META

Moderator: Sabrina Siddiqui
August 25, 2022
11:00 a.m. ET

Operator: Hello and welcome to the Quarterly Integrity Press Call. There will be prepared remarks and a Q&A to follow. To ask a question after the prepared remarks conclude, please press the "1" "4" on your telephone. As a reminder this call is being recorded. Now I'd like to turn the call over Sabrina Siddiqui, who will kick it off. Please go ahead.

Sabrina Siddiqui: Hi, everyone. Thank you for joining us. You should have received embargoed materials ahead of this call with our community standards enforcement report, widely viewed content report, and the oversight board quarter report.

To kick off our call, today you will hear from Vice President of Content Policy, Monika Bickert, and Director of Product for our Integrity team, Maxime Prades. We will then open up the call for questions. We are on the record and this call is embargoed until 9:00 a.m. pacific/12:00 p.m. eastern. With that, I'll hand it over to Monika.

Monika Bickert: Thanks, Sabrina, and hello, everyone. Thanks for joining us today. I'm Monika Bickert, our VP of Content Policy. Today we'll share details on our progress at keeping violating content low, updates on the oversight board and their impact on Meta's transparency, and more about the ongoing work to refine how we enforce our policies. As you all know, this work doesn't stop and we'll keep sharing these updates each quarter.

As part of our integrity efforts, we released our quarterly adversarial threat report earlier this month which provides an in-depth qualitative view into different types of adversarial threats that we tackle globally.

In addition and as a part of our continuous enforcement against Proud Boys, a hate group we banned in 2018, we recently uncovered and removed a network of about 480 Facebook and Instagram accounts, pages, and groups.

In total in 2022 we took down around 750 assets from the Proud Boys as part of the strategic network disruptions. We will continue to refine our targeted approach to adversarial threats as well as our scaled enforcement.

We have several milestones to share from the Oversight Board quarterly update. The Q2 update shows that Meta has responded to every single recommendation of the Oversight Board publicly and that we're committed to implement or further explore the feasibility of implementing 73 percent of the recommendations they've made to date.

In addition to the binding decisions that they've issued, all of which complied with, that's the content should be up or down decisions, the board has also issued 119 non-binding recommendations. We responded to all of those recommendations, but there were a few highlights that I would like to call out from the report.

First, we're expanding the board's scope. The board will soon have the ability to issue a new type of binding judgment on cases and that is to apply a warning screen. This builds on the board's existing ability to apply binding judgments for us to take down or leave up pieces of content.

Second, we're releasing data on a number of newsworthy allowances we've granted. And that means when we've allowed violating content to stay on our site because of its public interest value. This data along with examples of the allowances and additional details on the approach to newsworthy content can now be found in our updated transparency center.

And third, we're publishing our Crisis Policy Protocol as a result of the recommendation from the board. This language explains how we respond to crises around the world, including how we assess situations real-time and the measures that we might apply to respond to those situations with appropriate policy applications and refinements.

Finally, I'd also like to point you to our newsroom posts, where we share details on our ongoing efforts to combat misinformation on our platforms. And with that, I'll turn it over to Maxime.

Maxime Prades: Thanks, Monika. Hi everyone. My name is Maxime Prades and I lead the integrity and safety product teams here at Meta. I'd like to start by first turning to the Community Standards Enforcement Report. This report is an update on our progress at keeping people safe while also empowering their ability to express themselves.

Every quarter we published a prevalence of violations and the actions we take against them. Today, I want to highlight a couple of key takeaways from this quarter's report. Thanks to improvements in our A.I. technology, the

proactive detection rate of bullying and harassment content increased on Facebook from 67 percent in the first quarter to 76.8 percent in the second quarter.

On Instagram, the same detection rate increased from 83.8 percent in the first quarter to 87.4 percent in the second quarter. On Facebook, this past quarter, hate speech prevalence remained at 0.02 percent. In other words, that's 2 per 10,000 views. And we took action on 13.5 million pieces of content.

Similarly, on Instagram, hate speech prevalence was between 0.01 and 0.02 percent. That's 1 to 2 per 10,000 views. And we took actions on 3.8 million pieces of content. Next, I'd like to talk about one change in calculating appeals. As part of our work to constantly improve the metrics that we share, we have updated the appeals counting methodology.

Prior to this quarter's change, the way we reported appeals was by accounting for when users requested a review after we took content down. At the beginning of the COVID-19 pandemic, due to reduced review capacity worldwide, we were not always in the position to offer people the option to appeal.

However, we still gave people the option to tell us that they disagreed with our decision, which helped us review many of these instances and even restore content when it was appropriate.

Over time, we've improved the appeal experience to get better signals from people when we make mistakes. And we now account for this experience in our appeals metric. Today, we view capacity as recovered and people who select this option will have their account automatically submitted for additional review, except in some cases like severe violations or (push back).

And we continue to leverage that signal to gather important feedback that maybe used to improve our policies and automated detection system.

Lastly, I'd like to talk about how we're looking into ways that we can improve our proactive enforcement. We know we don't always get it right and we're always refining our policies and enforcement.

I wanted to highlight two major improvements in this report. First, in Q2 we found that using warning screens to discourage posting hate speech or bullying and harassment prevented some of this content, which could have validated our community standards from being posted in the first place.

