


UPHELD

2021-014-FB-UA

Alleged crimes in Raya Kobo


The Oversight Board has upheld Meta's original decision to remove a post alleging the involvement of ethnic Tigrayan civilians in atrocities in Ethiopia's Amhara region.

Policies and topics

 Freedom of expression, War and conflict

 Hate speech

Region and countries

 Subsaharan Africa

 Ethiopia

Platform

 Facebook

Attachments

[Amharic translation](#)

[Tigrinya translation](#)

[Public Comments 2021-014-FB-UA](#)

This decision is also available in [Amharic](#) and [Tigrinya](#).

ሙሉ ውሳኔውን በአማርኛ ለማንበብ፣ [እዚህ ይጫኑ](#)።

ብትግርኛ አተገብረ ውሳኔ ምሉእ ከተንብቦ አንተ ደሊኻ [ካብዚ ጠውቅ](#)።

Case summary

Note: On October 28, 2021, Facebook announced that it was changing its company name to Meta. In this text, Meta refers to the company, and Facebook continues to refer to the product and policies attached to the specific app.

The Oversight Board has upheld Meta’s original decision to remove a post alleging the involvement of ethnic Tigrayan civilians in atrocities in Ethiopia’s Amhara region. However, as Meta restored the post after the user’s appeal to the Board, the company must once again remove the

content from the platform.

About the case

In late July 2021, a Facebook user from Ethiopia posted in Amharic. The post included allegations that the Tigray People's Liberation Front (TPLF) killed and raped women and children, and looted the properties of civilians in Raya Kobo and other towns in Ethiopia's Amhara region. The user also claimed that ethnic Tigrayan civilians assisted the TPLF with these atrocities. The user claims in the post that he received the information from the residents of Raya Kobo. The user ended the post with the following words "we will ensure our freedom through our struggle."

After Meta's automatic Amharic language systems flagged the post, a content moderator determined that the content violated Facebook's Hate Speech Community Standard and removed it. When the user appealed this decision to Meta, a second content moderator confirmed that the post violated Facebook's Community Standards. Both moderators belonged to Meta's Amharic content review team.

The user then submitted an appeal to the Oversight Board. After the Board selected this case, Meta identified its original decision to remove the post as incorrect and restored it on August 27. Meta told the Board it usually notifies users that their content has been restored on the day they restore it. However, due to a human error, Meta informed this user that their post had been restored on September 30 – over a month later. This notification happened after the Board asked Meta whether it had informed the user that their content had been restored.

Key findings

The Board finds that the content violated Facebook's Community Standard on Violence and Incitement.

While Meta initially removed the post for violating the Hate Speech Community Standard, the company restored the content after the Board selected the case, as Meta claimed the post did not target the Tigray ethnicity and the user's allegations did not constitute hate speech. The Board finds this explanation for restoring the content to be lacking detail and incorrect.

Instead, the Board applied Facebook's Violence and Incitement Community Standard to this post. This Standard prohibits "misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm." The Board finds that the content in this case contains an unverifiable rumor according to Meta's definition of the term. While the user claims his sources are previous unnamed reports and people on-the-ground, he does not even provide circumstantial evidence to support his allegations. Rumors alleging that an ethnic group is complicit in mass atrocities, as found in this post, are dangerous and significantly increase the risk of imminent violence.

The Board also finds that removing the post is consistent with Meta's human rights responsibilities as a business. Unverifiable rumors in a heated and ongoing conflict could lead to grave atrocities, as was the case in Myanmar. In decision [2020-003-FB-UA](#), the Board stated that "in situations of armed conflict in particular, the risk of hateful, dehumanizing expressions accumulating and spreading on a platform, leading to offline action impacting the right to security of person and potentially life, is especially pronounced." Cumulative impact can amount to causation through a "gradual build-up of effect," as happened in the Rwandan genocide.

The Board came to its decision aware of the tensions between protecting freedom of expression and reducing the threat of sectarian conflict. The Board is aware of civilian involvement in the atrocities in various parts of Ethiopia, though not in Raya Kobo, and the fact that Meta could not verify the post's allegations at the time they were posted. The Board is also aware that true reports on atrocities can save lives in conflict zones, while unsubstantiated claims regarding civilian perpetrators are likely to heighten risks of near-term violence.

