Tennessee v. Meta
Case No. 23-1364-IV

DX2045

Meta

Back to Newsroom

META

Community Standards Enforcement Report, November 2019 Edition

November 13, 2019

By Guy Rosen, VP Integrity

LISTEN TO ARTICLE

Today we're publishing the fourth edition of our Community Standards Enforcement Report, detailing our work for Q2 and Q3 2019. We are now including metrics across ten policies on Facebook and metrics across four policies on Instagram.

These metrics include:

- Prevalence: how often content that violates our policies was viewed
- Content Actioned: how much content we took action on because it was found to violate our policies
- Proactive Rate: of the content we took action on, how much was detected before someone reported it to us
- Appealed Content: how much content people appealed after we took action
- Restored Content: how much content was restored after we initially took action

We also launched a new page today so people can view examples of how our Community Standards apply to different types of content and see where we draw the line.

Adding Instagram to the Report

For the first time, we are sharing data on how we are doing at enforcing our policies on Instagram. In this first report for Instagram, we are providing data on four policy areas: child nudity and child sexual exploitation; regulated goods — specifically, illicit firearm and drug sales; suicide and self-injury; and terrorist propaganda. The report does not include appeals and restores metrics for Instagram, as appeals on Instagram were only launched in Q2 of this year, but these will be included in future reports.

While we use the same proactive detection systems to find and remove harmful content across both Instagram and Facebook, the metrics may be different across the two services. There are many reasons for this, including: the differences in the apps' functionalities and how they're used – for example, Instagram doesn't have links, re-shares in feed, Pages or Groups; the differing sizes of our communities; where people in the world use one app more than another; and where we've had greater ability to use our proactive detection technology to date. When comparing metrics in order to see where progress has been made and where more improvements are needed, we encourage people to see how metrics change, quarter-over-quarter, for individual policy areas within an app.

What Else Is New in the Fourth Edition of the Report

- **Data on suicide and self-injury:** We are now detailing how we're taking action on suicide and self-injury content. This area is both sensitive and complex, and we work with experts to ensure everyone's safety is considered. We remove content that depicts or encourages suicide or self-injury, including certain graphic imagery and real-time depictions that experts tell us might lead others to engage in similar behavior. We place a sensitivity screen over content that doesn't violate our policies but that may be upsetting to some, including things like healed cuts or other non-graphic self-injury imagery in a context of recovery. We also recently strengthened our policies around self-harm and

made improvements to our technology to find and remove more violating content.

  - On Facebook, we took action on about 2 million pieces of content in Q2 2019, of which 96.1% we detected proactively, and we saw further progress in Q3 when we removed 2.5 million pieces of content, of which 97.3% we detected proactively.

  - On Instagram, we saw similar progress and removed about 835,000 pieces of content in Q2 2019, of which 77.8% we detected proactively, and we removed about 845,000 pieces of content in Q3 2019, of which 79.1% we detected proactively.

- **Expanded data on terrorist propaganda:** Our Dangerous Individuals and Organizations policy bans all terrorist organizations from having a presence on our services. To date, we have identified a wide range of groups, based on their behavior, as terrorist organizations. Previous reports only included our efforts specifically against al Qaeda, ISIS and their affiliates as we focused our measurement efforts on the groups understood to pose the broadest global threat. Now, we've expanded the report to include the actions we're taking against all terrorist organizations. While the rate at which we detect and remove content associated with Al Qaeda, ISIS and their affiliates on Facebook has remained above 99%, the rate at which we proactively detect content affiliated with any terrorist organization on Facebook is 98.5% and on Instagram is 92.2%. We will continue to invest in automated techniques to combat terrorist content and iterate on our tactics because we know bad actors will continue to change theirs.

- **Estimating prevalence for suicide and self-injury and regulated goods:** In this report, we are adding prevalence metrics for content that violates our suicide and self-injury and regulated goods (illicit sales of firearms and drugs) policies for the first time. Because we care most about how often people may see content that violates our policies, we measure prevalence, or the frequency at which people may see this content on our services. For the policy areas addressing the most severe safety concerns — child nudity and sexual exploitation of children, regulated goods, suicide and self-injury, and terrorist propaganda — the likelihood that people view content that violates these policies is very low, and we remove much of it before people see it. As a result, when we sample views of content in order to measure prevalence for these

policy areas, many times we do not find enough, or sometimes any, violating samples to reliably estimate a metric. Instead, we can estimate an upper limit of how often someone would see content that violates these policies. In Q3 2019, this upper limit was 0.04%. Meaning that for each of these policies, out of every 10,000 views on Facebook or Instagram in Q3 2019, we estimate that no more than 4 of those views contained content that violated that policy.

- It's also important to note that when the prevalence is so low that we can only provide upper limits, this limit may change by a few hundredths of a percentage point between reporting periods, but changes that small do not mean there is a real difference in the prevalence of this content on the platform.

Progress to Help Keep People Safe

Across the most harmful types of content we work to combat, we've continued to strengthen our efforts to enforce our policies and bring greater transparency to our work. In addition to suicide and self-injury content and terrorist propaganda, the metrics for child nudity and sexual exploitation of children, as well as regulated goods, demonstrate this progress. The investments we've made in AI over the last five years continue to be a key factor in tackling these issues. In fact, recent advancements in this technology have helped with rate of detection and removal of violating content.

