https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184

Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts.

AI has only minimal success in removing hate speech, violent images and other problem content, according to internal company reports

By Deepa Seetharaman Follow , *Jeff Horwitz* Follow and *Justin Scheck* Follow
Oct. 17, 2021 9:17 am ET

Facebook Inc. executives have long said that artificial intelligence would address the company's chronic problems keeping what it deems hate speech and excessive violence as well as underage users off its platforms.

That future is farther away than those executives suggest, according to internal documents reviewed by The Wall Street Journal. Facebook's AI can't consistently identify first-person shooting videos, racist rants and even, in one notable episode that puzzled internal researchers for weeks, the difference between cockfighting and car crashes.

On hate speech, the documents show, Facebook employees have estimated the company removes only a sliver of the posts that violate its rules—a low-single-digit percent, they say. When Facebook's algorithms aren't certain enough that

content violates the rules to delete it, the platform shows that material to users less often—but the accounts that posted the material go unpunished.

The employees were analyzing Facebook's success at enforcing its own rules on content that it spells out in detail internally and in public documents like its community standards.

> The documents reviewed by the Journal also show that Facebook two years ago cut the time human reviewers focused on hate-speech complaints from users and made other tweaks that reduced the overall number of complaints. That made the company more dependent on AI enforcement of its rules and inflated the apparent success of the technology in its public statistics.
>
> — Facebook senior engineer and research scientist

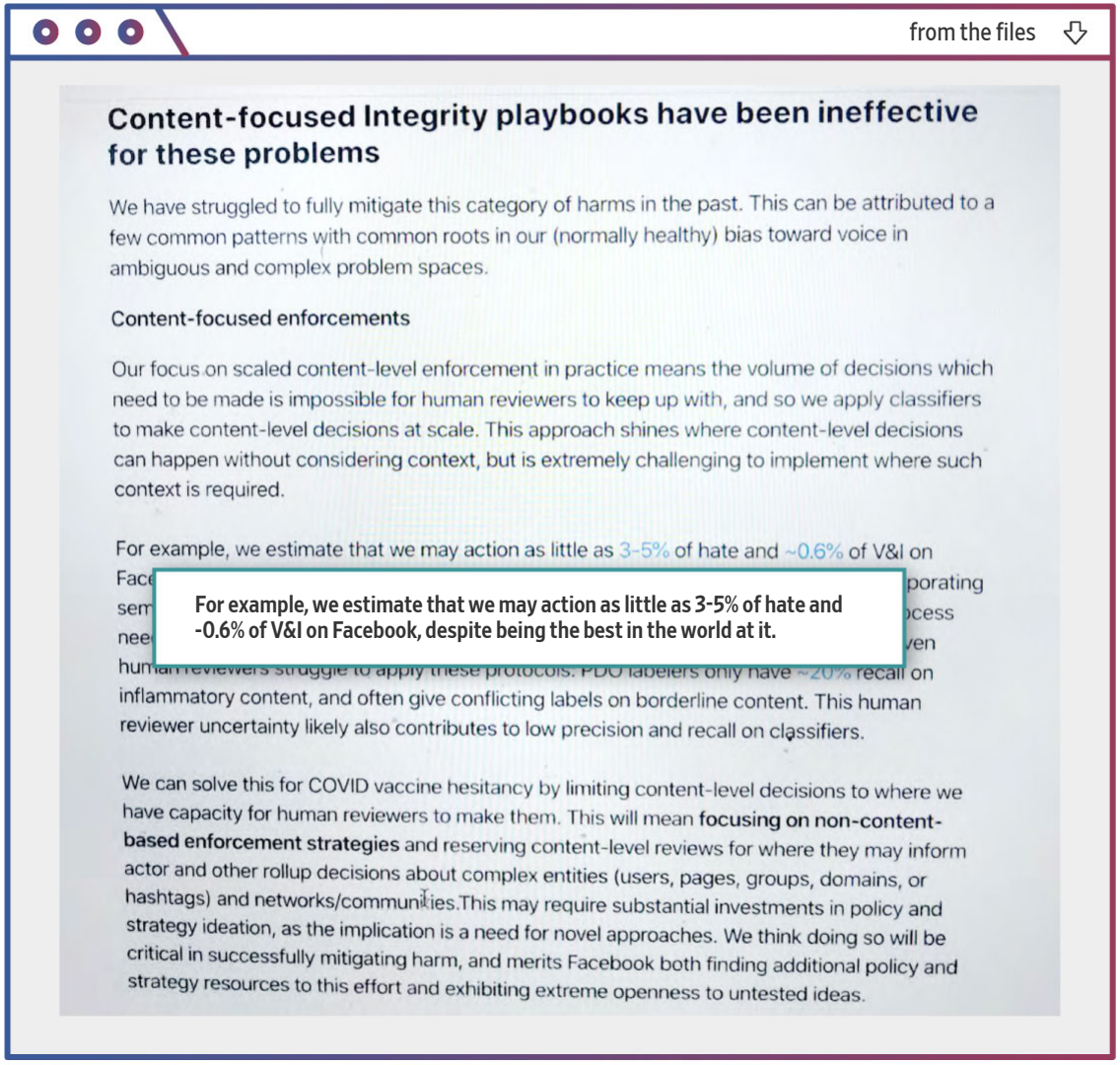
According to the documents, those responsible for keeping the platform free from content Facebook deems offensive or dangerous acknowledge that the company is nowhere close to being able to reliably screen it.

