Tennessee v. Meta
Case No. 23-1364-IV
PX-04245

October 25, 2024
EXHIBIT
Meta-Jin-44

.st0{fill:#1c2b33}.st1{clip-path:url(https://about.fb.com/#SVGID_1_)}.st2{clip-path:url(https://about.fb.com/#SVGID_3_)}.st3{fill:url(https://about.fb.com/#SVGID_

Back to Newsroom

Meta

Community Standards Enforcement Report, Third Quarter 2021

November 9, 2021

By Guy Rosen, VP of Integrity



- **Prevalence of hate speech on Facebook continued to decrease for the fourth quarter in a row. We've cut prevalence by more than half within the last year through improvements in our technology and changes we made to reduce problematic content in News Feed.**
- **For the first time, we're sharing prevalence metrics for bullying and harassment on both Facebook and Instagram, prevalence metrics for hate speech on Instagram and all metrics for violence and incitement.**
- **To validate that our metrics are measured and reported correctly, we will undergo an audit by EY covering the fourth quarter of this year, and results will be released in the Spring of 2022.**

Today we're publishing our Community Standards Enforcement Report for the third quarter of 2021. This report provides metrics on how we enforced our policies from July 2021 through September 2021. With these additional metrics we now report on 14 policy areas on Facebook and 12 on Instagram.

We're also sharing:

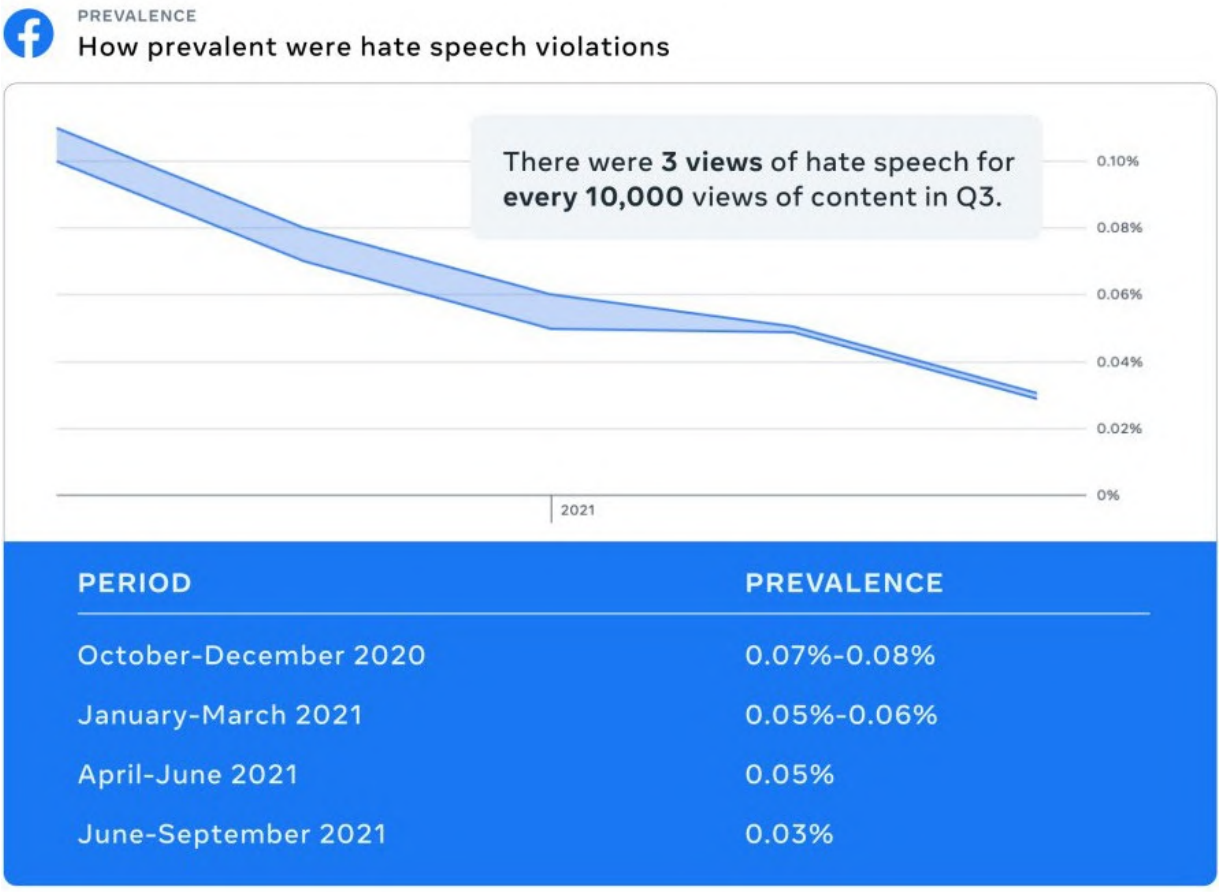
- The Widely Viewed Content Report and companion guide for the third quarter of 2021
- The Oversight Board Quarterly Report for the second and third quarters of 2021
- Our other Transparency Reports for the first half of 2021:
    - Content Restrictions
    - Internet Disruptions
    - Government Requests for User Data
    - Intellectual Property