The Oversight Board's decision

The Oversight Board upholds Meta's original decision to remove the post. As Meta restored the content after the user's appeal to the Board, the company must once again remove the content from the platform.

In a policy advisory statement, the Board recommends that Meta:

- Rewrite its value of "Safety" to reflect that online speech may pose risk to the physical security of persons and the right to life, in addition to the risks of intimidation, exclusion and silencing.
- Reflect in the Facebook Community Standards that in the contexts of war and violent conflict, unverified rumors pose higher risk to the rights of life and security of persons. This should be reflected at all levels of the moderation process.

- Commission an independent human rights due diligence assessment on how Facebook and Instagram have been used to spread hate speech and unverified rumors that heighten the risk of violence in Ethiopia. The assessment should review the success of measures Meta took to prevent the misuse of its products and services in Ethiopia. The assessment should also review the success of measures Meta took to allow for corroborated and public interest reporting on human rights atrocities in Ethiopia. The assessment should review Meta’s language capabilities in Ethiopia and if they are adequate to protect the rights of its users. The assessment should cover a period from June 1, 2020, to the present. The company should complete the assessment within six months from the moment it responds to these recommendations. The assessment should be published in full.

*Case summaries provide an overview of the case and do not have precedential value.

Full case decision

1. Decision summary

The Oversight Board upholds Meta’s original decision to remove the content. The post alleges the involvement of ethnic Tigrayan civilians in the atrocities against the people in Ethiopia’s Amhara region. Meta initially applied the Hate Speech Community Standard to remove the post from Facebook, but restored it after the Board selected the case. The Board finds Meta’s explanation for restoration lacking detail and incorrect. The Board finds that the content violated the prohibition on unverified rumors under the Violence and Incitement Community Standard.

2. Case description

In late July 2021, a Facebook user posted in Amharic on his timeline allegations that the Tigray People’s Liberation Front (TPLF) killed and raped women and children, as well as looted the properties of civilians in Raya Kobo and other towns in Ethiopia’s Amhara region. The user also claimed that ethnic Tigrayan civilians assisted the TPLF in these atrocities (for the specific translation and meaning of the Amharic post see Section 6 below). The user ended the post with the following words “we will ensure our freedom through our struggle.” The user claims in the post that he received the information from the residents of Raya Kobo.

The post was viewed nearly 5,000 times, receiving fewer than 35 comments and more than 140 reactions. It was shared over 30 times. The post remained on Facebook for approximately one day. Among the comments in Amharic were statements, as translated by the Board’s linguistic experts, stating that: “[o]ur only option is to stand together for revenge” and “are you ready, brothers and sisters, to settle this matter?” According to Meta, the user’s account that posted the content is located in Ethiopia, but not in the Tigray or Amhara regions. The user’s profile picture includes a hashtag signaling disapproval of the TPLF. Based on the information available to the Board, the user describes himself as an Ethiopian man from Raya.

The post was identified by Meta’s Amharic language automated system (classifier) as potentially violating its policies. Meta operates machine learning classifiers that are trained to automatically detect potential violations of the Facebook Community Standards. Meta announced that it is “using this technology to proactively identify hate speech in Amharic and Oromo, alongside over 40 other languages globally.” The Board understands that Ethiopia is a multilingual country, with Oromo, Amharic, Somali, and Tigrinya being the four most spoken languages in the country. Meta also reported that it hires moderators who can review content in Amharic, Oromo, Tigrinya, and Somali.

Meta uses a “‘biased sampling’ method that samples content to improve the Amharic classifier quality. This means that Amharic content with both low and high potential match scores is continuously sampled and enqueued for human review to improve classifier performance.” This content was selected for human review as part of that improvement process. Meta also explained that its automated system determined that this content had “a high number of potential views” and that it gave the post “a low violating score.” The low violating score means that the content does not meet the threshold for auto-removal by Meta’s automated system.