"The problem is that we do not and possibly never will have a model that captures even a majority of integrity harms, particularly in sensitive areas," wrote a senior engineer and research scientist in a mid-2019 note.

He estimated the company's automated systems removed posts that generated just 2% of the views of hate speech on the platform that violated its rules. "Recent

estimates suggest that unless there is a major change in strategy, it will be very difficult to improve this beyond 10-20% in the short-medium term," he wrote.

This March, another team of Facebook employees drew a similar conclusion, estimating that those systems were removing posts that generated 3% to 5% of the views of hate speech on the platform, and 0.6% of all content that violated Facebook's policies against violence and incitement.



Source: Internal report titled, "'Harmful Non-Violating Narratives" is a Problem Archetype in Need of Novel Solutions'

Facebook spokesman Andy Stone said that these percentages referred to posts that were removed using AI, and didn't include other actions the company takes to reduce how many people view hate speech, including ranking posts lower in news feeds. Facebook says by that measure, the prevalence of content that

violates its policies has been shrinking, and that is what the company considers its most important enforcement metric.

The statistics contrast starkly with the confidence in AI presented by Facebook's top executives, including CEO Mark Zuckerberg, who previously said he expected Facebook would use AI to detect "the vast majority of problematic content" by the end of 2019.

The company often says that nearly all of the hate speech it takes down was discovered by AI before it was reported by users. It calls this figure its proactive detection rate, and it had reached nearly 98% as of earlier this year.

Civil rights groups and academics have long been skeptical that the AI detection rate shows meaningful progress, saying it doesn't seem to match user experiences or their own studies. "They won't ever show their work," said Rashad Robinson, president of the civil rights group Color of Change, which helped organize an advertiser boycott of Facebook last year due to what it called the company's failure to control hate speech.

"We ask, what's the numerator? What's the denominator? How did you get that number?" he said. "And then it's like crickets."

In an interview, Facebook's head of integrity, Guy Rosen, said it was more important to look at other data points that show the amount of hate speech shrinking as a percentage of what people see on the platform overall. Facebook says five out of every 10,000 content

views contained hate speech, an improvement from roughly 10 of every 10,000 views in mid-2020, according to its latest public report on how it enforces its policies, for the second quarter of this year.

“Prevalence is the most important metric, and it represents not what we caught, but what we missed, and what people saw, and it’s the primary metric we hold ourselves accountable to,” Mr. Rosen said. “We’ve been successful in moving it down, and it’s the one that we really focus on.”

Mr. Stone, the spokesman, said Facebook executives have increasingly emphasized this measurement in their public comments. He said much of the improvement has come because AI ranks suspected content lower to give it less visibility.

Mr. Rosen also said the documents reviewed by the Journal were outdated, but that they had informed Facebook’s broader thinking about AI-driven content moderation.

Last month, the company said its AI systems were getting better at “proactively removing content that violates our standards on hate speech” and said it was removing 15 times more of this content than in 2017.

The documents are part of extensive internal communications reviewed by the Journal that offer an unprecedented look at Facebook’s struggles to manage the products and systems at the heart of its business success.