In July, we shared our first Oversight Board quarterly update covering the first quarter of 2021. In this update, we're reporting on the details of cases we referred to the board and updates on our responses to the board about the recommendations and decisions they made. Moving forward, we will include these updates along with the Community Standards Enforcement Report.

All of these reports are available in the Transparency Center.

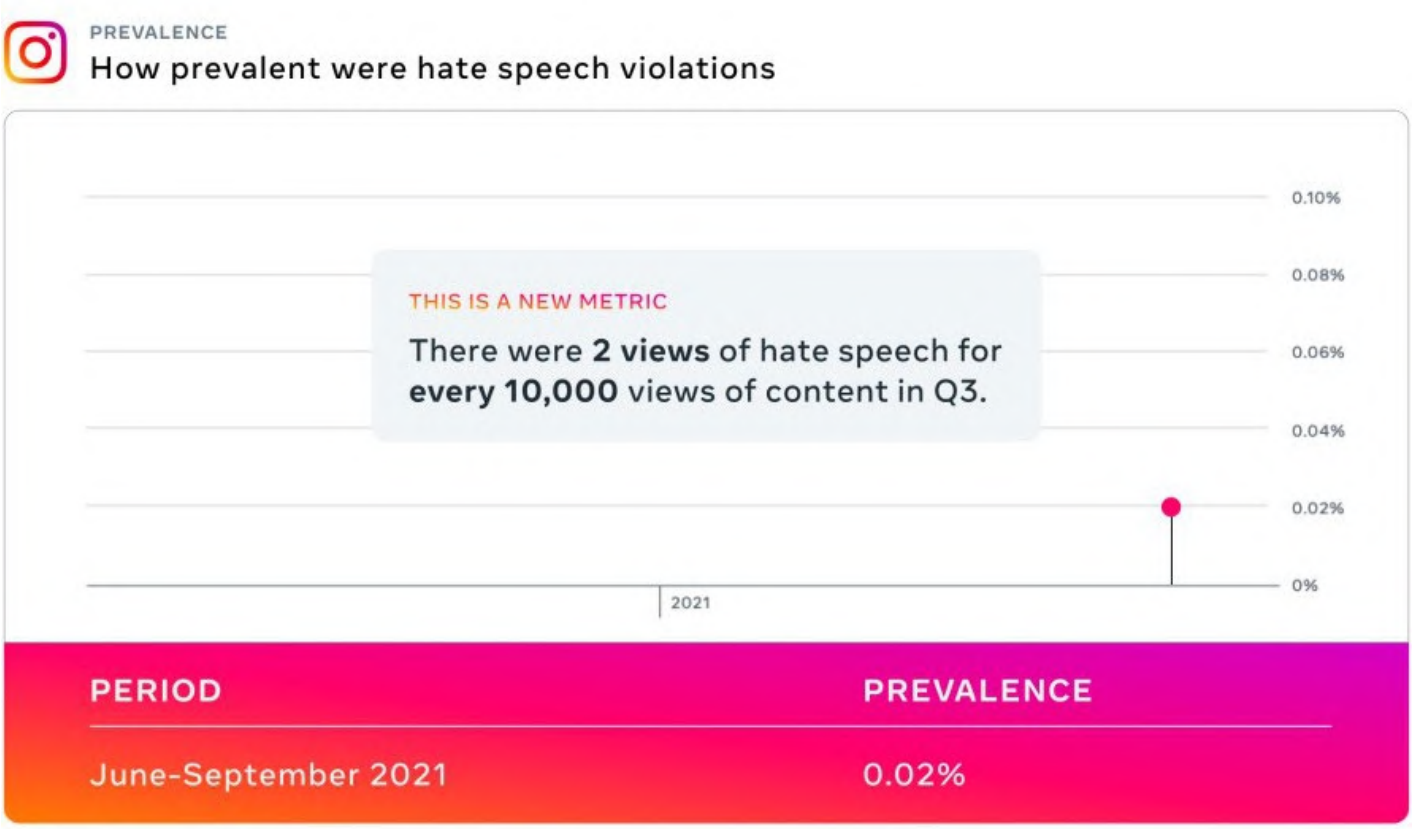
Highlights from the Report

Prevalence of hate speech on Facebook has continued to decrease for the fourth quarter in a row. In Q3, it was 0.03% or 3 views of hate speech per 10,000 views of content, down from 0.05%, or 5 views of hate speech per 10,000 views of content in Q2. We continue to see a reduction in hate speech due to our improvements in our technology and ranking changes that reduce problematic content in News Feed, including through improved personalization.



PERIOD	PREVALENCE
October-December 2020	0.07%-0.08%
January-March 2021	0.05%-0.06%
April-June 2021	0.05%
June-September 2021	0.03%

Hate speech prevalence on Instagram was 0.02% for Q3, and this is our first time reporting this number.



