Facebook's race-blind practices around hate speech came at the expense of Black users, new documents show

Researchers proposed a fix to the biased algorithm, but one internal document predicted pushback from 'conservative partners'

By Elizabeth Dwoskin, Nitasha Tiku and Craig Timberg

November 21, 2021 at 8:00 a.m. EST

Last year, researchers at Facebook showed executives an example of the kind of hate speech circulating on the social network: an actual post featuring an image of four female Democratic lawmakers known collectively as "The Squad."

The poster, whose name was scrubbed out for privacy, referred to the women, two of whom are Muslim, as "swami rag heads." A comment from another person used even more vulgar language, referring to the four women of color as "black c---s," according to internal company documents exclusively obtained by The Washington Post.

The post represented the "worst of the worst" language on Facebook — the majority of it directed at minority groups, according to a two-year effort by a large team working across the company, the document said. The researchers urged executives to adopt an aggressive overhaul of its software system that would primarily remove only those hateful posts before any Facebook users could see them.

But Facebook's leaders balked at the plan. According to two people familiar with the internal debate, top executives including Vice President for Global Public Policy Joel Kaplan feared the new system would tilt the scales by protecting some vulnerable groups over others. A policy executive prepared a document for Kaplan that raised the potential for backlash from "conservative partners," according to the document. The people spoke to The Post on the condition of anonymity to discuss sensitive internal matters.

Facebook researchers used a post containing hate speech targeting members of "The Squad" — Democratic Reps. Rashida Tlaib (Mich.), Ilhan Omar (Minn.), Alexandria Ocasio-Cortez (N.Y.) and Ayanna Pressley (Mass.) — as an example to survey a group on what they perceive as harmful language. (Tom Williams/CQ Roll Call via Getty Images)

The previously unreported debate is an example of how Facebook's decisions in the name of being neutral and race-blind in fact come at the expense of minorities and particularly people of color. Far from protecting Black and other minority users, Facebook executives wound up instituting half-

measures after the “worst of the worst” project that left minorities more likely to encounter derogatory and racist language on the site, the people said.

“Even though [Facebook executives] don’t have any animus toward people of color, their actions are on the side of racists,” said Tatenda Musapatike, a former Facebook manager working on political ads and CEO of the Voter Formation Project, a nonpartisan, nonprofit organization that uses digital communication to increase participation in local state and national elections. “You are saying that the health and safety of women of color on the platform is not as important as pleasing your rich White man friends.”

The Black audience on Facebook is in decline, according to data from a study Facebook conducted earlier this year that was revealed in documents obtained by whistleblower Frances Haugen. According to the February report, the number of Black monthly users fell 2.7 percent in one month to 17.3 million adults. It also shows that usage by Black people peaked in September 2020. Haugen’s legal counsel provided redacted versions of the documents to Congress, which were viewed by a consortium of news organizations including The Post.

Civil rights groups have long claimed that Facebook’s algorithms and policies had a disproportionately negative impact on minorities, and particularly Black users. The “worst of the worst” documents show that those allegations were largely true in the case of which hate speech remained online.

But Facebook didn’t disclose its findings to civil rights leaders. Even the independent civil rights auditors Facebook hired in 2018 to conduct a major study of racial issues on its platform say they were not informed of the details of research that the company’s algorithms disproportionately harmed minorities. Laura Murphy, president of Laura Murphy and Associates, who led the civil rights audit process, said Facebook told her that “the company does not capture data as to the protected group(s) against whom the hate speech was directed.”

“I am not asserting nefarious intent, but it is deeply concerning that metrics that showed the disproportionate impact of hate directed at Black, Jewish, Muslim, Arab and LGBTQIA users were not shared with the auditors,” Murphy said in a statement. “Clearly, they have collected some data along these lines.”

The auditors, in the report they released last year, still concluded that Facebook’s policy decisions were a “tremendous setback” for civil rights.

Facebook’s own civil rights auditors say its policy decisions are a ‘tremendous setback’

Facebook spokesman Andy Stone defended the company’s decisions around its hate speech policies and how it conducted its relationship with the civil rights auditors.

“The Worst of the Worst project helped show us what kinds of hate speech our technology was and was not effectively detecting and understand what forms of it people believe to be the most insidious,” Stone said in a statement.

He said progress on racial issues included policies such as banning white nationalist groups, prohibiting content promoting racial stereotypes — such as people wearing blackface or claims that Jews control the media — and reducing the prevalence of hate speech to 0.03 percent of content on the platform.