A content moderator from the Amharic content review team determined that the post violated Facebook’s Hate Speech Community Standard and removed it. This Standard prohibits content targeting a person or group of people based on their race, ethnicity, or national origin with “violent speech.” Meta stated that it notified the user that his post violated Facebook’s Hate Speech Community Standard, but not the specific rule that was violated. The user then appealed the decision to Meta, and, following a second review by another moderator from the Amharic content review team, Meta confirmed that the post violated Facebook’s policies.

The user then submitted an appeal to the Oversight Board. As a result of the Board selecting the case, Meta identified the post's removal as an "enforcement error" and restored it on August 27. Meta stated that it usually notifies users about content restoration on the same day. However, due to a human error, Meta informed this user of restoration on September 30. This happened after the Board asked Meta whether the user had been informed that their content had been restored.

The case concerns unverified allegations that Tigrayans living in Raya Kobo town were collaborating with the TPLF to commit atrocities including rape against the Amhara ethnic group. These allegations were posted on Facebook in the midst of an ongoing civil war in Ethiopia that erupted in 2020 between the Tigray region's forces and Ethiopian federal government forces and military and its allies (International Crisis Group, [Ethiopia's Civil War: Cutting a Deal to Stop the Bloodshed](#) October 26, 2021).

According to expert briefings received by the Board, Facebook is an important, influential and popular online medium for communication in Ethiopia. The expert briefings also noted there is little to no coverage on the conflict-affected areas in Ethiopian media, and Ethiopians use Facebook to share and receive information about the conflict.

In its recent history, Ethiopia has seen recurring ethnic conflict involving, among others, Tigrayan groups (ACCORD, [Ethnic federalism and conflict in Ethiopia](#), 2017). The Board is aware of allegations of serious violations of human rights and humanitarian law in the Tigray region and in other parts of the country, including in Afar, Amhara, Oromo and Somali regions by the involved parties in the current conflict ([UN Special Advisor on the Prevention of Genocide, on the continued deterioration of the situation in Ethiopia statement](#), July 30, 2021; Office of the United Nations High Commissioner for Human Rights (UN OHCHR), [Ethiopia: Bachelet urges end to 'reckless' war as Tigray conflict escalates](#), November 3, 2021). Furthermore, according to the recently published [joint investigation](#) by the Ethiopian Human Rights Commission (EHRC) and the UN OHCHR, Tigrayan and Amharan civilians were involved in human rights violations in late 2020. However, the scope of the investigation did not cover violations during July 2021 in areas mentioned in the user's post (EHRC and UN OHCHR report, [Joint Investigation into Alleged Violations of International Human Rights, Humanitarian and Refugee Law Committed by all Parties to the Conflict in the Tigray Region of the Federal Democratic Republic of Ethiopia](#), November 3, 2021).

According to a Reuters report, local officials from Amhara region claimed that the Tigrayan forces killed 120 civilians in a village in Amhara region on September 1 and 2 ([Reuters](#), September 8). The Tigrayan forces later issued a statement rejecting what they called a "fabricated allegation" by the Amhara regional government. These allegations could not be independently confirmed. The Board is also aware of allegations that Tigrayans are ethnically profiled, harassed and are increasingly subject to hate speech ([Remarks of the UN High Commissioner for Human Rights Michelle Bachelet in Response to Questions on Ethiopia](#), December 9, 2020; [NGOs Call for UN Human Rights Council Resolution on Tigray](#), June 11, 2021).

3. Authority and scope

The Board has authority to review Meta's decision following an appeal from the user whose post was removed (Charter Article 2, Section 1). The Board may uphold or reverse that decision, and its decision is binding on Meta (Charter Article 3, Section 5, and Article 4). The Board's decisions may include policy advisory statements with non-binding recommendations that Meta must respond to (Charter Article 3, Section 4). According to its Charter, the Oversight Board is an independent body designed to protect free expression by making principled, independent decisions about important pieces of content. It operates transparently, exercising neutral, independent judgement and rendering decisions impartially.