The Journal’s series, based on the documents and interviews with current and former employees, describes how the company’s rules favor elites; how its algorithms foster discord; that it has long known drug cartels and human

traffickers use its services openly; and how Facebook is used by antivaccine activists, among other issues. An article about Instagram's effects on teenage girls' mental health spurred a Senate hearing in late September.

Examples of content that Facebook's AI should have detected but missed include close-up videos of a person shooting someone, and videos of car crashes with "dismemberment and visible innards," according to the documents. Other violations of Facebook's policies that slipped through AI were violent threats directed at transgender children.

A memorial site for the shooting victims in Christchurch, New Zealand, in 2019. The attack was live streamed on Facebook.

PHOTO: VINCENT THIAN/ASSOCIATED PRESS

Facebook says it has spent about $13 billion on "safety and security" since 2016, or nearly 4% of its revenue in that time. Mr. Rosen said that in 2016, Facebook's content-moderation system relied largely on user complaints and that the company has since built AI tools to find the objectionable content.

In 2018, Mr. Zuckerberg told a Senate committee that he was optimistic that within five to 10 years, Facebook would have the AI tools to proactively detect most hate speech. “Over the long term, building AI tools is going to be the scalable way to identify and root out most of this harmful content,” he said at the time.

In July 2020, he told Congress: “In terms of fighting hate, we’ve built really sophisticated systems.”

A Facebook executive testified at the late-September Senate hearing that the company is using AI to keep kids under 13 off Instagram.

Facebook’s artificial-intelligence systems comb through billions of posts looking for items that might match the company’s definitions of content that violates its rules. The screening algorithms, called classifiers, are the bedrock of the company’s content-moderation system.

Building these classifiers is labor intensive and complex, requiring an army of humans to mark a vast number of posts based on a set of rules. Engineers then take these examples and train their systems to determine the probability that other posts violate the rules.

Facebook’s algorithms can automatically remove hate speech when they reach a certain level of confidence that the post violates policies, or they can push lower on feeds more questionable posts to limit their spread.

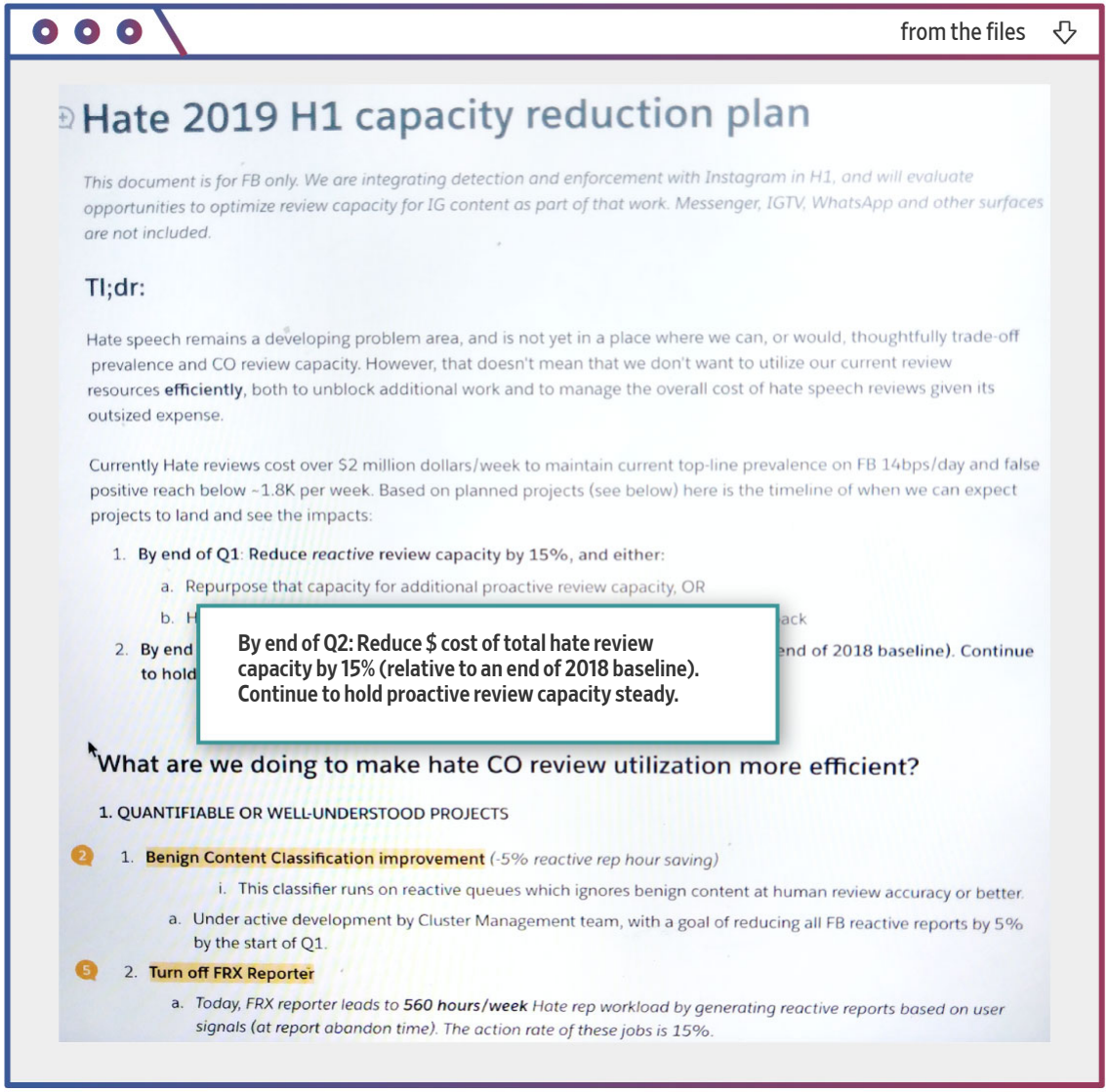
In some areas, such as with spam, Facebook's classifiers work relatively well. But they often fall short in sensitive and controversial areas, especially when Facebook's rules are complex and cultural context matters, according to the documents and people familiar with the matter.

