Facebook staff say core products make misinformation worse

October 25 2021, by Daniel Zuidijk and Michael Riley

Credit: CC0 Public Domain

For years, Facebook has fought back against allegations that its platforms play an outsized role in the spread of false information and harmful content that has fueled conspiracies, political divisions and

distrust in science, including COVID-19 vaccines.

But research, analysis and commentary contained in a vast trove of internal documents indicate that the company's own employees have studied and debated the issue of misinformation and harmful content at length, and many of them have reached the same conclusion: Facebook's own products and policies make the problem worse.

In 2019, for instance, Facebook created a fake account for a fictional, 41-year-old North Carolina mom named Carol, who follows Donald Trump and Fox News, to study misinformation and polarization risks in its recommendation systems. Within a day, the woman's account was directed to "polarizing" content and within a week, to conspiracies including QAnon.

"The content in this account (followed primarily via various recommendation systems!) devolved to a quite troubling, polarizing state in an extremely short amount of time," according to a Facebook memo analyzing the fictional U.S. woman's account. When a similar experiment was conducted in India, a test account representing a 21-year-old woman was in short order directed to pictures of graphic violence and doctored images of India air strikes in Pakistan.

Memos, reports, internal discussions and other examples contained in the documents suggest that some of Facebook's core product features contribute to the spread of false and polarizing information globally and that suggestions to fix them can face significant internal challenges. Facebook's efforts to quell misinformation and harmful content, meanwhile, have sometimes been undercut by political considerations, the documents indicate.

"We have evidence from a variety of sources that hate speech, divisive political speech, and misinformation on Facebook and the family of apps

are affecting societies around the world," an employee noted in an internal discussion about a report entitled "What is Collateral Damage?"

"We also have compelling evidence that our core product mechanisms, such as virality, recommendations and optimizing for engagement, are a significant part of why these types of speech flourish on the platform."

The documents were disclosed to the U.S. Securities and Exchange Commission and provided to Congress in redacted form by whistle-blower Frances Haugen's legal counsel. The redacted versions were obtained by a consortium of news organizations, including Bloomberg. The documents represent a selection of information produced mostly for internal Facebook audiences. The names of employees are redacted, and it's not always clear when they were created. Some of the documents have been previously reported by the Wall Street Journal, BuzzFeed News and other media outlets.

Facebook has pushed back against the initial allegations, noting that Haugen's "curated selection" of documents "can in no way be used to draw fair conclusions about us." Facebook Chief Executive Mark Zuckerberg said the allegations that his company puts profit over user safety are "just not true."

"Every day our teams have to balance protecting the ability of billions of people to express themselves openly with the need to keep our platform a safe and positive space," Joe Osborne, a Facebook spokesman said in a statement. "We continue to make significant improvements to tackle the spread of misinformation and harmful content. To suggest we encourage bad content and do nothing is just not true."

The experimental account for the North Carolina woman is just the kind of research the company does to improve and help inform decisions such as removing QAnon from the platform, according to a Facebook

statement. The increase in polarization predates social media and despite serious academic research there isn't much consensus, the company said, adding that what evidence there is doesn't support that idea that Facebook—or social media more generally—is the primary cause.

Still, while the social media giant has undoubtedly made progress in disrupting and disclosing the existence of interference campaigns orchestrated by foreign governments—and collaborated with external organizations to address false claims—it has often failed to act against emerging political movements such as QAnon or vaccine misinformation until they have spread widely, according to critics.

The documents reflect a company culture that values open debate and disagreement and is driven by the relentless collection and analysis of data. But the resulting output, which often lays bare the company's shortcomings in stark terms, could create a serious challenge ahead: a whistleblower complaint filed to the SEC, which is included in the cache of documents, alleges, "Facebook knows that its products make hate speech and misinformation worse" and that it has misrepresented that fact repeatedly to investors and the public.

Those alleged misrepresentations include Zuckerberg's March appearance before Congress, where he expressed confidence that his company shared little of the blame for the worsening political divide in the U.S. and across the globe. "Now, some people say that the problem is the social networks are polarizing us," Zuckerberg told the lawmakers. "But that's not at all clear from the evidence or research."

But the documents often tell a different story.

"We've known for over a year now that our recommendation systems can very quickly lead users down the path to conspiracy theories and groups," a Facebook employee wrote on their final day in August 2020.

Citing examples of safeguards the company had rolled back or failed to implement, the employee wrote, "During the time that we hesitated, I've seen folks from my hometown go further and further down the rabbit hole of QAnon and COVID anti-mask/anti-vax conspiracy on FB. It has been painful to observe."

Facebook said in its statement selecting anecdotes from departing employees doesn't tell the story of how changes happen at the company. Projects go through rigorous reviews and debates, according to the statement, so that Facebook can be confident in any potential changes and its impact on people. In the end, the company ended up implementing many of the ideas raised in this story, according to the statement.

Like other major social media platforms, Facebook has for years struggled with the problem of false information in part because it doesn't necessarily contain slurs or particular phrases that can be easily screened. In addition, figuring out what posts are false and potentially harmful isn't an exact science—a problem made even more difficult by different languages and cultural contexts.