4. Relevant standards

The Oversight Board considered the following standards in its decision:

1. Facebook's Community Standards

In the policy rationale for Facebook's Hate Speech Community Standard, Meta states that hate speech is not allowed on the platform "because it creates an environment of intimidation and exclusion and, in some cases, may promote real-world violence." The Community Standard defines hate speech as "a direct attack against people — rather than concepts or institutions — on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease." The rationale further defines an attack "as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation." Facebook's Hate Speech Community Standard describes three tiers of attacks. Under Tier 1, Facebook's Community Standards prohibit "content targeting a person or group of people

(including all subsets except those described as having carried out violent crimes or sexual offenses) on the basis of their aforementioned protected characteristic(s) with “dehumanizing speech.” Such speech can take the form of generalizations or unqualified behavioral statements about people sharing a protected characteristic being “violent and sexual criminals” or “other criminals.”

The rationale for Facebook’s Violence and Incitement Community Standard states that Meta “aim[s] to prevent potential offline harm that may be related to content” on the platform. Specifically, Meta prohibits content containing “misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm.” As part of its Internal Implementation Standards, Meta considers an unverifiable rumor to be information which is extremely hard or impossible to trace the source of the information, or cannot be confirmed or debunked in a meaningful timeframe because it is extremely hard or impossible to trace the source of the information. Meta also considers information that is devoid of enough specificity for the claim to be debunked to be an unverifiable rumor. Meta notes that it requires additional context to enforce this policy found in Facebook’s Violence and Incitement Community Standard.

II. Meta’s values

Meta’s values are outlined in the introduction to Facebook’s Community Standards. The value of “Voice” is described as “paramount”:

The goal of our Community Standards has always been to create a place for expression and give people a voice. [...] We want people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable.

Meta limits “Voice” in service of four other values, and two are relevant here:

“Safety”: *We are committed to making Facebook a safe place. Expression that threatens people has the potential to intimidate, exclude or silence others and isn’t allowed on Facebook.*

“Dignity”: *We believe that all people are equal in dignity and rights. We expect that people will respect the dignity of others and not harass or degrade others.*

III. Human rights standards

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. In 2021, Meta announced its [Corporate Human Rights Policy](#), where it re-committed to respecting human rights in accordance with the UNGPs. The Board’s analysis in this case was informed by the following human rights standards:

- [Freedom of expression](#): Article 4 and 19 International Covenant on Civil and Political Rights ([ICCPR](#)); Human Rights Committee, General Comment No. 34, 2011 ([General Comment 34](#)); UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, [Online Content Regulation](#), A/HRC/38/35.
- [Right to life](#): ICCPR Article 6, as interpreted by General Comment No. 36, Human Rights Committee (2018) ([General Comment 36](#)); UN Working Group on the issue of human rights and transnational corporations and other business enterprises, Business, human rights and conflict-affected regions: towards heightened action report ([A/75/212](#)).

5. User statement

The user stated in his appeal to the Board that he posted this content to protect his community which is in danger and that Meta must help communities in war zones. He stated the post is not hate speech “but is truth.” He also stated that the TPLF targeted his community of one million people and left them without food, water and other basic necessities. The user speculated that his post was reported “by members and supporters of that terrorist group,” and claimed to “know well most of the rules” and that he has “never broken any rules of Facebook.”

6. Explanation of Meta’s decision

Meta explained in its rationale that the content was originally removed as an attack under Facebook’s Hate Speech Community Standard, specifically for violating its policy prohibiting “violent speech” targeted at Tigrayan people based on their ethnicity. Meta informed the Board that its moderators do not record their reasons for removing content, beyond indicating the Community Standard violated. Therefore, Meta did not confirm if the moderators who reviewed the post initially and on appeal applied the same rule within Facebook’s Hate Speech policy to remove the post.

Meta stated in its rationale that, as a result of the Board selecting the case, the company determined that its “decision was an error” and restored the post. Meta also stated that the content did not violate its rules because it did not target the Tigray ethnicity and the user’s allegations about the TPLF or Tigrayans did not rise to the level of hate speech.

Meta confirmed in its response to the Board’s question that its Amharic automated systems are in place and that these are audited and refreshed every six months. Meta also explained that it was the original text in Amharic that led the automated system to identify the content as potentially violating. Similarly, Meta confirmed that the two content moderators were Amharic speakers and that they based their review on the original text in Amharic.

