How Facebook got addicted to spreading misinformation

The company's AI algorithms gave it an insatiable habit for lies and hate speech. Now the man who built them can't fix the problem.

MIT Technology Review, Karen Hao, March 11 2021

Find the original article here: https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Joaquin Quiñonero Candela, a director of AI at Facebook, was apologizing to his audience.

It was March 23, 2018, just days after the revelation that Cambridge Analytica, a consultancy that worked on Donald Trump's 2016 presidential election campaign, had surreptitiously siphoned the personal data of tens of millions of Americans from their Facebook accounts in an attempt to influence how they voted. It was the biggest privacy breach in Facebook's history, and Quiñonero had been previously scheduled to speak at a conference on, among other things, "the intersection of AI, ethics, and privacy" at the company. He considered canceling, but after debating it with his communications director, he'd kept his allotted time.

As he stepped up to face the room, he began with an admission. “I’ve just had the hardest five days in my tenure at Facebook,” he remembers saying. “If there’s criticism, I’ll accept it.”

The Cambridge Analytica scandal would kick off Facebook’s largest publicity crisis ever. It compounded fears that the algorithms that determine what people see on the platform were amplifying fake news and hate speech, and that Russian hackers had weaponized them to try to sway the election in Trump’s favor. Millions began deleting the app; employees left in protest; the company’s market capitalization plunged by more than $100 billion after its July earnings call.

In the ensuing months, Mark Zuckerberg began his own apologizing. He apologized for not taking “a broad enough view” of Facebook’s responsibilities, and for his mistakes as a CEO. Internally, Sheryl Sandberg, the chief operating officer, kicked off a two-year civil rights audit to recommend ways the company could prevent the use of its platform to undermine democracy.

Finally, Mike Schroepfer, Facebook’s chief technology officer, asked Quiñonero to start a team with a directive that was a little vague: to examine the societal impact of the company’s algorithms. The group named itself the Society and AI Lab (SAIL); last year it combined with another team working on issues of data privacy to form Responsible AI.

Quiñonero was a natural pick for the job. He, as much as anybody, was the one responsible for Facebook’s position as an AI powerhouse. In his six years at Facebook, he’d created some of the first algorithms for targeting users with content precisely tailored to their interests, and then he’d diffused those algorithms across the company. Now his mandate would be to make them less harmful.

Facebook has consistently pointed to the efforts by Quiñonero and others as it seeks to repair its reputation. It regularly trots out various leaders to speak to the media about the ongoing reforms. In May of 2019, it granted a series of interviews with Schroepfer to the New York Times, which rewarded the company with a humanizing profile of a sensitive, well-intentioned executive striving to overcome the technical challenges of filtering out misinformation and hate speech from a stream of content that amounted to billions of pieces a day. These challenges are so hard that it makes Schroepfer emotional, wrote the Times: “Sometimes that brings him to tears.”

In the spring of 2020, it was apparently my turn. Ari Entin, Facebook’s AI communications director, asked in an email if I wanted to take a deeper look at the company’s AI work. After talking to several of its AI leaders, I decided to focus on Quiñonero. Entin happily obliged. As not only the leader of the Responsible AI team but also the man who had made Facebook into an AI-driven company, Quiñonero was a solid choice to use as a poster boy.

He seemed a natural choice of subject to me, too. In the years since he’d formed his team following the Cambridge Analytica scandal, concerns about the spread of lies and hate speech on Facebook had only grown. In late 2018 the company admitted that this activity had helped fuel a genocidal anti-Muslim campaign in Myanmar for several years. In 2020 Facebook started belatedly taking action against Holocaust deniers, anti-vaxxers, and the conspiracy movement QAnon. All these dangerous falsehoods were metastasizing thanks to the AI capabilities Quiñonero had helped build. The algorithms that underpin Facebook’s business weren’t created to filter out what was false or inflammatory; they were designed to make people share and engage with as much content as possible by showing them things they were most likely to be outraged or titillated by. Fixing this problem, to me, seemed like core Responsible AI territory.