Facebook approached the civil rights audit with "transparency and openness" and was proud of the progress it has made on issues of race, Stone said.

Stone noted that the company had implemented parts of the "worst of the worst" project. "But after a rigorous internal discussion about these difficult questions, we did not implement all parts as doing so would have actually meant fewer automated removals of hate speech such as statements of inferiority about women or expressions of contempt about multiracial people," he added.

Algorithmic bias

Facebook researchers first showed the racist post featuring The Squad — Reps. Alexandria Ocasio-Cortez (D-N.Y.), Ilhan Omar (D-Minn.), Rashida Tlaib (D-Mich.) and Ayanna Pressley (D-Mass.) — to more than 10,000 Facebook users in an online survey in 2019. (The Squad now has six members.) The users were asked to rate 75 examples of hate speech on the platform to determine what they considered the most harmful.

Other posts among the examples included a post that said, "Many s---hole immagruntz on welfare send money back to their homejungles." An image of a chimpanzee in a long-sleeve shirt was captioned, "Here's one of Michelle Obama." Another post in the survey said, "The only humanitarian assistance needed at the border is a few hundred motion-sensor machine gun turrets. Problem solved."

The 10 worst examples, according to the surveyed users, were almost all directed at minority groups, documents show. Five of the posts were directed at Black people, including statements about mental inferiority and disgust. Two were directed at the LGBTQ community. The remaining three were violent comments directed at women, Mexicans and White people.

People outside Philadelphia's police headquarters in 2019 demand the removal of officers from street duty after the commissioner announced an external review of racist or offensive social media posts. (Bastiaan Slabbers/NurPhoto/Getty Images)

These findings about the most objectionable content held up even among self-identified White conservatives that the market research team traveled to visit in Southern states. Facebook researchers sought out the views of White conservatives in particular because they wanted to overcome potential objections from the company's leadership, which was known to appease right-leaning viewpoints, two people said.

Yet racist posts against minorities weren't what Facebook's own hate speech detection algorithms were most commonly finding. The software, which the company introduced in 2015, was supposed to detect and automatically delete hate speech before users saw it. Publicly, the company said in 2019 that its algorithms proactively caught more than 80 percent of hate speech.

Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show

But this statistic hid a serious problem that was obvious to researchers: The algorithm was aggressively detecting comments denigrating White people more than attacks on every other group, according to several of the documents. One April 2020 document said roughly 90 percent of "hate speech" subject to content takedowns were statements of contempt, inferiority and disgust directed at White people and

men, though the time frame is unclear. And it consistently failed to remove the most derogatory, racist content. The Post previously reported on a portion of the project.

Researchers also found in 2019 that the hate speech algorithms were out of step with actual reports of harmful speech on the platform. In that year, the researchers discovered that 55 percent of the content users reported to Facebook as most harmful was directed at just four minority groups: Blacks, Muslims, the LGBTQ community and Jews, according to the documents.

One of the reasons for these errors, the researchers discovered, was that Facebook's "race-blind" rules of conduct on the platform didn't distinguish among the targets of hate speech. In addition, the company had decided not to allow the algorithms to automatically delete many slurs, according to the people, on the grounds that the algorithms couldn't easily tell the difference when a slur such as the n-word and the c-word was used positively or colloquially within a community. The algorithms were also over-indexing on detecting less harmful content that occurred more frequently, such as "men are pigs," rather than finding less common but more harmful content.

"If you don't do something to check structural racism in your society, you're going to always end up amplifying it," one of the people involved with the project told The Post. "And that is exactly what Facebook's algorithms did."

Facebook agrees to overhaul targeted advertising system for job, housing and loan ads after discrimination complaints

"This information confirms what many of us already knew: that Facebook is an active and willing participant in the dissemination of hate speech and misinformation," Omar said in a statement. "For years, we have raised concerns to Facebook about routine anti-Muslim, anti-Black, and anti-immigrant content on Facebook, much of it based on outright falsehoods. It is clear that they only care about profit, and will sacrifice our democracy to maximize it."

For years, Black users said that those same automated systems also mistook posts about racism as hate speech — sending the user to "Facebook jail" by blocking their account — and made them disproportionate targets of hate speech that the company failed to control. But when civil rights leaders complained, those content moderation issues were routinely dismissed as merely "isolated incidents" or "anecdotal," said Rashad Robinson, president of Color of Change, a civil rights group that regularly sought more forceful action by the company against hate speech and incitements to violence on Facebook, and has argued that Kaplan should be fired.