Meta explained in its submissions that its regional team provided the cultural and linguistic context in developing the case file for this appeal. For example, Meta’s decision rationale presented to the Board is based on the regional team’s translation of the content. Meta’s regional team translated the supposedly violating part of the user’s post as “Tigrean” teachers, health professionals and merchants “are leading the way for the rebel TPLF forces to get women raped and loot properties.”

The Board requested and received an additional English translation of the text from its own linguistic experts and Meta provided an additional translation of the text by its external linguistic vendor. The two versions confirmed that the prevailing meaning of the text indicates that Tigrayan civilians assisted in the atrocities committed by the TPLF. For the purposes of this decision, the Board notes the version provided by Meta’s external vendor. That version reads as follows: “As reported previously and per information obtained from people living in the area who make a living as teachers, as health professionals, as merchants, as daily labour workers, and low [wage] workers, we are receiving direct reports that the Tigreans, who know the area very well, are leading the rebel group door-to-door exposing women to rape and looting property.” Moreover, the Amharic comments on the post stated that: “[o]ur only option is to stand together for revenge” and “are you ready, brothers and sisters, to settle this matter?”

Meta confirmed in its response to the Board’s question that Ethiopia is “designated as a Tier 1 At-Risk Country.” According to Meta, this is the highest risk level. Meta noted that it has designated Ethiopia as “a crisis location” for its content policy and integrity work. As such, it established “a top-level” Integrity Product Operations Center (IPOC) for Ethiopia’s June 2021 elections and another IPOC to monitor post-election developments in September. Both IPOCs ended in the same month that they were set up. Meta also stated that it has been treating Ethiopia as a “top-level crisis” by its Operations, Policy and Product teams. Meta stated that its “Crisis Response Cross-functional team” that focuses on Ethiopia convenes weekly to understand and mitigate ongoing risk. Meta added that this work does not change the way it reviews content that does not pose such a risk, and that the work did not affect its determination in this case.

In response to the Board’s question, Meta explained that its Trusted Partner(s) did not escalate the post for additional review for violations of misinformation and harms policies. Meta also noted that as there was no third-party fact check there “was no evidence to suggest that the claims made in the post were false or unverifiable rumor(s).”

7. Third-party submissions

The Oversight Board received 23 public comments related to this case. Six of the comments were from Sub-Saharan Africa, specifically Ethiopia, one from the Middle East and North Africa, one was from Asia Pacific and Oceania, five were from Europe and 10 were from the United States and Canada. The Board received comments from stakeholders including academia, private individuals and civil society organizations focusing on freedom of expression and hate speech in Ethiopia. The submissions covered themes including whether the content should stay on the platform, difficulties in distinguishing criticism of the TPLF from hate speech against the Tigrayan people, and Meta’s lack of content moderators who speak Ethiopian languages.

To read public comments submitted for this case, please click [here](#).

8. Oversight Board analysis

The case concerns allegations made during an ongoing civil and ethnic war in a region with a history of lethal ethnic conflict. There is a tension between protecting freedom of expression and reducing the threat of sectarian conflict. This tension can only be resolved through attention to the specifics of a given conflict. The Board is aware of civilian involvement in the atrocities in various parts of Ethiopia, though not in Raya Kobo (see relevant context of the conflict in Ethiopia in Section 2 above). Meta stated that it had no evidence that the content was false or unverifiable rumor. The Board notes that at the time of the posting Meta could not and did not proactively verify the allegations. It was not possible to verify the allegations given the communication blackout in Amhara region. The situation in Amhara was beyond access for international observers

and journalists. The Board is also aware that true reports on atrocities can be lifesaving in conflict zones by putting potential victims on notice of potential perpetrators. However, in an ongoing heated conflict, unsubstantiated claims regarding civilian perpetrators are likely to pose heightened risks of near-term violence.

8.1. Compliance with Community Standards

Meta restored the content because it found that the content is not hate speech (see Section 6 above). The Board finds that explanation to be lacking detail and incorrect. The Board finds the Violence and Incitement Community Standard relevant to this case. The Board concludes that the post violates the Violence and Incitement policy's prohibition on "misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm" (see Section 4, for a definition of an "unverifiable rumor").

