Oversight Board

Public Comment Appendix for

Colombian police cartoon

Case number

Case description

In September 2020, a Facebook user in Colombia posted a picture of a cartoon as a comment on another user's post. The cartoon resembles the official crest of the National Police of Colombia and depicts three figures wearing police uniforms and holding batons over their heads. The figures appear to be kicking and beating another figure who is lying on the ground with blood beneath their head. A book and a pencil are shown next to the figure on the ground. The text on the crest reads in Spanish, "República de Colombia - Policía Nacional - Bolillo y Pata," which Meta's regional markets team translated to "National Police – Republic of Colombia – Baton and Kick." At the time the content was posted, there had recently been protests in Colombia against police violence.

According to Meta, in January 2022, 16 months after the content was originally posted to Facebook, the company removed the content as it matched with a picture in its "media matching bank" of content violating Facebook's Dangerous Individuals and Organizations Community Standard. Meta's mention of a "media matching bank" seems to refer to a system which helps Meta find duplicates of harmful media content and prevent them being shared. The Board is currently seeking more information from Meta about how the "media matching bank" is created, controlled, and used.

The user appealed this decision, and Meta maintained its decision to remove the content, but upheld the decision based on a different Facebook Community Standard, the Violence and Incitement Community Standard. At the time of removal, the content had received three views and no reactions. No users reported the content.
In their statement to the Board, the user primarily expresses confusion about how their content violated Meta's policies. They describe the content as reflecting reality in Colombia.

As a result of the Board selecting this case, Meta identified the removal of the content as an "enforcement error" and restored it. Meta explained to the Board that the removal decisions were wrong as the content did not contain a reference to any dangerous individual or organization, nor did it contain a threat of violence or statement of intent to commit violence. Meta also confirmed that, while they found the Violent and Graphic Content Community Standard relevant, the company did not regard the content to have violated this standard as "fictional imagery" is not prohibited.

The Board would appreciate public comments that address:

- Whether Meta's policies on Dangerous Individuals and Organizations, Violence and Incitement, and Violent and Graphic Content sufficiently respect expressions of political dissent, including against state/police violence.
- How Meta's use of "media matching bank" and automation could be improved to avoid the removal of non-violating content and enhance detection of violating content.
- Insights on the socio-political context in Colombia, particularly regarding the restriction of information and discussion on social media of protests and criticism of police violence.
- Insights into the role of social media globally in criticizing or documenting instances of police violence.

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.

Oversight Board

Public Comment Appendix for

Colombian police cartoon

Case number

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.

Oversight Board

Public Comment Appendix for
Colombian police cartoon

Case number

4

Number of Comments

Regional Breakdown

0	0	0	2
Asia Pacific & Oceania	Central & South Asia	Europe	Latin America & Caribbean
0	0	1	
Middle East and North Africa	Sub-Saharan Africa	United States & Canada	

Colombian police cartoon	PC-10437	Latin America and Caribbean
Case number	Public comment number	Region
PRISCILLA	RUIZ (et. al)	Spanish
Commenter's first name	Commenter's last name	Commenter's preferred language
Article 19 Oficina México y Centroamérica		Yes
Organization		Response on behalf of organization

____

Short summary provided by the commenter

La libertad de expresión protege las expresiones independiente de su tono, lo que implica aquellas que "chocan, irritan o inquietan a los funcionarios públicos o a un sector cualquiera de la población". La crítica permitida por parte de la sociedad es mayor cuando esta se refiere al gobierno en lugar de un ciudadano privado, o incluso de un político. La sátira está protegida por el derecho internacional y tiene características de exageración y distorsión de la realidad que buscan agitar y provocar.