"The classifiers are like elementary school students and they need teachers (human reviewers) to grow into PhDs," one Facebook engineer wrote in a discussion about hate-speech costs on Facebook's internal employee platform in August 2019. Based on one measure of success, the engineer wrote, "our classifiers are still pretty naive."

In one example, AI labeled a video of a carwash as a first-person shooter video, according to the documents. In another, it mistook a video of a shooting for a car crash.

Some employees say Facebook is misusing the classifiers, which they say are more effective as tools to flag broad problem areas than as the main tool for removing specific content problems.

In 2019, documents reviewed by the Journal show, Facebook introduced "hate speech cost controls" to save money on its human content review operations. Review of hate speech by human staff was costing $2 million a week, or $104 million a year, according to an internal document covering planning for the first half of that year.



Source: 2019 document titled 'Hate 2019 H1 capacity reduction plan'

"Within our total budget, hate speech is clearly the most expensive problem," a manager wrote of the effort in a separate document, declaring that the cost of policing slurs and the denigration of minority groups, which Facebook rules bar, "adds up to real money."

Mr. Stone, the spokesman, said the funds were shifted to hire more people to train Facebook's algorithms and that the overall budget stayed steady.

Roughly 75% of the costs came from employing people to review user complaints, the vast majority of which were deemed, after review, to not be hate speech, the documents show. In 2019, beyond simply cutting the number of contractor hours dedicated to reviewing hate speech, the company began employing an algorithm that led them to ignore a larger percentage of user reports that the system deemed unlikely to be violations.

It also introduced "friction" to the content reporting process, adding hoops for aggrieved users to jump through that sharply reduced how many complaints about content were made, according to the documents.

"We may have moved the needle too far," the author of one of the documents acknowledged of the company's efforts to make it less likely that users would complete their reports on hate speech to the company.

The moves helped boost the company's proactive detection rate, meaning, a greater proportion of the content that was removed was flagged by AI—the figure that is now nearly 98%. In December 2017, 24% of removed hate speech was detected by AI, and the rest from user reports, according to Facebook's quarterly public report on how it enforces its policies.

Mr. Stone said the moves to ignore user reports deemed unlikely to be violations and the addition of friction weren't intended to change the proactive detection rate but instead were intended to make the system more efficient. He added that some of that additional friction has since been rolled back.

The performance of Facebook's automated systems illustrates how difficult it is for Facebook and other tech companies to build systems that reliably and comprehensively detect content that breaks their rules.

"This is one of the hardest problems in machine learning," said J. Nathan Matias, an assistant professor at Cornell University. "It's also an area that so many companies and policy makers have just decided was going to be the solution—without understanding the problem."