The content falls within Meta's Internal Implementation Standards' definition of an unverifiable rumor (see Section 4, for a definition of an "unverifiable rumor"). The Board finds that rumors alleging the complicity of an ethnic group in mass atrocities are dangerous and significantly increase the risk of imminent violence during an ongoing violent conflict such as presently in Ethiopia. The Board understands that Tigrayans in Ethiopia, like other ethnic groups, are already subject to imminent, and, in some instances actual, violence and physical harm.

8.2. Compliance with Meta's values

The Board finds that Meta's decision to restore and allow the content is inconsistent with its values of "Dignity" and "Safety." The Board recognizes that "Voice" is Meta's paramount value, but the company allows for expression to be limited to prevent abuse and other forms of online and offline harm.

In the context of this case, "Voice" that exposes human rights violations is of utmost importance. However, the form that an expression takes in the midst of a violent conflict is also important. Speech that seemingly seeks to bring attention to alleged human rights violations while making unverified claims during an ongoing violent conflict that an ethnic group is complicit in atrocities runs the risk that it will justify or generate retaliatory violence. This is particularly pertinent in Ethiopia in the current crisis.

8.3. Compliance with Meta's human rights responsibilities

The Board finds that removing the content in this case is consistent with Meta's human rights responsibilities as a business under UNGP Principle 13, which requires companies to "avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur." In a heated and ongoing conflict, unverifiable rumors may lead to grave atrocities, which the experience in Myanmar has indicated. To mitigate such a risk a transparent system of moderating content in conflict zones, including a policy regarding unverifiable rumors, is a necessity.

Freedom of Expression and Article 19 of the ICCPR

Article 19 of the ICCPR provides broad protection for freedom of expression through any media and regardless of frontiers. However, the Article allows this right to be restricted under certain narrow and limited conditions, known as the three-part test of legality (clarity), legitimacy, and necessity, which also includes an assessment of proportionality. Although the ICCPR does not create obligations for Meta as it does for states, Meta has committed to respecting human rights as [set out in the UNGPs](#). This commitment encompasses internationally recognized human rights as defined, among other instruments, by the ICCPR. The UN Special Rapporteur on freedom of opinion and expression has suggested that Article 19, para. 3 of the ICCPR provides a useful framework to guide platforms' content moderation practices ([A/HRC/38/35](#), para. 6)

1. Legality (clarity and accessibility of the rules)

The requirement of legality demands that any restriction on freedom of expression is: (a) adequately accessible so that individuals have a sufficient indication of how the law limits their rights; and (b) that the law must be formulated with sufficient precision so that individuals can regulate their conduct. Further, a law may not confer unfettered discretion for the restriction of freedom of expression on those charged with its execution ([General Comment 34](#), para. 25).

The term "unverifiable rumor" is not defined in the public facing Community Standards. When Meta fails to explain key terms and how its policies are applied, users may find it difficult to understand if their content violates Facebook's Community Standards. However, as applied to the facts of this case in which an unverified allegation was made in the midst of an ongoing violent conflict, the Board finds that the term

“unverifiable rumor” provides sufficient clarity. The rumor was not verifiable for Meta, nor for the user who was not present in Raya Kobo. International observers and journalists also could not verify the rumor given the ongoing conflict and the communications blackout. In such circumstances it is foreseeable for users that such a post falls within the prohibition.

II. Legitimate aim

Restrictions on freedom of expression must pursue a legitimate aim, which includes the protection of the rights of others, among other aims. The Human Rights Committee interpreted the term “rights” to include human rights as recognized in the ICCPR and more generally in international human rights law ([General Comment 34](#), para. 28). The Facebook Community Standard on Violence and Incitement exist in part to prevent offline harm that may be related to content on Facebook. Restrictions based this policy thus serve the legitimate aim of the protection of the right to life.

III. Necessity and proportionality

The principle of necessity and proportionality under international human rights law requires that restrictions on expression “must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected” ([General Comment 34](#), para. 34). The principle of proportionality demands a consideration for the form of expression at issue ([General Comment 34](#), para. 34).