Facebook relies on artificial intelligence to scan its vast user base for potential problems and then sends flagged posts to a collection of fact-checking organizations spread around the world. If the fact checkers rate something as false, Facebook adds a warning label and reduces the distribution so fewer people can see it, according to a March 2021 post by Guy Rosen, vice president of integrity.

The most serious kinds of disinformation, including false claims about COVID-19 vaccines, may be removed. It's a process that is complicated by crushing volume from nearly 3 billion users.

Facebook has provided some details on ways it has succeeded at curbing

misinformation. For instance, it disabled more than 1.3 billion accounts between October and December 2020—amid a contentious U.S. presidential election. And over the past three years, the company removed more than 100 networks for coordinated inauthentic behavior, when groups of pages or people work together to mislead people, according to Rosen's post.

And yet, aside from the challenges of trying to monitor a colossal volume of data, the company's system for screening and removing false and potentially harmful claims has significant flaws, according to the documents. For instance, political concerns can shape how Facebook reacts to false postings.

In one September 2019 incident, a decision to remove a video posted by the anti-abortion group Live Action was overturned "after several calls from Republican senators."

The video, which claimed incorrectly that "abortion was never medically necessary," was reposted after Facebook declared it "not eligible for fact-checking," according to one of the documents.

"A core problem at Facebook is that one policy org is responsible for both the rules of the platform and keeping governments happy," a former employee is quoted as saying in one December 2020 document. "It is very hard to make product decisions based upon abstract principles when you are also measured on your ability to keep innately political actors from regulating/investigating/prosecuting the company."

In addition, politicians, celebrities and certain other special users are exempt from many of the company's content review procedures, through a process called "whitelisting." For example, videos by and of President Donald Trump were repeatedly flagged on Instagram for incitement to violence in the run up to the Jan. 6 Capitol riots, the documents indicate.

"By providing this special exemption to politicians, we are knowingly exposing users to misinformation that we have the processes and resources to mitigate," according to a 2019 employee post entitled "The Political Whitelist Contradicts Facebook's Core State Principles."

Facebook employees repeatedly cite policies and products at Facebook that they believe have contributed to misinformation and harmful conduct, according to the documents. Their complaints are sometimes backed by research or proposals to fix or minimize the problems

For instance, employees have cited the fact that misinformation contained in comments to other posts is scrutinized far less carefully than the posts themselves, even though comments have a powerful sway over users. The "aggregate risk" from vaccine hesitancy in comments may be higher than from posts, "and yet we have under-invested in preventing vaccine hesitancy in comments compared to our investment in content," concluded an internal report entitled "Vaccine Hesitancy is Twice as Prevalent in English Vaccine Comments compared to English Vaccine Posts."

In its statement, Facebook said it demoted comments that match known misinformation, are shared by repeat offenders or violate its community standards.

Many of the employees' suggestions pertain to Facebook's algorithms, including a change in 2018 that was intended to encourage more meaningful social interactions but ended up fueling more provocative, low-quality content.

The company changed the ranking for its News Feed to prioritize meaningful social interactions and deprioritize things like viral videos, according to its statement. That change led to a decrease in time spent on Facebook, according to the statement, which noted it wasn't the kind of

thing a company would do if it was simply trying to drive people to use the service more.

In internal surveys, Facebook users report their experience on the platform has worsened since the change, and they say it doesn't give them the kind of content they would prefer to see. Political parties in Europe asked Facebook to suspend its use, and several tests by the company indicate that it quickly led users to content supporting conspiracy theories or denigrating other groups.

"As long as we continue to optimize for overall engagement and not solely what we believe individual users will value, we have an obligation to consider what the effect of optimizing for business outcomes has on the societies we engage in," one employee argued in a report called "We are Responsible for Viral Content," posted in December 2019.

Similarly, after the New York Times published an op-ed in January 2021, shortly after the raid on the U.S. Capitol, explaining how Facebook's algorithms entice users to share extreme views by rewarding them with likes and shares, an employee noted that the article mirrored other research and called it "a problematic side-effect of the architecture of Facebook as a whole."

"In my first report 'Qurios about QAnon,' I recommended removing /disallowing social metrics such as likes as a way to remove the 'hit' that comes from watching those likes grow."

Instagram had also previously experimented with removing likes from their posts, which culminated in a May 26 announcement that the company would begin giving users of the platform the ability to hide likes if they chose.

The documents do provide some details, albeit incomplete, of the

company's efforts to reduce the spread of misinformation and harmful content. In a literature review published in January 2020, the author detailed how the company already banned "the most serious, repeat violators" and limited "access to abuse-prone features" to discourage the distribution of harmful content.

Teams within the company were assigned to look for ways to make improvements, with at least two documents indicating that a task force had been created to consider "big ideas to reduce the prevalence of bad content in the News Feed" to focus on "soft actions" that stopped short of removing content.It's not clear how many of those recommendations were instituted and if so, whether they were successful.

In the goodbye note from August 2020, the Facebook employee praised colleagues as "amazing, brilliant and extraordinary." But the employee also rued how many of their best efforts to curtail misinformation and other "violating content" had been "stifled or severely constrained by key decision-makers – often based on fears of public and policy stakeholder responses."

"While mountains of evidence is (rightly) required to support a new intervention, none is required to kill (or severely limit) one," the employee wrote.