Facebook reports progress on curbing hateful, abusive content

November 19 2020

Facebook says its improved automated tools have helped reduce the prevalence of hateful content on the massive social network

Facebook said Thursday it has made progress in curbing hate speech and other abusive content on its platform with improved automated tools

complementing its human reviewers.

Releasing its transparency report for the third quarter, the social media giant said it took action against more than 70 million pieces of content on its core social network and Instagram which included hate speech, bullying or harassment, graphic violence, child sexual exploitation and suicide or self-injury.

Facebook for the first time released a statistic on "prevalence" of hate speech, amounting to 0.10 to 0.11 percent of viewed posts on the platform.

"You can think of prevalence as an air quality test," said Guy Rosen, vice president of integrity at Facebook, in a conference call with journalists.

Rosen said Facebook chose this metric as a gauge of the health of the platform because "a small amount of content can go viral and get a lot of distribution."

The release comes with Facebook under rising pressure from governments and activists to crack down on hateful and abusive content while keeping its platform open to divergent viewpoints.

Facebook said it took action on some 22 million pieces of hate speech content in the July-September period, up from 15 million in the prior quarter. It said it increased enforcement for other kinds of violations as well.

Rosen said automated systems using artificial intelligence have become more effective and now detect some 95 percent of hate speech.

But he noted that human reviewers are still needed for finding more subtle forms of abusive content which may not be detected by

computerized systems.

The news comes a day after some 200 Facebook contract moderators signed a petition calling for better safety conditions as Facebook begins to call workers back to the office amid the global pandemic.

Rosen said that "the majority of our review workforce is still working from home" but that Facebook is not asking these people to review the most offending content.

"This is really sensitive content. This is not something you want people reviewing from home with their family around," he said.

Rosen said a major effort in content moderation this year involved misinformation about the US election and the COVID-19 pandemic.

He said Facebook removed some 265,000 posts between March 1 and the November 3 election for violating voter interference policies and displayed warnings on 180 million posts whose claims were debunked by independent fact-checkers.

Facebook also took down some 12 million posts between March and October "containing misinformation that may lead to imminent physical harm" including on fake coronavirus cures or treatments, Rosen said, and displayed warnings on another 160 million pieces of content.