In Q3, the prevalence of violence and incitement on Facebook was 0.04-0.05%, or between 4 and 5 views per 10,000 views of content, and it was 0.02%, or 2 views per 10,000 views of content, on Instagram. We removed 13.6 million pieces of content on Facebook for violating our violence and incitement policy, and we proactively detected 96.7% of this content before anyone reported it to us. On Instagram, we removed 3.3 million pieces of this content with a proactive detection rate of 96.4%. There is a range of content that we might remove under our violence and incitement policy where someone may advocate for violence or has made a statement of intent to commit violence. Due to the potentially harmful nature of content attempting to incite violence, we over-index on safety and remove such content even if it is unclear whether the content is in jest. This could range from something serious such as instructions on how to use weapons to cause injury to a joke where one friend says to another “I’ll kill you!”. In instances where necessary, we also work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety.

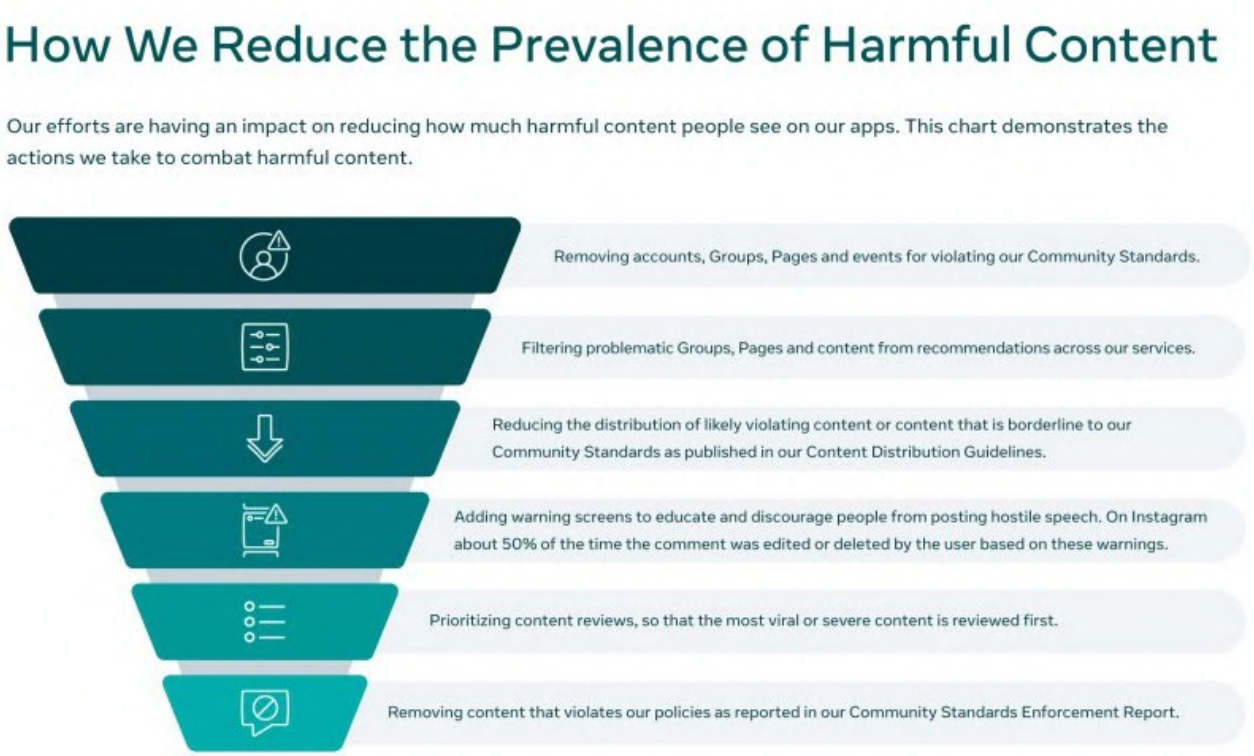
In Q3, the prevalence of bullying and harassment content was 0.14-0.15% or between 14 and 15 views of bullying and harassment content per 10,000 views of content on Facebook, and 0.05-0.06% or between 5 and 6 views per 10,000 views of content on Instagram. We