Full Comment

La libertad de expresión protege las expresiones independiente de su tono, lo que implica aquellas que "chocan, irritan o inquietan a los funcionarios públicos o a un sector cualquiera de la población". La crítica permitida por parte de la sociedad es mayor cuando esta se refiere al gobierno en lugar de un ciudadano privado, o incluso de un político. La sátira está protegida por el derecho internacional y tiene características de exageración y distorsión de la realidad que buscan agitar y provocar. El Tribunal Europeo de Derechos Humanos (TEDH) ha indicado que las expresiones sin protección del derecho internacional son aquellas que alientan el odio o violencia o con la intención de destruir las libertades y derechos dispuestos en el Convenio Europeo de Derechos Humanos. Son incompatibles con el Convenio las expresiones con incitación a participar en luchas y acciones, a formar parte de grupos armados terroristas, a mostrar a miembros de esos grupos como héroes. En todo caso, es fundamental analizar la forma en que se realizan las declaraciones y su capacidad directa o indirecta de llevar a consecuencias dañinas. El Plan de Acción de Rabat establece factores que sirven para delimitar cuándo una expresión no está protegida: i) Las características del emisor; ii) la intención, iii) el contenido y la forma, iv) la extensión, v) la posibilidad de materialización y vi) su inminencia. La política de violencia e incitación no establece criterios para el análisis de contexto.

De una lectura textual, se da a entender que se aplica el algoritmo de manera automática sobre expresiones calificadas a priori como prohibidas, sin realizar un análisis de contexto. En el caso que nos ocupa, implicó la omisión de la ponderación de que se trataba de una sátira y del interés público. Las políticas hacen referencia, en algunas partes, a la posibilidad de que se materialicen acciones violentas, pero sin hacer referencia al análisis del riesgo ni a la probabilidad de materialización. La política de contenido violento y gráfico tiene un alcance limitado sobre contenidos que, aunque sensibles, no necesariamente constituyen incitaciones a la violencia. Es importante que su remoción aplique estándares como los mencionados en esta sección. La política de Personas y organizaciones peligrosas da definiciones claras de exaltación, apoyo sustancial y representación. Adicionalmente, indica de forma explícita los tipos de organizaciones peligrosas que son cubiertas. Se observa, además, una excepción de sátira, burla o crítica, lo que permitiría que su aplicación no se dé de forma que restrinja contenidos que contribuyen al debate público. *El análisis completo se encuentra en el documento adjunto al formato.

Link to Attachment

PC-10437

Colombian police cartoon	PC-10430	United States and Canada
Case number	Public comment number	Region
Kelsey	Zorzi	English
Commenter's first name	Commenter's last name	Commenter's preferred language
ADF International		Yes
Organization		Response on behalf of organization

____

Short summary provided by the commenter

In order to sufficiently respect expressions of political dissent, these three policies should define violence as physical force intended to damage person or property and should replace the unmodified word "harm" with "physical harm to person or property." In addition, the Dangerous Individuals and Organizations policy should be revised to clarify that the individuals and organizations in question are "dangerous" because they threaten the physical security of people or property and not merely because they express morally objectionable positions.

Full Comment

RE: Whether Meta's policies on Dangerous Individuals and Organizations, Violence and Incitement, and Violent and Graphic Content sufficiently respect expressions of political dissent, including against state/police violence. Background: The right to freedom of expression, recognized in Article 19 of the International Covenant on Civil and Political Rights, is among the most foundational international human rights. Any policy limitations on the right to freedom of speech should be limited and well-defined so that it is clear what constitutes a violation. Limitations on speech should also be clearly defined to prevent them from being so broadly interpreted that they subsume the freedom. International law is least protective of speech directly inciting violence. Individuals or organizations that incite or advocate violence are the most likely to be clearly recognizable and so within the competence of content moderators to police. While the three policies under review seemingly aim to allow for censorship of content inciting violence, their lack of clear definitions could lead to over-inclusive censorship that prohibits certain political viewpoints based on subjective interpretation. Dangerous Individuals and Organizations Policy: The policy should be revised to make clear that the individuals and organizations the policy intends to restrict are "dangerous" because they threaten the physical security of people or property. Currently the policy can be