I began video-calling Quiñonero regularly. I also spoke to Facebook executives, current and former employees, industry peers, and external experts. Many spoke on condition of anonymity because they’d signed nondisclosure agreements or feared retaliation. I wanted to know: What was Quiñonero’s team doing to rein in the hate and lies on its platform?

Joaquin Quiñonero Candela outside his home in the Bay Area, where he lives with his wife and three kids.

But Entin and Quiñonero had a different agenda. Each time I tried to bring up these topics, my requests to speak about them were dropped or redirected. They only wanted to discuss the Responsible AI team's plan to tackle one specific kind of problem: AI bias, in which algorithms discriminate against particular user groups. An example would be an ad-targeting algorithm that shows certain job or housing opportunities to white people but not to minorities.

By the time thousands of rioters stormed the US Capitol in January, organized in part on Facebook and fueled by the lies about a stolen election that had fanned out across the platform, it was clear from my conversations that the Responsible AI team had failed to make headway against misinformation and hate speech because it had never made those problems its main focus. More important, I realized, if it tried to, it would be set up for failure.

The reason is simple. Everything the company does and chooses not to do flows from a single motivation: Zuckerberg's relentless desire for growth. Quiñonero's AI expertise supercharged that growth. His team got pigeonholed into targeting AI bias, as I learned in my reporting, because preventing such bias helps the company avoid proposed regulation that might, if passed, hamper that growth. Facebook leadership has also repeatedly weakened or halted many initiatives meant to clean up misinformation on the platform because doing so would undermine that growth.

In other words, the Responsible AI team's work—whatever its merits on the specific problem of tackling AI bias—is essentially irrelevant to fixing the bigger problems of misinformation, extremism, and political polarization. And it's all of us who pay the price.

"When you're in the business of maximizing engagement, you're not interested in truth. You're not interested in harm, divisiveness, conspiracy. In fact, those are your friends," says Hany Farid, a professor at the University of California, Berkeley who collaborates with Facebook to understand image- and video-based misinformation on the platform.

"They always do just enough to be able to put the press release out. But with a few exceptions, I don't think it's actually translated into better policies. They're never really dealing with the fundamental problems."

In March of 2012, Quiñonero visited a friend in the Bay Area. At the time, he was a manager in Microsoft Research's UK office, leading a team using machine learning to get more visitors to click

on ads displayed by the company's search engine, Bing. His expertise was rare, and the team was less than a year old. Machine learning, a subset of AI, had yet to prove itself as a solution to large-scale industry problems. Few tech giants had invested in the technology.

Quiñonero's friend wanted to show off his new employer, one of the hottest startups in Silicon Valley: Facebook, then eight years old and already with close to a billion monthly active users (i.e., those who have logged in at least once in the past 30 days). As Quiñonero walked around its Menlo Park headquarters, he watched a lone engineer make a major update to the website, something that would have involved significant red tape at Microsoft. It was a memorable introduction to Zuckerberg's "Move fast and break things" ethos. Quiñonero was awestruck by the possibilities. Within a week, he had been through interviews and signed an offer to join the company.

His arrival couldn't have been better timed. Facebook's ads service was in the middle of a rapid expansion as the company was preparing for its May IPO. The goal was to increase revenue and take on Google, which had the lion's share of the online advertising market. Machine learning, which could predict which ads would resonate best with which users and thus make them more effective, could be the perfect tool. Shortly after starting, Quiñonero was promoted to managing a team similar to the one he'd led at Microsoft.

Quiñonero started raising chickens in late 2019 as a way to unwind from the intensity of his job.