"They would regularly push back against that," Robinson said. "They would say, 'That's simply not true, Rashad.' They'd say, 'Do you have data to support that?' "

Malkia Devich-Cyril, a Black and queer activist, and the former executive director of the Center for Media Justice, who ran two Black Lives Matter pages on Facebook in 2016, said they had to stop managing the pages because they were "harassed relentlessly," including receiving death threats.

Malkia Devich-Cyril, seen at the premiere of "13th" in 2016, had to stop managing two Black Lives Matter pages on Facebook because they were "harassed relentlessly," including receiving death threats. (Charles Sykes/Invision/AP)

“It sickened me,” Devich-Cyril said. “As an activist — whose calling is to stand on the front lines and fight for change — it created in me a kind of fear. If that kind of chill factor in a democratic state is what Facebook is going for, they have achieved it.”

One set of rules for everyone

In December 2019, researchers on the “worst of the worst,” which came to be known as Project WoW, were ready to deliver their findings from two years of work to key company leaders, including Kaplan and head of global policy management Monika Bickert.

They were proposing a major overhaul of the hate speech algorithm. From now on, the algorithm would be narrowly tailored to automatically remove hate speech against only five groups of people — those who are Black, Jewish, LGBTQ, Muslim or of multiple races — that users rated as most severe and harmful. (The researchers hoped to eventually expand the algorithm’s detection capabilities to protect other vulnerable groups, after the algorithm had been retrained and was on track.) Direct threats of violence against all groups would still be deleted.

Facebook’s vice president for global public policy, Joel Kaplan, left, and Mark Zuckerberg, the company’s CEO, in 2018. (Christophe Morin/Bloomberg News)

Facebook users could still report any post they felt was harmful, and the company’s content moderators would take a second look at it.

The team knew that making these changes to protect more vulnerable minorities over others would be a hard sell, according to the people familiar with the situation. Facebook largely operates with one set of standards for billions of users. Policies that could benefit a particular country or group were often dismissed because they were not “scalable” around the globe, and could therefore interfere with the company’s growth, according to many former and current employees.

Civil rights groups flagged dozens of anti-Muslim pages and groups to Facebook that stayed up, lawsuit alleges

In February 2020, Kaplan and other leaders reviewed the proposal — and quickly rejected the most substantive changes. They felt the changes too narrowly protected just a few groups, while leaving out others, exposing the company to criticism, according to three of the people. For example, the proposal would not have allowed the automatic deletion of comments against Mexicans or women. The document prepared for Kaplan referenced that some “conservative partners” might resist the change because they think that “hate targeted toward trans people is an expression of opinion.”

When asked for comment on Kaplan bending to conservatives, Facebook’s Stone said that Kaplan’s objection to the proposal was because of the types of hate speech it would no longer automatically delete.

Kaplan, the company’s most influential Republican, was widely known as a strong believer in the idea that Facebook should appear “politically neutral,” and his hard-line free speech ideology was in lockstep with company CEO Mark Zuckerberg. (Facebook recently changed its corporate name to Meta.) He bent

over backward to protect conservatives, according to previous reporting in The Post, numerous former insiders and the Facebook Papers.

But Kaplan and the other executives did give the green light to a version of the project that would remove the least harmful speech, according to Facebook's own study: programming the algorithms to stop automatically taking down content directed at White people, Americans and men. The Post previously reported on this change when it was announced internally later in 2020.

"Facebook seems to equate protecting Black users with putting its thumb on the scale," said David Brody, senior counsel for the Lawyers' Committee for Civil Rights Under Law, when The Post presented him the company's research. "The algorithm that disproportionately protected White users and exposed Black users — that is when Facebook put its thumb on the scale."

Complaint alleges that Facebook is biased against black workers

This year, Facebook conducted a consumer product study on "racial justice" that found Black users were leaving Facebook. It found that younger Black users in particular were drawn to TikTok. It appeared to confirm a study from three years ago called Project Vibe that warned that Black users were "in danger" of leaving the platform because of "how Facebook applies its hate speech policy."

"The degree of death threats on these platforms, specifically Facebook, that my colleagues have suffered is untenable," said Devich-Cyril, who added that today they rarely post publicly about politics on Facebook. "It's too unsafe of a platform."