In assessing whether the restriction on user’s speech served its aim as required by Meta’s human rights responsibilities, the Board considered what Meta has done to prevent and mitigate the risk to life from the spread of unverifiable rumors about parties to the Ethiopian conflict (see Section 6 above, for a description of Meta’s work in Ethiopia).

The user alleged that Tigrayan civilians were accomplices in grave atrocities committed by Tigrayan forces. The user’s sources for this claim are previous unnamed reports and sources on-the-ground, but he did not provide even circumstantial evidence supporting the allegations which he could have added without putting at risk his sources.

The Board is aware of the importance of shedding light on human rights violations in a conflict situation. Reporting on atrocities is an important activity serving the right of others to be informed. “Journalism is a function shared by a wide range of actors... [including] bloggers and others who engage in forms of self-publication ... on the internet...” ([General Comment 34](#), para. 44). Those who engage in forms of self-publication share the responsibilities related to the watchdog function and when reporting on human rights they must meet standards of accuracy. Moreover, information on atrocities may save lives, especially where social media are the ultimate source of information. However, the above qualities are absent in the user’s post as it does not contain information about actual threat to life and it does not contain specific information that can be used in the documentation of human rights violation. As formulated, the content could contribute to ethnic hatred. Ethnic conflict situations call for heightened scrutiny over how people should report on and discuss human rights violations committed by parties to the conflict. These considerations apply to Facebook posts, which can be used to spread unverifiable rumors at great speed.

The Board notes that some Ethiopian government officials have instigated or spread hate speech targeting Tigrayans (see, for example, Amnesty International, [Ethiopia: Sweeping emergency powers and alarming rise in online hate speech as Tigray conflict escalates](#), [DW 2020 report](#)). There is no evidence that this post formed part of such deliberate efforts to fan discord, but the present content has to be considered in view of reports that some Ethiopian government officials and public figures instigated or spread hate speech. Good faith postings or information on matters of public concern can enable vulnerable populations to better protect themselves. Additionally, a better understanding of important events may help in the pursuit of accountability. However, unverified rumors can feed into hateful narratives and contribute to their acceptance, especially in the absence of counter-speech efforts.

The Board finds that in a country where there is an ongoing armed conflict and an assessed inability of governmental institutions to meet their human rights obligations under international law, Meta may restrict freedom of expression that it otherwise would not (see ICCPR Article 4 on derogations in times of public emergencies). The principle of proportionality must take account of “the form of expression at issue as well as the means of its dissemination” ([General Comment 34](#), para. 34). On its own, an unverifiable rumor may not directly and immediately cause harm. However, when such content appears on an important, influential and popular social media platform during an ongoing conflict, the risk and likelihood of harm become more pronounced. The Board came to a similar conclusion in decision [2020-003-FB-UA](#). There, the Board found that “in situations of armed conflict in particular, the risk of hateful, dehumanizing expressions accumulating and spreading on a platform, leading to offline action impacting the right to security of person and potentially life, is especially pronounced.” Furthermore, cumulative impact can amount to causation through a “gradual build-up of effect,” as happened in Rwanda where calls to genocide were

repeated (see the *Nahimana* case, [Case No. ICTR-99-52-T](#), paras 436, 478, and 484-485). A direct call for violence is absent from the post in this case, although there is a reference to “our struggle.” Moreover, the content has been viewed by thousands of Amharan speakers in the 24 hours that it remained online. Some of them left comments that include calls for vengeance (see Section 2 above).

The right to life entails a due diligence responsibility to undertake reasonable positive measures, which do not impose disproportionate burdens on freedom of expression, in response to foreseeable threats to life originating from private persons and entities, whose conduct is not attributable to the state. Specific measures of protection shall be taken towards persons in situations of vulnerability (members of ethnic and religious minorities) whose lives have been placed at particular risk because of specific threats or pre-existing patterns of violence ([General Comment 36](#), paras 21 and 23). With the respective differences having been considered, these considerations are relevant to Meta’s responsibilities to protect human rights, because business is required to “seek to prevent or mitigate adverse human rights impacts that are directly linked to [its] operations, products or services” (UNGPs, Principle 13).