removed 9.2 million pieces of bullying and harassment content on Facebook, with a proactive rate of 59.4%. And we removed 7.8 million pieces of bullying and harassment content on Instagram with a proactive rate of 83.2%. Bullying and harassment is a unique challenge and one of the most complex issues to address because context is critical. Read more about our approach and efforts to lower the prevalence of bullying and harassment.

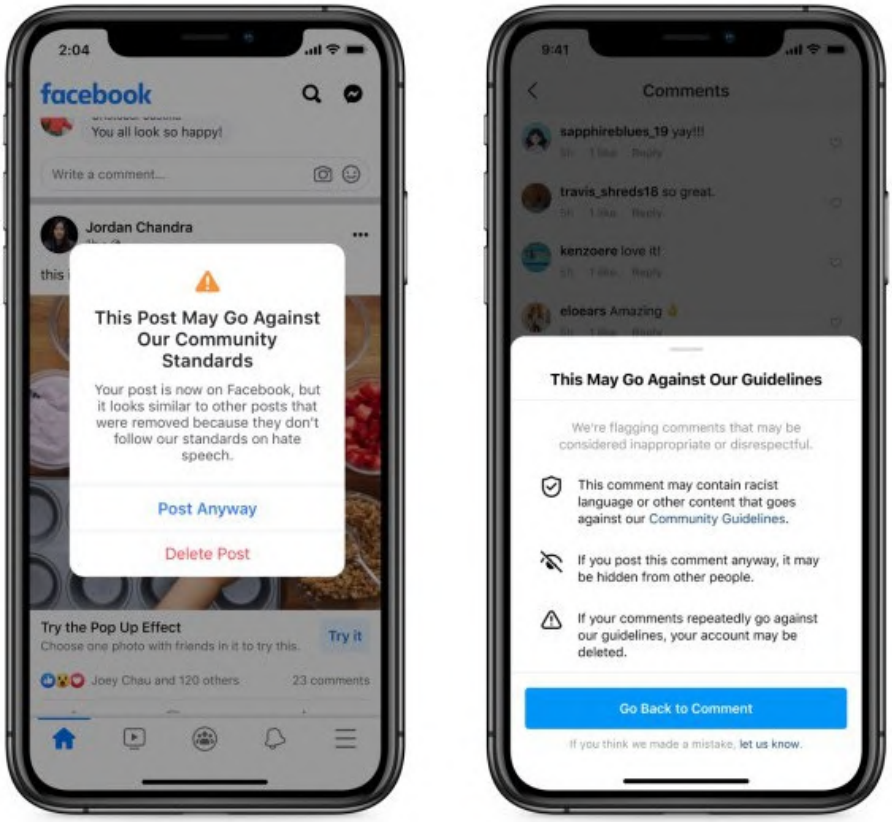
How We Reduce Harmful Content People See on Our Apps

We scale our enforcement to review millions of pieces of content across the world every day and use our technology to help detect and prioritize content that needs review. We have global review teams that review content in over 70 languages and AI technology for hate speech in over 50 languages and we continue to build technologies like RIO, WPIE and XLM-R that can help us identify harmful content faster, across languages and content type (i.e. text, image, etc.). These efforts and our continued focus on AI research help our technology scale quickly to keep our platforms safe. For countries that are experiencing or at risk for conflict, we have made multi-year investments to build teams that work with local communities, develop policies, improve our technologies, and respond to real-world developments.

We reduce prevalence of violating content in a number of ways, including improvements in detection and enforcement and reducing problematic content in News Feed. These tactics have enabled us to cut hate speech prevalence by more than half on Facebook in the past year alone, and we're using these same tactics across policy areas like violence and incitement and bullying and harassment. To better address hate speech, bullying and harassment and violence and incitement — all of which require understanding of language, nuance and cultural norms — we deployed a new cross-problem AI system to consolidate learnings for all three to better address each violation area.



