4) Evidence or analysis of statements from armed groups or the military in Ethiopia on social media that have incited or instigated violence, including any violations or abuses of international law. Tigrayans living in the city of Gondar in Amhara region has been subjected to abuses and the investigation by TBIJ links some of these abuses to hate campaigns on social media, particularly Facebook. The report refers to the killing of a Tigrayan jeweller, a resident of Gondar and along with other 11 people in November 2021. According to people interviewed by the TBIJ, hate speech against Tigrayans on Facebook “played a key role in not only his killing, but many others.” It’s also worth recalling that the PM and the official Facebook page of the Ethiopian national army have used references such as “cancers”, and “day-time hyenas” to portray Tigrayan. (5) The information environment in Ethiopia during the conflict, including access to the internet and independent sources of reporting, and how this should influence Meta's approach to moderating content from parties to the conflict. The Tigray region has been cut
off from all means of communications for more than a year. This complete blockade of region coupled with the Ethiopian government’s assault on media and journalists, has allowed the Federal to take complete control of the narrative using state media, and social media. In this regard, platforms such as Facebook offered Tigrayans abroad the possibility to raise awareness about the war and counter the narrative of the Federal government. Thus, given that Tigray is cut off from the outside world, extreme caution is required in supressing the only voices for Tigrayans.

Link to Attachment
PC-10433
Meta's "cross-check", "automated systems", "user reports", "initial review", "another review" of the content in question all share the same flaw: they are done in a black box and only the civil society groups contacted by Meta's team can provide feedback. The real solution is to empower civil society, multilaterals and NGOs by exposing additional API endpoints to enable these groups to access Meta's content moderation pipeline in real-time and provide feedback.

Whenever there is a given piece of content up for review, Meta proceeds in a black-box approach where it controls all information and decision-making. It may reach out to Trusted Partners such as CSOs, NGOs, and UN organizations for additional feedback, but Trusted Partners are not made aware of each other's comments and Meta does not inform Trusted Partners of its own internal decision-making process by its policy teams. When a user reports a given piece of content, only the reporting user sees the updated status of the review. Metadata of reported content and Meta's own content moderation process is not made readily available to civil society. Things could be done different. Namely, Meta could expand the API endpoints exposed to its Trusted Partners through Crowdtangle by exposing additional endpoints from its content moderation pipeline, i.e. - GET /reports where a Trusted Partner could retrieve reported content in real time by language, country, search criteria, just like for the existing Crowdtangle /posts endpoint. Importantly, extra params would include current status of the content (taken down, left up), the current decision status (pending, decision rendered, on intermediate appeal, reversed on intermediate appeal, affirmed on intermediate appeal, on appeal to OSB, reversed by OSB, affirmed by OSB, etc.) and user (user account suspended, user account banned) and the *grounds* on which the content decision was made (which Community Guideline(s) violated or not, and provided rationale). Reporting
user accounts could be anonymized if the user clicks an "anonymous" checkbox in the GUI on report submission. The post id and post data (from the GET /posts endpoint) of the reported post would be provided in the returned JSON object. - POST /report/:id where a Trusted Partner could submit their own recommended decision and rationale for a given piece of content, their language and country focus, and their organizational name. - GET /trustedPartnerReports where other Trusted Partners can retrieve posted decisions + rationales via the previous POST endpoint from Trusted Partner(s) in order to share analysis, so Trusted Partners aren't in the dark as to what their colleagues around the world and in-country are doing. - POST /comment where Trusted Partners could file complaints, praise, or general comments on how Meta's content moderation process is going, with ideally these comments shared to both Meta and OSB in real time. This is a way Trusted Partners can make general comments on Meta's system independent of a specific content piece. - GET /comments where Trusted Partners could see the comments made by other Trusted Partners. Webhooks. Crowdtangle features Notifications in their GUI (https://help.crowdtangle.com/en/articles/2580674-notifications) but not programmatically. Exposing webhooks such that Trusted Partners could receive real-time updates of change in decision (reversed, affirmed, etc.) and content (banned, left up, etc.) statuses would be huge because Trusted Partners could be updated of such changes in real time without having to keep calling Crowdtangle's APIs. Developing and exposing these API endpoints and webhooks would be a game changer and globally democratize content moderation and review. Instead of all of us waiting around for Meta to complete its black-box decision making, with the OSB having some limited, high-touch review, civil society and multilaterals could track decisions *as they are being made*, in real time, at scale, across country context and language. Data scientists could analyze content moderation datasets for bias in Meta's decision-making. Computer scientists could build competing NLP models. Journalists could expose strengths and weaknesses of Meta's current system. All of these initiatives could make much more muscular OSB's own review. I highly recommend the OSB consider issuing the foregoing as a policy recommendation for Meta to implement.
Meta should be focusing its resources on government agents and other entities that deliberately post false content on Facebook to entice hate and violence against victims of war as is the case in Ethiopia. The govt of Ethiopia has waged a war against the minority population of Tigray which includes information propaganda and hate that has resulted in the death of thousands of innocent civilians. Meta's focus should not be on words or statements made by the victims of hate in the context of war. By all accounts, the content of the Tigray Bureau are trivial and inconsequential. Meta should not remove content by the Tigray Bureau as the victims in Tigray have limited opportunity for their voice and atrocities happening to them to be heard.