The discrepancy between Facebook's public claims about the effectiveness of its AI and the reality of the user experience has long puzzled researchers and other heavy users of the platform.

In 2016, pop star Selena Gomez flew to Facebook's Menlo Park headquarters to pose for pictures with Mr. Zuckerberg and Facebook's Chief Operating Officer Sheryl Sandberg to celebrate her status as the most-followed account on Instagram. Not long after, she was startled to read a user comment on one of her Instagram posts: "Go kill yourself," according to the star's spokesman.

She grew increasingly concerned about the spread of hate speech on these platforms, and in September 2020 she sent an Instagram message that she later posted on her account to Mr. Zuckerberg and Ms. Sandberg, saying the company had a "serious problem" with hate, misinformation, racism and bigotry.

selenagomez
343M followers

View profile

View more on Instagram

2,883,698 likes

selenagomez

When you meet the boss and talk tech in a very unusual office @zuck

view all 16,594 comments

Add a comment...

Ms. Gomez then followed up by email to ask why Facebook allowed hate groups to thrive on the site, according to emails reviewed by the Journal and previously reported by the Associated Press. Ms. Sandberg responded that Facebook's AI had detected 91% of the 1.5 million posts it removed for violating its rules against using symbols or phrases from hate groups.

Ms. Gomez wrote back that Ms. Sandberg hadn't addressed her broader questions, sending screenshots of Facebook groups that promoted violent ideologies.

"You refuse to even mention, let alone address, the problem Facebook has with white supremacists and bigots," Ms. Gomez wrote in an Oct. 10, 2020, email to Ms. Sandberg and other executives, adding that there were plenty of Facebook groups "full of hate and lies that might lead to people being hurt or, even worse, killed."

Ms. Gomez declined requests for further comment.

Mr. Stone said Ms. Sandberg has publicly highlighted Facebook's hate-speech prevalence figures this year.

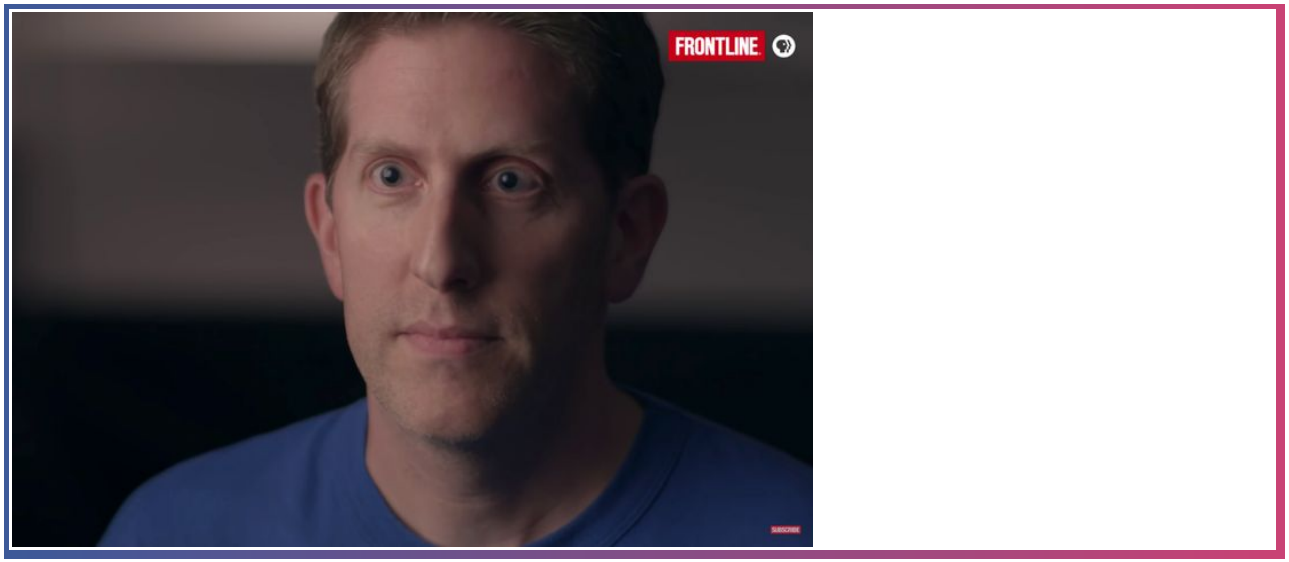
Fadi Quran, a researcher at the human-rights group Avaaz, which advocates for citizen action in areas such as climate change and poverty, said he has repeatedly asked Facebook employees if they understood how much hate speech was on their platform and how much they acted on. “They said verbatim that that was almost impossible, and they can only report with certainty on what they detect,” he said.