Unlike traditional algorithms, which are hard-coded by engineers, machine-learning algorithms "train" on input data to learn the correlations within it. The trained algorithm, known as a machine-learning model, can then automate future decisions. An algorithm trained on ad click data, for example, might learn that women click on ads for yoga leggings more often than men. The resultant model will then serve more of those ads to women. Today at an AI-based company like Facebook, engineers generate countless models with slight variations to see which one performs best on a given problem.<

Facebook's massive amounts of user data gave Quiñonero a big advantage. His team could develop models that learned to infer the existence not only of broad categories like "women" and "men," but of very fine-grained categories like "women between 25 and 34 who liked Facebook pages related to yoga," and targeted ads to them. The finer-grained the targeting, the better the chance of a click, which would give advertisers more bang for their buck.

Within a year his team had developed these models, as well as the tools for designing and deploying new ones faster. Before, it had taken Quiñonero's engineers six to eight weeks to build, train, and test a new model. Now it took only one.

News of the success spread quickly. The team that worked on determining which posts individual Facebook users would see on their personal news feeds wanted to apply the same techniques. Just as algorithms could be trained to predict who would click what ad, they could also be trained to predict who would like or share what post, and then give those posts more prominence. If the model determined that a person really liked dogs, for instance, friends' posts about dogs would appear higher up on that user's news feed.

Quiñonero's success with the news feed—coupled with impressive new AI research being conducted outside the company—caught the attention of Zuckerberg and Schroepfer. Facebook now had just over 1 billion users, making it more than eight times larger than any other social network, but they wanted to know how to continue that growth. The executives decided to invest heavily in AI, internet connectivity, and virtual reality.

They created two AI teams. One was FAIR, a fundamental research lab that would advance the technology's state-of-the-art capabilities. The other, Applied Machine Learning (AML), would integrate those capabilities into Facebook's products and services. In December 2013, after months of courting and persuasion, the executives recruited Yann LeCun, one of the biggest names in the field, to lead FAIR. Three months later, Quiñonero was promoted again, this time to lead AML. (It was later renamed FAIAR, pronounced "fire.")

> "That's how you know what's on his mind. I was always, for a couple of years, a few steps from Mark's desk."
>
> *Joaquin Quiñonero Candela*

In his new role, Quiñonero built a new model-development platform for anyone at Facebook to access. Called FBLearner Flow, it allowed engineers with little AI experience to train and deploy machine-learning models within days. By mid-2016, it was in use by more than a quarter of Facebook's engineering team and had already been used to train over a million models, including models for image recognition, ad targeting, and content moderation.

Zuckerberg's obsession with getting the whole world to use Facebook had found a powerful new weapon. Teams had previously used design tactics, like experimenting with the content and frequency of notifications, to try to hook users more effectively. Their goal, among other things, was to increase a metric called L6/7, the fraction of people who logged in to Facebook six of the previous seven days. L6/7 is just one of myriad ways in which Facebook has measured "engagement"—the propensity of people to use its platform in any way, whether it's by posting things, commenting on them, liking or sharing them, or just looking at them. Now every user interaction once analyzed by engineers was being analyzed by algorithms. Those algorithms were creating much faster, more personalized feedback loops for tweaking and tailoring each user's news feed to keep nudging up engagement numbers.

Zuckerberg, who sat in the center of Building 20, the main office at the Menlo Park headquarters, placed the new FAIR and AML teams beside him. Many of the original AI hires were so close that his desk and theirs were practically touching. It was "the inner sanctum," says a former leader in the AI org (the branch of Facebook that contains all its AI teams), who recalls the CEO shuffling

people in and out of his vicinity as they gained or lost his favor. “That’s how you know what’s on his mind,” says Quiñonero. “I was always, for a couple of years, a few steps from Mark's desk.”

With new machine-learning models coming online daily, the company created a new system to track their impact and maximize user engagement. The process is still the same today. Teams train up a new machine-learning model on FBLearner, whether to change the ranking order of posts or to better catch content that violates Facebook’s community standards (its rules on what is and isn’t allowed on the platform). Then they test the new model on a small subset of Facebook’s users to measure how it changes engagement metrics, such as the number of likes, comments, and shares, says Krishna Gade, who served as the engineering manager for news feed from 2016 to 2018.

