CENTER FOR DEMOCRACY & TECHNOLOGY

May 24, 2022

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,[1] and Meta participates in the Global Internet Forum to Counter Terrorism's (GIFCT) shared hash database.[2] 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".[3])

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:

[1] https://about.fb.com/news/2017/06/how-we-counter-terrorism/
[2] https://gifct.org/tech-innovation/
[3] https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf, p. 9.

1. Identify an image (through, for example, user reporting or human review of automatically flagged content) to be detected.
2. 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".
3. 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.
4. 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

The takedown of the cartoon at the center of this case raises questions about the purpose of the media matching bank and the impact of its use on users' speech.

Hash-matching tools can be useful for identifying identical content across a site, but they are a blunt tool that cannot, on their own, account for the complexity of human expression or the many contexts in which identical images or videos can be posted. And because the media matching bank is likely used at scale to assess all content posted on Meta's services, an erroneous addition to the Bank could result in widespread and persistent removal of content that does not, in fact, violate Meta's policies.

What happens when an image is added to the Bank erroneously? First, every time a user attempts to post that image, it will, at best, be flagged for review and subject to scrutiny. (It is unclear whether Meta ever automates removal of content that matches a hash in its media matching bank.) By flagging a user's post for review, that post is treated as a potential rule violation, increasing its chance of being taken down. It is not clear what information content moderators receive about the genesis of a flag, but if moderators are informed that the flagged content matches material already logged as violating the "Dangerous Individuals and Organizations" policy, this may further influence the moderator's review of the new post. Content added to the Bank is thus routinely scrutinized, and the risk of an erroneous removal decision is amplified. Meta acknowledges that its media matching technology can be error-prone: In its Q1 2022 Community Standards Enforcement Report, Meta reported that it had restored more than 413,000 pieces of content that were mistakenly flagged "due to a bug in [its] media-matching technology". Deployed at scale, these tools can increase the risk of erroneous content removal and shift the burden onto users to appeal and defend the legitimacy of their posts.

The Board should explore key questions about Meta's use of the Bank

In reviewing this case, the Board should address a number of significant questions. The first set of questions concern the facts of this particular case. For example, why was the cartoon included in the Bank in the first place? Why was there a 16-month delay between the time the content was posted and the time it was removed? What role, if any, did the Columbian government play in having the cartoon included in the Bank and/or removed? This question is especially

important given the political situation in Colombia. This cartoon depicted police violence during a period where Colombia was facing a surge of protests around the use of police force. According to Reuters, Colombian police and other government authorities requested takedowns of several pieces of content and accounts linked to dissident groups. The 2021 Freedom on the Net report notes that the Colombian government "has requested that platforms restrict content considered to incite violence."

The Board should also explore questions about the Bank more generally. For example, what criteria must be met for content to be included in the Bank? Which staff are able to add content hashes to the Bank? What processes exist to review the addition of content to the Bank? What led to the erroneous inclusion of over 400,000 pieces of content in the Bank? Is content that is found to match a hash included in the Bank ever subject to automated removal or is human review required? Is content reported to Meta by government officials eligible for inclusion in the Bank? If so, what safeguards are in place to prevent such referrals from becoming a tool for extralegal government censorship that would be significantly vulnerable to abuse?

Recommendations for transparency and due process around Meta's use of the Media Matching Bank

Meta should provide substantially more clarity about the policies and procedures around its media matching bank and how this Bank overlaps with the GIFCT shared hash database. In the wake of the terror attack in Buffalo, NY, a Meta spokesperson has told the press that multiple versions of the video of the violent incident and manifesto shared by the alleged gunman have been added to "a database to help the platform detect and remove content." The Verge also reports that "lists to external platforms hosting the content are also permanently blocked." While it is clear that Meta is using hash-matching technology to keep terrorism-related content off of its services, the details of this aspect of Meta's content moderation systems are murky. A critical first step is publicly articulating clear bright lines around the purpose and contents of the Bank.

As a comparison, the GIFCT Transparency Report describes the scope of its shared hash database, which is limited to content related to organizations on the UN Sanctions list.[4] GIFCT also describes a "severity framing taxonomy" that provides a bit of additional information about the nature of the content that may end up reflected in its hash database. GIFCT also discloses the number of pieces of content reflected in its hash database (320,000 unique images and videos) and the number of hashes (2.3 million).

At the very least, Meta should also disclose:

[4] https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf

- The purpose and scope of the Bank, including whether it is limited to content that violates the Dangerous Individuals and Organizations policy or includes other policy categories
- The relationship between Meta's Bank and the GIFCT shared hash database
- The criteria for inclusion of a hash in the Bank
- The process by which staff may add hashes to the Bank, including any verification/review steps
- The size of the Bank and the rate at which new hashes are added to it
- Whether content is added to the Bank as a result of law enforcement or other government referrals and, if so, the process for vetting such referrals
- The general procedures for use of the Bank in Meta's content moderation, including the process for determining what action to take on content that is flagged as a match and what role human review plays.

In addition to more robust disclosures, Meta should incorporate additional safeguards to ensure context is considered in the sharing of the matching material. Only the most egregious content that is a violation of Meta's policies in every context should be removed automatically following a match with a hash in the Bank. When context is relevant to the determination of a policy violation, that content should be sent to a human moderator for review. Meta should also ensure that human moderators are able to challenge the inclusion of a hash in the Bank, if the moderator believes the content has been included erroneously.

Finally, given the inevitability of error in a content moderation operating at scale, Meta should enable users not only to appeal the removal of their content or accounts, but to appeal the inclusion of their content in the Bank. Meta should disclose to the user when their content is flagged or removed due to a match with a hash in the Bank; without this information, the user will not be able to seek a full remedy for the erroneous removal of their content, and the risk that the same content will be erroneously removed when other users post it will persist.

For more information, contact Emma Llansó, Director, Free Expression Project, ellanso@cdt.org and Aliya Bhatia, Policy Analyst, Free Expression Project, abhatia@cdt.org.