“By hiding the problem and giving the opposite impression—that the issue is under control—they’re actually complicit in allowing those community violations to go forward with minimal accountability,” he said.

Mr. Stone said Facebook provided Mr. Quran with public prevalence figures and other metrics.

In its quarterly public reports on how it enforces its policies, Facebook measures the prevalence of certain types of content, like hate speech, by the number of views that content attracts. The company says this is a more accurate way of measuring the true impact of a piece of content that violates its policies. In other words, hate speech viewed a million times is more of a problem than hate speech viewed just once.

The company doesn’t publicly report what percentage of hate-speech views it removes. Internally, the company calculates this figure by applying their hate-speech classifiers to a sample of posts and then having humans review the same posts to see how much the classifiers missed, according to a person with direct knowledge of the estimates. The number is then used as an estimate for the amount of hate-speech views removed across the whole platform.



Guy Rosen, Facebook's head of integrity, during a 2018 interview on Frontline.

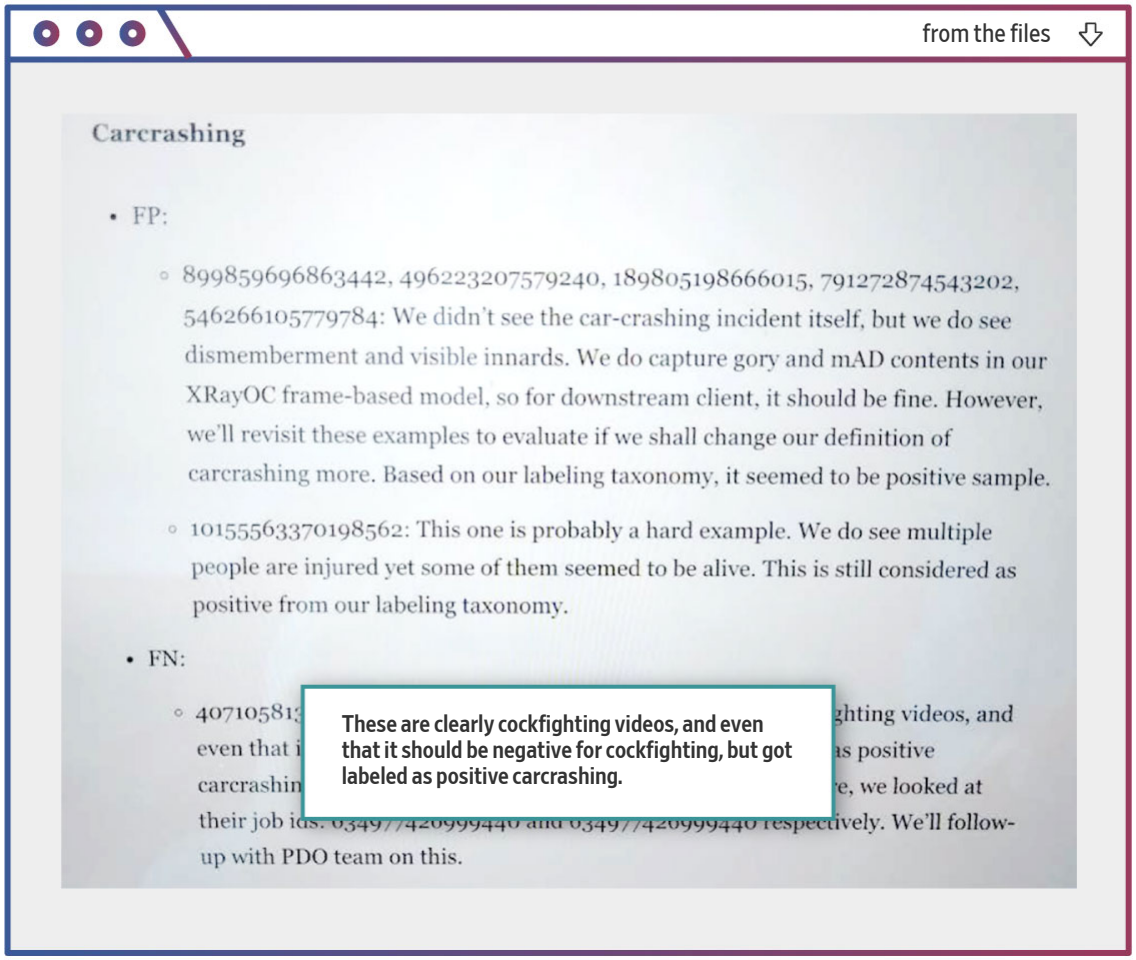
PHOTO: PHOTO COURTESY OF FRONTLINE

Mr. Rosen, the integrity chief, said in the interview that the company's quarterly public reports are evidence it is taking these problems seriously.

In mid-2018, an engineer noticed a troubling trend: "a lot of car crashing and cockfighting in prevalence data," he wrote in a 2019 internal report. Facebook users were finding in their feeds videos of crashing cars and fighting roosters, which would normally violate Facebook's rules. Data scientists weren't sure why.

The engineer and a team of colleagues trained an artificial intelligence system to recognize videos of cockfights and car crashes and weed them out. "However," the

engineers wrote in a memo, “the problem didn’t really get solved.”



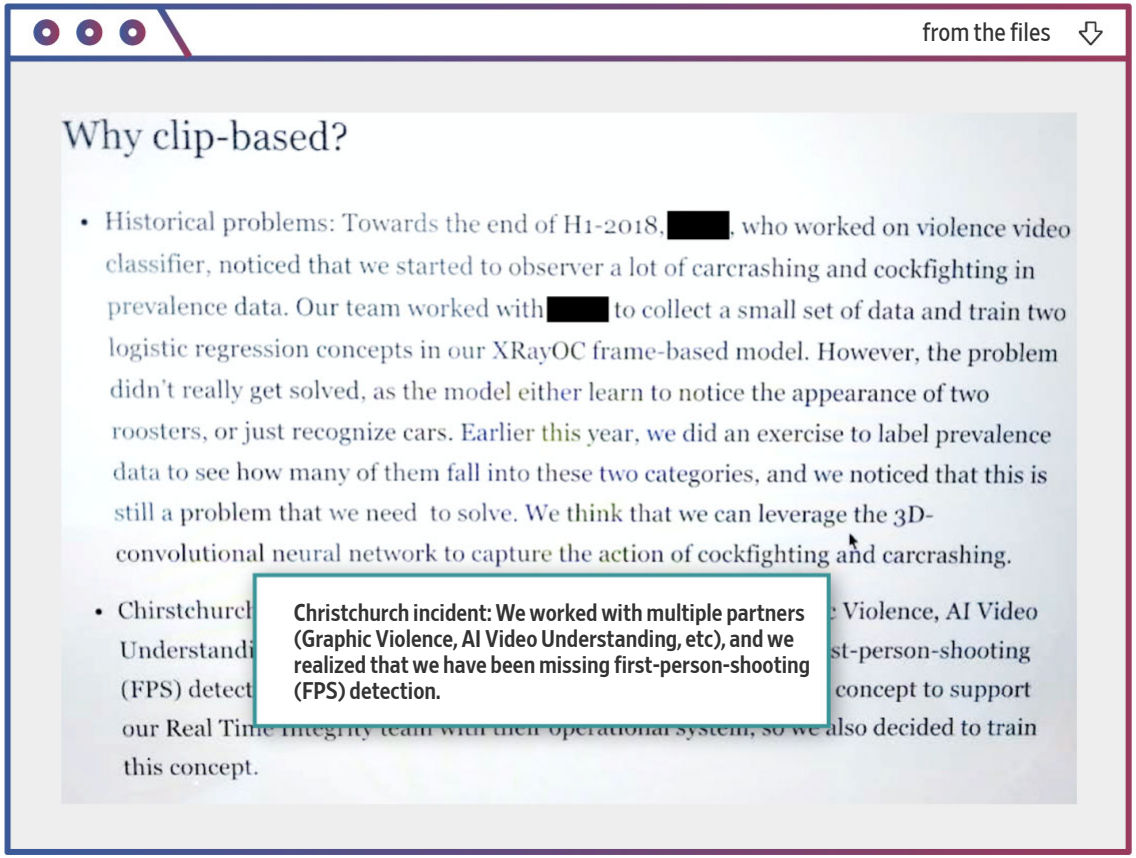
Source: June 2019 internal note titled ‘XRayOC 2019a clip-based model’

Facebook also set rules for cockfighting that the AI had trouble following. Mild cockfights were deemed acceptable, but those in which the birds were seriously hurt were banned. But the computer model couldn’t distinguish fighting roosters from non-fighting roosters.

