FACEBOOK

Community Standards Enforcement Report

HIGHLIGHTS - MAY 2020

We want Facebook and Instagram to be places where people can express themselves and have a voice. To help us achieve this, we have policies in place that promote expression while enabling our users to connect and share safely.

The Community Standards Enforcement Report measures how we are doing at enforcing our policies.

This report does not reflect the full impact of the content review changes we made during the COVID-19 pandemic. We anticipate we'll see the impact of those changes in our next report and beyond, and will continue to be transparent about it.

What's New

1

INTRODUCED DATA FOR ORGANIZED HATE

For the first time, we're sharing enforcement data on organized hate, under the dangerous organizations category.

2

REPORTED ADDITIONAL INSTAGRAM POLICIES

We've added 4 new policy areas from Instagram: adult nudity & sexual activity, bullying & harassment, hate speech, violent & graphic content.

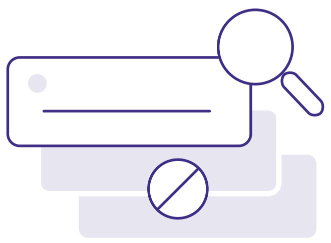
3

ADDED MORE INSTAGRAM METRICS

We're sharing appealed content and restored content data for Instagram for the first time.

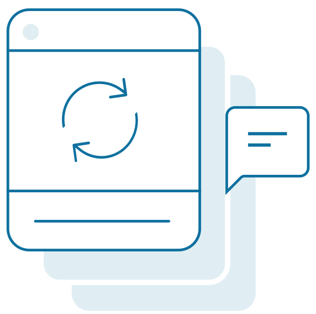
What We've Done

Continued to Invest in Our Technology to Keep People Safe



IMPROVED SUICIDE & SELF-INJURY DETECTION TECHNOLOGY FOR INSTAGRAM

Proactive rate increased more than 12 points from our last report, from 77.5% in Q2 2019, to 89.7% in Q1 2020, due to new detection technology to take down more violating content. Content actioned increased by 40% as a result.



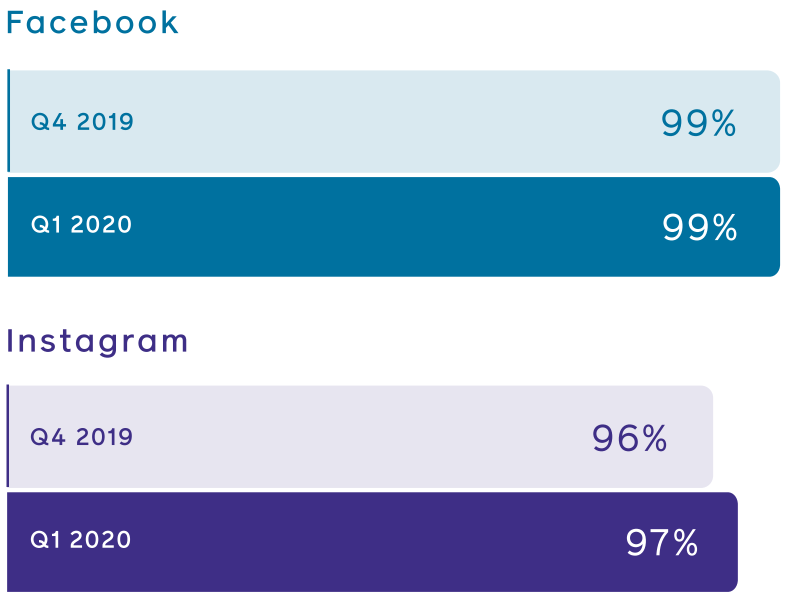
EXPANDED HATE SPEECH ENFORCEMENT ON FACEBOOK

Proactive rate increased from our last report, from 70.9% in Q2 2019 to 88.8% Q1 2020, because of improved detection technology, as well as expanding to new languages to keep reducing bad experiences globally. In addition, content actioned increased by 70% from Q4 2019 to Q1 2020.

ENHANCED DETECTION OF CHILD NUDITY & SEXUAL EXPLOITATION ON FACEBOOK AND INSTAGRAM

Since our last report, proactive rate on Facebook has remained consistently above 99% as we improved our ability to detect and remove old violating content. On Instagram, proactive rate increased from our last report, from over 96% in Q4 2019 to over 97% Q1 2020, due to improvements to our automated systems.

PROACTIVE RATE



What We're Working On

BULLYING & HARASSMENT

Instagram remains committed to curbing bullying and harassment, and protecting young people using the service. We made progress in our work combatting online bullying and harassment by introducing several new features* to help people control their experiences and limit unwanted interactions. We are sharing enforcement data for bullying and harassment on Instagram for the first time in this report, including that we took action on 1.5 million pieces of content in each of Q4 2019 and Q1 2020. We have not yet been able to provide prevalence data for bullying and harassment on either Facebook or Instagram, an important part of understanding progress in this area. Because identifying online bullying can be highly dependent on language and context, and often reflects the nature of personal relationships, prevalence in this policy area is particularly difficult to measure. We will continue working on finding the best way to estimate this metric.

SUICIDE & SELF-INJURY

We continue to develop comprehensive policies to handle sensitive cases and create an environment to allow people to heal and connect, as well as invest in new technology to detect and act on self-harm content faster in potentially life-saving situations.

*https://about.instagram.com/blog/announcements/stand-up-against-bullying-with-restrict

METRICS

PREVALENCE = estimate of how often content that violates our policies is seen by a user

CONTENT ACTIONED = number of pieces of content (such as posts, photos, videos or comments) or accounts we take action on for going against our policies

PROACTIVE RATE = percentage of all content or accounts acted on that we found and flagged before users reported them to us

APPEALED CONTENT = number of pieces of content that people appealed after we took action on them for going against our policies

RESTORED CONTENT = number of pieces of content we restored after we originally took action on them