read to restrict individuals and organizations that are "dangerous" in the sense that they express morally objectionable—bigoted, ignorant, misleading, etc.—viewpoints. For example, some might see an organization that harshly criticizes Modi's Hindu Nationalism as a "hate organization" that makes "statements… that attack individuals based on… religious affiliation." Censoring organizations and individuals that are "dangerous" in this broader sense of the word will unavoidably lead to prohibition of political viewpoints based on subjective disagreement and interpretation. Content moderators cannot be expected to have the competence or bandwidth to fairly and objectively navigate these issues, as they are likely to involve controversial ideological viewpoints, facts that are undiscoverable or unknowable to content moderators, judgments that rely on speculations about the impact of statements, and myriad other complex social complexities specific to each of the geographic regions in which Meta operates. If, despite these challenges, Meta does intend to censor individuals and organizations that express morally objectionable viewpoints, it should develop a separate policy that clearly delineates what it is doing and why. Clarifying that the individuals and organizations referenced in the policy under consideration are "dangerous" because they threaten the physical security of people or property would require five changes: 1. Insert a definition of "violence" as "physical force intended to damage person or property." 2. Replace the unmodified word "harm" with "physical harm to person or property." 3. Clarify references to "hate" to refer to hate-motivated incitement to the use of physical force to harm person or property. 4. Define Violence-Inducing Conspiracy Networks (VICNs) as groups that promote theories advocating for the use of physical force to harm person or property. 5. In the paragraph defining Tier 1, change "repeatedly dehumanizing or advocating for harm against people based on protected characteristics" to "repeatedly advocating for physical harm to person or property of people based on protected characteristics." Violence and Incitement Policy; Violence and Graphic Content Policy: These two policies should be clarified through the inclusion of a definition of "violence" as "physical force intended to damage person or property" and a definition of "harm" as "physical damage to person or property." These two policies are less likely to lead to the suppression of political dissent than the Dangerous Individuals and Organizations policy because they are more clearly focused exclusively on physical force that damages person or property. Still, while the policies in their entirety are clearly not intended to police psychological harm, the lack of definitions for "violence" and "harm" make it possible to misapply them. For example, organizing a rally to sharply (perhaps even distastefully) criticize a political party could be considered "a threatening call-to-action" that "encourages others to… join in carrying out… harmful acts," where the harmful acts are the visitation of psychological harm on the supporters of the party. Policing psychological harm will unavoidably lead to the suppression of political viewpoints based on subjective disagreement and interpretation. As noted above, content moderators cannot be expected to have the competence to fairly and objectively navigate these issues, as they are likely to involve controversial ideological viewpoints, facts that are undiscoverable or unknowable to content moderators, judgments that rely on speculations about the impact of statements, and myriad other social complexities specific to each of the geographic regions in

which Meta operates. Ensuring that these policies are interpreted by moderators to narrowly focus on physical harm would require two changes: 1. Insert a definition of "violence" as “physical force intended to damage person or property.” 2. Insert a definition of “harm” as “physical damage to person or property.”

Link to Attachment

PC-10430

Colombian police cartoon	PC-10434	-
Case number	Public comment number	Region
-	-	English
Commenter's first name	Commenter's last name	Commenter's preferred language
-		Yes
Organization		Response on behalf of organization

____

Short summary provided by the commenter

The Case Summary describes the Bank as a system that enables Meta to find duplicates "of content violating Facebook's Dangerous Individuals and Organizations Community Standard." This suggests that the Bank consists of hashes of previously identified content that violates the Dangerous Organizations policy (rather than being a tool for detecting novel examples of content that violate the policy). This case raises questions of how this cartoon came to be included in the Bank, and what Meta's procedures are for reviewing or allowing appeals of content's inclusion in the Bank. In this comment, we explain the likely technical underpinnings of Meta's media matching bank and the implications of its use for user speech.