For **child nudity and sexual exploitation of children**, we made improvements to our processes for adding violations to our internal database in order to detect and remove additional instances of the same content shared on both Facebook and Instagram, enabling us to identify and remove more violating content.

On Facebook:

- In Q3 2019, we removed about 11.6 million pieces of content, up from Q1 2019 when we removed about 5.8 million. Over the last four quarters, we proactively detected over 99% of the content we remove for violating this policy.

While we are including data for Instagram for the first time, we have made progress increasing content actioned and the proactive rate in this area within the last two quarters:

- In Q2 2019, we removed about 512,000 pieces of content, of which 92.5% we detected proactively.
- In Q3, we saw greater progress and removed 754,000 pieces of content, of which 94.6% we detected proactively.

For **our regulated goods policy** prohibiting illicit firearm and drug sales, continued investments in our proactive detection systems and advancements in our enforcement techniques have allowed us to build on the progress from the last report.

On Facebook:

- In Q3 2019, we removed about 4.4 million pieces of drug sale content, of which 97.6% we detected proactively — an increase from Q1 2019 when we removed about 841,000 pieces of drug sale content, of which 84.4% we detected proactively.
- Also in Q3 2019, we removed about 2.3 million pieces of firearm sales content, of which 93.8% we detected proactively — an increase from Q1 2019 when we removed about 609,000 pieces of firearm sale content, of which 69.9% we detected proactively.

On Instagram:

- In Q3 2019, we removed about 1.5 million pieces of drug sale content, of which 95.3% we detected proactively.
- In Q3 2019, we removed about 58,600 pieces of firearm sales content, of which 91.3% we detected proactively.

New Tactics in Combating Hate Speech

Over the last two years, we've invested in proactive detection of hate speech so that we can detect this harmful content before people report it to us and sometimes before anyone sees it. Our detection techniques include text and image matching, which means we're identifying images and identical strings of text that have already been removed as hate speech, and machine-learning classifiers that look at things like language, as well as the reactions and comments to a post, to assess how closely it matches common phrases, patterns and attacks that we've seen previously in content that violates our policies against hate.

Initially, we've used these systems to proactively detect potential hate speech violations and send them to our content review teams since people can better assess context where AI cannot. Starting in Q2 2019, thanks to continued progress in our systems' abilities to correctly detect violations, we began removing some posts automatically, but only when content is either identical or near-identical to text or images previously removed by our content review team as violating our policy, or where content very closely matches common attacks that violate our policy. We only do this in select instances, and it has only been possible because our automated systems have been trained on hundreds of thousands, if not millions, of different examples of violating content and common attacks. In all other cases when our systems proactively detect potential hate speech, the content is still sent to our review teams to make a final determination. With these evolutions in our detection systems, our proactive rate has climbed to 80%, from 68% in our last report, and we've increased the volume of content we find and remove for violating our hate speech policy.

While we are pleased with this progress, these technologies are not perfect and we know that mistakes can still happen. That's why we continue to invest in systems that enable us to improve our accuracy in removing content that violates our policies while safeguarding content that discusses or condemns hate speech. Similar to how we review decisions made by our content review team in order to monitor the accuracy of our decisions, our teams routinely review removals by our automated systems to make sure we are enforcing our policies correctly. We also continue to review content again when people appeal and tell us we made a mistake in removing their post.

Updating our Metrics

Since our last report, we have improved the ways we measure how much content we take action on after identifying an issue in our accounting this summer. In this report, we are updating metrics we previously shared for content actioned, proactive rate, content appealed and content restored for the periods Q3 2018 through Q1 2019.

During those quarters, the issue with our accounting processes did not impact how we enforced our policies or how we informed people about those actions; it only impacted how we counted the actions we took. For example, if we find that a post containing one photo violates our policies, we want our metric to reflect that we took action on one piece of content — not two separate actions for removing the photo and the post. However, in July 2019, we found that the systems logging and counting these actions did not correctly log the actions taken. This was largely due to needing to count multiple actions that take place within a few milliseconds and not miss, or overstate, any of the individual actions taken.

We'll continue to refine the processes we use to measure our actions and build a robust system to ensure the metrics we provide are accurate. We share more details about these processes here.

Downloads

Community Standards Enforcement Report Highlights

Press Call Audio Recording

Press Call Transcript

Categories:
Integrity and Security, Meta, Public Policy, Safety and Expression

Tags:
Community Standards and Enforcement, Community Standards Enforcement Report, Facebook app, Instagram

DOWNLOAD ALL IMAGES

Share this article

RECENT NEWS

AI Breakthroughs, Our Most Advanced Glasses, and More: Meta's 2025 Highlights

One Year In: Meta's Richland Parish Data Center Supports Louisiana Economy With $875M In Contracts

Advancing Water Stewardship in Our Data Center Communities

Watch Reels on the Big Screen With Instagram for TV

RELATED NEWS

META

More Speech and Fewer Mistakes

January 7, 2025

META

What We Saw on Our Platforms During 2024's Global Elections

December 3, 2024

Follow Meta Newsroom |

Press Resources | PRESS@META.COM MEDIA GALLERY

Meta

Meta Store

Community

Our actions

About us

Site terms and policies

App support

© 2026 Meta Community Standards Data Policy Terms Cookie policy

United States

(English)