Secondly, we're expanding a test called Flagged by Facebook, which allows group admin to better shape their group culture and take context into account. Admins can keep some content in their groups that might otherwise be flagged

for bullying and harassment. As with all of our integrity work, the views on our platforms is never static and neither are we in how we address it. And with that, I'll turn it back over to Sabrina.

Sabrina Siddiqui: Thank you. We'll open it up now for questions.

Operator: Thank you. If you would like to register a question please press the "1" "4" on your telephone. You will hear a three toned prompt to acknowledge your request. If your question has been answered and you would like to withdraw your registration, please press the "1" followed by the "3." One moment please for the first question. Our first question comes from Queenie Wong with CNET. Please proceed.

Queenie Wong: Hi, so there's a lot of interest in whether or not Facebook -- or how Facebook will sort of handle the Trump suspension after January 20223? And I was wondering if you could comment a little and I was wondering if you could comment a little bit about how you're looking at that suspension and sort of what experts you're looking at to assess the risks of public safety?

Monika Bicker: Hi, this is Monika. We put out a statement on this at the time that we suspended then President Trump's accounts. And the -- there's no new information. We announced then that the account would be suspended for two years and that started on January 7th of 2021. And as we said then, at the end of this period we will look to experts to assess whether the risk to public safety has receded.

And that is going to include looking at external factors, including incidents of violence and restrictions on peaceful assembly and other traditional markers of civil unrest. If at that time we determine there's still a serious risk to public safety that we will extend the restriction for a period of time and then we'll continue to evaluate.

Operator: Our next question comes from Beatrice Peterson with ABC News. Please proceed.

Beatrice Peterson: Hi. So my question is about the Proud Boys. I was wondering if I can get some more information about whether they were using alias pages or the pages that you took down were just flat out publicly calling themselves Proud Boys?

Monika Bickert: What I can share with you now is that this was a network of pages and groups, Instagram and Facebook accounts, so some of these were accounts, not just pages or groups. But all of them were linked to this network that was the Proud Boys.

This last strategic disruption is just part of our overall effort this year. We've removed about 750 assets belonging to the Proud Boys.

Operator: Out next question comes from Katie Paul with Reuters. Please proceed.

Katie Paul: Hi, there. Thanks, as always, for doing this. I am wondering if you could talk about how you're planning to do disclosure accounting around Reels as it becomes a bigger part of what appears on Facebook and Instagram? Is that going to show up in the post section or do you have any new plans around kind of a (time) for that content.

And particularly in terms of what gets recommended to people because that has a different kind of quality to it that's difficult to track from the outside of Messenger. And then kind of – I suppose adjacent to that question is about so many of the posts that reach the top 20 are still ones that ultimately get taken down for inauthentic behavior in a lot of cases.

In one case I see – I'm wondering if you can talk about any controls that exist in the system or any way you might account for that publicly in a report like this to maybe mitigate against that so it doesn't reach quite so many people before it is taken down? Thanks, so much.

Sabrina Siddiqui: Thank you for that question. We'll take that first part over to Maxime and then we will also have, after Maxime gets to you, David Agranovich, our Director of Threat Disruption will answer that second part. Maxime?

Maxime Prades: Yes, thank you so much for that question. I think the first thing to know, is when we remove Reels content they are included in our content action numbers. And then as we make changes to our products we constantly assess how they should reflect in our transparency reports, and are constantly keeping a metrics of data to reflect the prevalence of violations and removals. I'll kick over to David.

David Agranovich: Thank you so much for the question. I think a few ways of thinking about this, one is spam, right, predates the internet but it's an incredibly adversarial space. One of the goals of publishing a widely viewed content report was both to bring transparency to the type of content that's getting virality on the platform.

But it was also to hold ourselves accountable for our own integrity interventions, our ability to suspect and remove this type of activity. And kind of create a virtuous cycle where as we're discovering these things we can tell people about it. If other folks are seeing suspicious activity from the links that we're posting on the widely viewed content report, it can enable our further investigative work.

So a few of the interventions here, one, when we're seeing links like this that aren't getting caught by our existing automated detection or scaled systems, it helps us build better detection and identify where those gaps were.

A good example of that is actually the portion of the widely viewed content report where we talked a bit about one of the examples, one of the links that we blocked for being involved in essentially merchandise spam, right, just trying to post links to t-shirt websites all over the platform.

When we initially blocked that link, it was in response to a flag from the Integrity Institute, which had identified a spamming behavior from the link.

When we started digging into it further, in part because it had been flagged in the report, we were able to identify a much larger set of accounts and assets on the platform that the spammers were using.

We were able to identify more about how they were trying to evade that detection and close up some of the gaps in our own detections, and we were able to go beyond what other people had been able to identify and actually link that activity to a specific company engaged in a spamming activity.

So in short, the report is there to drive us to better and more effective enforcement, and so improving our automated systems using our investigative capabilities to understand our gaps, and then try to close those gaps through a (inaudible) process.

Sabrina Siddiqui: Thanks, everyone. As a reminder, the embargo lifts at 9:00 AM Pacific, 12:00 PM Eastern. If you have any questions, please feel free to reach out to the (press@) so that we can get some answered for you. Thank you all for joining the press call.

Operator: This concludes the integrity press call. We thank you for joining. You may now disconnect your line. Have a great day, everybody.

END