We're also using warning screens to educate and discourage people from posting something that may include hostile speech such as bullying and harassment violating our Community Standards. The screens appear after someone has typed a post or comment explaining that the content may violate our rules and may be hidden or distribution reduced. Repeatedly posting this content could result in an account being disabled or deleted.

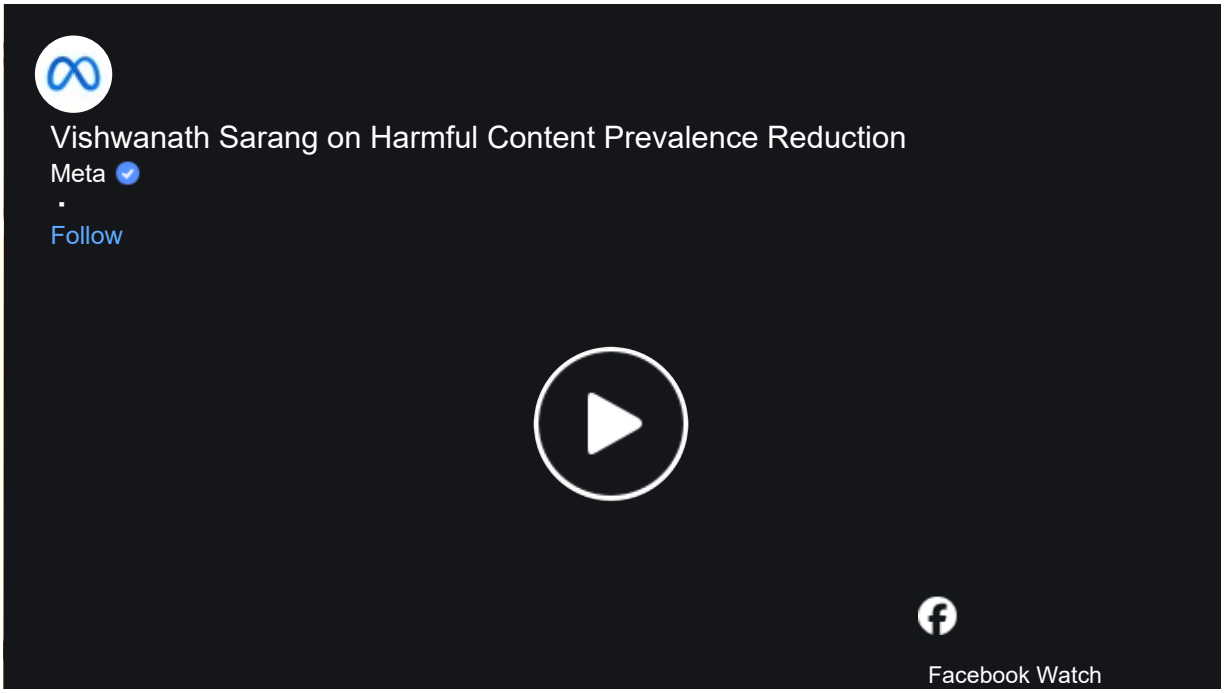


Looking Ahead

Abuse of our products isn't static — and neither is the way we approach our integrity work. We're continuing to evolve how we approach integrity, embedding integrity teams with product teams across Facebook, Instagram, Messenger, WhatsApp and teams that are going to build the metaverse in the years to come. Product teams tackle integrity issues as part of building and launching new products, and that includes building the metaverse.

We also remain committed to research and have just launched a Facebook Open Research and Transparency (FORT) early access program for an API designed for academic researchers, to give them more access to data on our platforms in a secure way.

We know we're never going to be perfect in catching every piece of harmful content. But we're always working to improve, share more meaningful data and continue to ground our decisions in research.



Downloads

November 9 Press Call Audio

Press Call Transcript

Categories: Meta, Public Policy, Safety and Expression

Tags:
Combating Hate Speech, Community Standards and Enforcement, Community Standards Enforcement Report

RELATED NEWS

Meta

Labeling AI-Generated Images on Facebook,

Instagram and Threads

Nick Clegg offers a new approach to identifying and labeling AI-generated content.

February 6, 2024

Virtual reality

Smart glasses

About us

Our community

Our actions

Support

.st0{fill:#1c2b33}.st1{clip-path:url(https://about.fb.com/#SVGID_1_)}.st2{clip-path:url(https://about.fb.com/#SVGID_3_)}.st3{fill:url(https://about.fb.com/#SVGID_6

Follow Us

© 2024 Meta
Community Standards
Data Policy
Terms
Cookie policy