Full Comment

The Center for Democracy & Technology welcomes the opportunity to provide comments on case 2022-004-FB-UA, regarding the use of Meta's "media matching bank" (Bank), in a decision regarding the flagging of and then takedown of a cartoon depicting police violence in Colombia. A cartoon shared by a user in Colombia was taken down (and has since been restored) for matching an image logged in Meta's Bank. The Case Summary describes the Bank as a system that enables Meta to find duplicates "of content violating Facebook's Dangerous Individuals and Organizations Community Standard." This suggests that the Bank consists of hashes of previously identified content that violates the Dangerous Organizations policy (rather than being a tool for detecting novel examples of content that violate the policy). This case raises questions of how this cartoon came to be included in the Bank, and what Meta's procedures are for reviewing or allowing appeals of content's inclusion in the Bank. In this comment, we explain the likely technical underpinnings of Meta's media matching bank, the implications of its use for user speech, and how Meta should provide more insight into its use of the Bank in its moderation system. Hashing and its limitations Meta does not provide much public information about the media matching bank referenced in the Case Summary. Meta has previously described using image matching as a tool to detect

known terrorist content, and Meta participates in the Global Internet Forum to Counter Terrorism's (GIFCT) shared hash database. The media matching bank involved in the removal of the Colombian user's post of a political cartoon is likely an in-house database of hashes of content that Meta seeks to detect, and likely to block, across its services. (It is unclear whether Meta's Bank includes the GIFCT hashes or if Meta administers the two databases separately. The cartoon about the Colombian police likely does not qualify for inclusion in the GIFCT database, which is limited to content related to organizations and individuals on the UN Sanctions list or content from live-streamed "content incidents".) Meta's media matching bank likely uses perceptual hashing to match content previously identified as violating the platform's policies with newer or recently posted content. The typical process is as follows: Identify an image (through, for example, user reporting or human review of automatically flagged content) to be detected. Run a perceptual hashing function on this image to generate an alphanumeric string of characters that are effectively a unique identifier of the image. This is the "hash". Run the same hashing algorithm on a newly uploaded image (or on an image that was previously uploaded to the site) to generate that file's hash. Compare the hashes. If they match, the new content is almost certainly identical to the previously identified content. (CDT's report, Do You See What I See? Capabilities and Limitations of Automated Multimedia Content Analysis, explains this process in more detail.) Using perceptual hashing allows the hash-matching system to identify content as a match even if it contains slight modifications such as adding a watermark or rotating the image. More significant edits, such as text superimposed on the image, are likely to produce images that generate a different enough hash that they will not be identified as a match. Essentially, hashing will not be able to detect and flag content that isn't effectively identical to the original image or video clip. Hash-matching tools cannot identify matches for content that is not already reflected in its database or bank—they cannot flag new content for review. Matching tools are also distinct from predictive machine learning tools that attempt to assess the likelihood that a post meets criteria the tool has been trained to identify. That means matching tools cannot predict whether content is likely to violate a site's policies. Matching models can only assess a piece of content based on whether it matches another in its database. Hash-matching is also context-agnostic: it merely identifies that identical content has been uploaded and does not answer questions regarding the permissibility of that content, e.g. if it was posted as part of critical commentary, news reporting, or some other exculpatory context. So, for example, hash-matching has been relatively effective as the backbone of PhotoDNA, the tool that Meta and other tech companies use to detect, remove, and report child sexual abuse material (CSAM). The publishing and sharing of CSAM is generally illegal across jurisdictions, regardless of the context in which it is shared. Identifying a matching hash of known CSAM is a strong signal that the flagged content is CSAM and must be removed. But for most other kinds of content, context is key. Content that depicts a mass atrocity or graphic violence may be used for glorification of terrorism in one context, and used by a journalist in another to report on an active conflict. A hash-matching tool in that case may flag a post by a reporter or activist that includes material that has already been hashed, like an image of a group of bodies killed in a

terrorist attack. But this re-contextualized sharing of that image may be critical to the documentation of the crisis. Recently, members of the Congressional Oversight Committee, the Committee on Foreign Affairs and relevant sub-committees in the U.S. Congress wrote to Meta CEO Mark Zuckerberg to ensure that critical documentation of human rights violations in Ukraine were not taken down by various moderation systems. If hash-matching tools are used without human review and adequate training of moderators, they can lead to the automated and widespread removal of important, newsworthy content—including content that does not violate Meta's content policies. Hash-based moderation can wrongly suppress users' speech

Link to Attachment

PC-10434