The United Nations Working Group on the issue of human rights and transnational corporations and other business enterprises declared that the UNGPs impose on businesses a heightened responsibility to undertake due diligence in a conflict setting (“Business, human rights and conflict-affected regions: towards heightened action,” [A/75/212](#), paras 41-54). Consequently, the legitimate aim to protect the right to life of others means that Meta has a heightened responsibility in the present conflict setting. The Board will consider this in its proportionality analysis.

The Board notes the steps Meta has taken so far in Ethiopia. The content was originally removed. However, Meta’s automated system determined that this content had “a low violating score,” thus Meta did not automatically remove the post. Even if the specific content was removed originally, Meta ultimately determined that the removal was an enforcement error. Meta also told the Board that the treatment of Ethiopia as Tier 1 At-Risk Country does not impact classifier performance or its ability to identify the content as potentially violating. The Board therefore concludes that without additional measures, Meta cannot properly fulfill its human rights responsibilities. The fact that Meta restored the content corroborates this concern.

In the present case the content was posted during an armed conflict. In such situations Meta has to exercise heightened due diligence to protect the right to life. Unverified rumors are directly connected to an imminent threat to life and Meta must prove that its policies and conflict-specific measures that it took in Ethiopia are likely to protect life and prevent atrocities (see Section 6 for Meta’s response to the Ethiopian conflict). In the absence of such measures, the Board has to conclude that the content must be removed. To prevent innumerable posts feeding into that narrative through unverified rumors, removal is the required measure in this case during an ongoing violent ethnic conflict.

A minority of the Board highlighted its understanding of the limited nature of this decision. In the context of an ongoing violent conflict, a post constituting an unverified rumor of ethnically-motivated violence by civilians against other civilians poses serious risks of escalating an already violent situation, particularly where Meta cannot verify the rumor in real time. Such increased risks triggered Meta’s human rights responsibility to engage in heightened due diligence with respect to content moderation involving the conflict. While it had various types of high alerts in place, Meta confirmed that such systems did not affect its determination in this case, which is difficult to understand given the risks of near-term violence. As noted in a previous decision of the Board ([2021-001-FB-FBR](#)), it is difficult to assess if measures short of content removal would constitute the least burden on a user’s speech to achieve a legitimate aim when Meta does not provide relevant information about whether its own design decisions and policies have amplified potentially harmful speech.

9. Oversight Board decision

The Oversight Board upholds Meta’s original decision to remove the content. Given that Meta subsequently restored the content after the user’s appeal to the Board, it must now remove the content once again from the platform.

10. Policy advisory statement

Content policy

1. Meta should rewrite Meta’s value of “Safety” to reflect that online speech may pose risk to the physical security of persons and the right to life, in addition to the risks of intimidation, exclusion and silencing.
2. Facebook’s Community Standards should reflect that in the contexts of war and violent conflict, unverified rumors pose higher risk to the rights of life and security of persons. This should be reflected at all levels of the moderation process.

Transparency

3. Meta should commission an independent human rights due diligence assessment on how Facebook and Instagram have been used to spread hate speech and unverified rumors that heighten the risk of violence in Ethiopia. The assessment should review the success of measures Meta took to prevent the misuse of its products and services in Ethiopia. The assessment should also review the success of measures Meta took to allow for corroborated and public interest reporting on human rights atrocities in Ethiopia. The assessment should review Meta's language capabilities in Ethiopia and if they are adequate to protect the rights of its users. The assessment should cover a period from June 1, 2020, to the present. The company should complete the assessment within six months from the moment it responds to these recommendations. The assessment should be published in full.

***Procedural note:**

The Oversight Board's decisions are prepared by panels of five Members and approved by a majority of the Board. Board decisions do not necessarily represent the personal views of all Members.

For this case decision, independent research was commissioned on behalf of the Board. An independent research institute headquartered at the University of Gothenburg and drawing on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world, provided expertise on socio-political and cultural context. The company Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world, provided linguistic expertise. Duco Advisers, an advisory firm focusing on the intersection of geopolitics, trust safety, and technology, also provided research.