To train the company's systems, the engineers employed sophisticated machine-learning programs with names like "Deep Vision" and fed hours and hours of cockfighting videos into them. Teaching the AI to flag a severely injured bird and ignore a less injured one proved difficult.

"This is hard to catch," the engineers wrote.

In two cases where the engineers did get the AI to flag a cockfight, they turned up another problem: "These are clearly cockfighting videos," but they were labeled as car crashes, the researchers wrote.



Note: A name has been redacted on this document.
Source: June 2019 internal note titled 'XRayOC 2019a clip-based model'

The same team hit obstacles around shootings recorded by the perpetrator, known as "first-person shooter" videos, the internal memo says. Three months before the memo was written, a man in Christchurch, New Zealand, used Facebook to live stream his fatal shooting of 51 people in two mosques.

In some cases, the AI didn't recognize shootings. In others, it mislabeled innocuous videos, such as paintball games, or the carwash, the researchers wrote.

The AI must also be trained in foreign languages.

According to a December 2020 memo, Facebook employees debated creating a hate-speech classifier for various Arabic dialects. But the lack of training data—such as samples of the various dialects—was a problem, especially since they were having trouble with standard Arabic. “As it stands, they have barely enough content to train and maintain the Arabic classifier currently—let alone breakdowns,” one employee wrote in a document.

In January, a Facebook employee reported that hate speech was one of the top “abuse categories” in Afghanistan, but the company took action against just 0.23% of the estimated hate speech posts in the country.

The employee said that the company’s “seriously scant” list of slurs in the languages spoken in Afghanistan meant it could be missing many violating posts.

In March, employees gearing up for regional elections in India said hate speech was a major risk in Assam, where there is growing violence against Muslims and other ethnic groups. “Assam is of particular concern because we do not have an Assamese hate-speech classifier,” according to one planning document.

Indian students and doctors protest in Assam state, India. A Facebook employee warned that hate speech related to ethnic violence in Assam was a major risk on the platform.

PHOTO: DAVID TALUKDAR/AGENCE FRANCE-PRESSE/GETTY IMAGES

While Facebook removes a tiny fraction of the content that violates its rules, executives are particularly sensitive to what it calls “over-enforcement,” or taking down too many posts that don’t actually violate hate-speech rules, according to people familiar with the matter. The emphasis on preventing those mistakes has pushed company engineers to train models that, in effect, allow for more hate speech on the platform to avoid false positives, according to the people.

Its own internal research shows that Facebook users world-wide are more concerned about lack of enforcement. In March 2020, Facebook found that users, on average, rated seeing violating content like hate speech as a more negative experience than having their content taken down by mistake, according to the documents.

Globally, users ranked inaccurate content removals last among a series of problems, while hate speech and violence topped the list. American users were more concerned by inaccurate removals, but still rated the problem behind hate speech and violence, the survey shows.

In a late 2020 note, a departing data scientist noted that Facebook has a policy of allowing groups to sanction hate speech five times before they are removed from the platform. Because Facebook's systems miss so much hate speech, the groups are likely to get away with far more, the data scientist wrote.

— Facebook data scientist

"When you consider that we miss 95% of violating hate speech, you realize that it might actually take 100 violations for that group to accrue its five strikes," he said in the note, which was previously reported by BuzzFeed.

The outgoing data scientist noted that despite intense investment by Facebook, the company's success rate at removing banned content remained dismal. "Each half [year] we make incremental progress on the amount of content we're able to proactively detect," he wrote. "But an incremental increase on a very small number is still a very small number."

"We might just be the very best in the world at it," he wrote, "but the best in the world isn't good enough to find a fraction of it."

—Design by Andrew Levinson. A color filter has been used on some photos.

Write to Deepa Seetharaman at Deepa.Seetharaman@wsj.com, Jeff Horwitz at Jeff.Horwitz@wsj.com and Justin Scheck at justin.scheck@wsj.com

Appeared in the October 18, 2021, print edition as 'Facebook Staff Express Doubt on Power of AI.'