The case as presented by Meta is quite intriguing at many levels. Meta has identified a content posted in Amharic by the Tigray Communications Affairs Bureau (TCAB) which in part states “...it goes on to say that the armed forces must surrender to the TPLF if they hope to save their lives, and if they refuse, they will die...” We are truly struggling to find the fault in the statement by TCAB. As you are well aware, there are two main choices for any soldier that engages in a war situation: surrender to your enemy, or potentially die fighting. With all due respect to Meta, which part of this statement is controversial? If anything, TCAB is urging fighters to surrender so they can save their lives rather than die fighting. Meta goes on to point to another partial statement from TCAB where the TPLF tells the army fighting against them to metaphorically “turn its guns” to the real enemy—meaning they should be fighting against the genocidal government of Ethiopia that is causing death of millions of people and destruction throughout the country. By any yardstick, the content referred by Meta to the Board is neither hate speech nor inciting violence which begs the question: Why is Meta not focusing on content that has caused war,
massacres, and death of ethnic minority Tigrayans in Ethiopia? In investigating this case, we ask the Board to examine the potential political motivation of the 10 users who reported this content to Facebook for violating the Violence and Incitement, Dangerous Individuals and Organizations, and Hate Speech policies. Unfortunately, this is an information war where the truth has become a primary victim. It is well known now that the government of Ethiopia has a track record of lying and misrepresentation, falsely accusing reputable organizations such as US news media and western governments of causing death in Ethiopia. This case is no different. As widely reported by international media, the government of Ethiopia has been, through its agents, posting content that has incited the murder and massacres of ethnic Tigrayans. Many government officials in Ethiopia including the prime minister himself have publicly used genocidal and dehumanizing words against ethnic Tigrayans such as “cancers” and “weeds”. As the Guardian in its publication of February 20, 2022, put it: “Facebook ‘lets vigilantes in Ethiopia incite ethnic Killing’.” A case in point is the testimony by former Facebook employee and whistleblower Frances Haugen who testified before the US Congress how inciting content on Facebook is causing violence and wars in Ethiopia. Wilfully or otherwise, we believe Meta has misplaced its focus and it is using its vast resources and time investigating a case, which in the grand scheme of things, can be considered trivial and inconsequential, especially under an armed conflict scenario as it is the case here. The statement by TCAB was made in the context of war. In focusing on this matter, Meta is allowing itself to be distracted from more important matters. In particular, it is ignoring dangerous postings by the Ethiopian government agents and others that are responsible for wars and violence in Ethiopia. In conclusion, we urge the Board to let the original determination by Meta stand based on the fact that the content was not violating Meta policies on inciting hate or violence.

Link to Attachment
PC-10427
<table>
<thead>
<tr>
<th>Case number</th>
<th>Public comment number</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Withheld</td>
<td>Withheld</td>
<td>Europe</td>
</tr>
<tr>
<td>Commenter’s first name</td>
<td>Commenter’s last name</td>
<td>English</td>
</tr>
<tr>
<td>Withheld</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Organization</td>
<td>Response on behalf of organization</td>
<td>No</td>
</tr>
</tbody>
</table>

---

Short summary provided by the commenter

See file attached.

Full Comment

See file attached.

Link to Attachment

PC-10425
This response is on the Ethiopia case, re 'Tigray Communication Affairs Bureau' (2022-006-FB-MR).
I submit that this content indeed is calling for serious violence against the legitimate, democratically elected federal government of Ethiopia by an insurgent movement (TPLF) within the country and known for horrendous human rights abuses against civilians since it started its armed campaigns on 3-4 November 2020.
The complaint is justified.

As the Oversight Board said on the website, the content to which objections were made was posted in Amharic by the Tigray Communication Affairs Bureau page, which states that it is the official page of the Tigray Regional State (Ethiopia) Communication Affairs Bureau. Its contents are mendacious and threatening. - the number of losses suffered by the Federal National Defense Forces in this war (that was started without warning by the Tigray TPLF on 4 November 2020 with a mass murder on sleeping federal army soldiers) are not known. The figures are incorrect. At the same, the statement does not mention their own losses. This is propaganda tactics. - the statement that "the armed forces [of the federal Ethiopian government] must surrender to the TPLF if they hope to save their lives, and if they refuse, they will die" is a threat by an illegal, insurgent movement (that was declared 'terrorist' by the Ethiopian Parliament in May 2021) against the legitimate national defense forces. That 'Tigray Bureau' is not in a position - neither legally, political or morally - to utter such a threat; they are not an independent national government but a regional one, bound by federal laws. They flouted these laws and Ethiopian national security by starting an armed rebellion targeting not only federal troops but also civilians and continuing a war until this this day. - the threats in that statement
encouraging the national army to “turn its gun” against "the Prime Minister's group" "in order to make amends with the people it has harmed" constitute a strange if not preposterous call for sedition and for violence against the legitimate political leaders, civil servants and civilians on the side of the federal government and those identified by the Tigray Bureau with it. It is therefore calling directly for violent action against the sovereign national government (At the same time, the Bureau keeps silent about the horrendous violence that the organization/movement (TPLF), of which the Bureau is the mouthpiece, has perpetrated against the civilian populations and against federal prisoners-of-war). Conclusion: The content of the Facebook statement of the Tigray Communication Affairs Bureau violates the Violence and Incitement, Dangerous Individuals and Organizations, and Hate Speech policy of Meta and is rightly removed.