If a model reduces engagement too much, it’s discarded. Otherwise, it’s deployed and continually monitored. On Twitter, Gade explained that his engineers would get notifications every few days when metrics such as likes or comments were down. Then they’d decipher what had caused the problem and whether any models needed retraining.

But this approach soon caused issues. The models that maximize engagement also favor controversy, misinformation, and extremism: put simply, people just like outrageous stuff. Sometimes this inflames existing political tensions. The most devastating example to date is the case of Myanmar, where viral fake news and hate speech about the Rohingya Muslim minority escalated the country’s religious conflict into a full-blown genocide. Facebook admitted in 2018, after years of downplaying its role, that it had not done enough “to help prevent our platform from being used to foment division and incite offline violence.”

While Facebook may have been oblivious to these consequences in the beginning, it was studying them by 2016. In an internal presentation from that year, reviewed by the Wall Street Journal, a company researcher, Monica Lee, found that Facebook was not only hosting a large number of extremist groups but also promoting them to its users: “64% of all extremist group joins are due to our recommendation tools,” the presentation said, predominantly thanks to the models behind the “Groups You Should Join” and “Discover” features.

> “The question for leadership was: Should we be optimizing for engagement if you find that somebody is in a vulnerable state of mind?”
>
> *A former AI researcher who joined in 2018*

In 2017, Chris Cox, Facebook’s longtime chief product officer, formed a new task force to understand whether maximizing user engagement on Facebook was contributing to political polarization. It found that there was indeed a correlation, and that reducing polarization would mean taking a hit on engagement. In a mid-2018 document reviewed by the Journal, the task force proposed several potential fixes, such as tweaking the recommendation algorithms to suggest a more diverse range of groups for people to join. But it acknowledged that some of the ideas were “antigrowth.” Most of the proposals didn’t move forward, and the task force disbanded.

Since then, other employees have corroborated these findings. A former Facebook AI researcher who joined in 2018 says he and his team conducted “study after study” confirming the same basic idea: models that maximize engagement increase polarization. They could easily track how strongly users agreed or disagreed on different issues, what content they liked to engage with, and how their stances changed as a result. Regardless of the issue, the models learned to feed users increasingly extreme viewpoints. “Over time they *measurably* become more polarized,” he says.

The researcher's team also found that users with a tendency to post or engage with melancholy content—a possible sign of depression—could easily spiral into consuming increasingly negative material that risked further worsening their mental health. The team proposed tweaking the content-ranking models for these users to stop maximizing engagement alone, so they would be shown less of the depressing stuff. "The question for leadership was: Should we be optimizing for engagement if you find that somebody is in a vulnerable state of mind?" he remembers. (A Facebook spokesperson said she could not find documentation for this proposal.)

But anything that reduced engagement, even for reasons such as not exacerbating someone's depression, led to a lot of hemming and hawing among leadership. With their performance reviews and salaries tied to the successful completion of projects, employees quickly learned to drop those that received pushback and continue working on those dictated from the top down.

One such project heavily pushed by company leaders involved predicting whether a user might be at risk for something several people had already done: livestreaming their own suicide on Facebook Live. The task involved building a model to analyze the comments that other users were posting on a video after it had gone live, and bringing at-risk users to the attention of trained Facebook community reviewers who could call local emergency responders to perform a wellness check. It didn't require any changes to content-ranking models, had negligible impact on engagement, and effectively fended off negative press. It was also nearly impossible, says the researcher: "It's more of a PR stunt. The efficacy of trying to determine if somebody is going to kill themselves in the next 30 seconds, based on the first 10 seconds of video analysis—you're not going to be very effective."

Facebook disputes this characterization, saying the team that worked on this effort has since successfully predicted which users were at risk and increased the number of wellness checks performed. But the company does not release data on the accuracy of its predictions or how many wellness checks turned out to be real emergencies.

That former employee, meanwhile, no longer lets his daughter use Facebook.

Quiñonero should have been perfectly placed to tackle these problems when he created the SAIL (later Responsible AI) team in April 2018. His time as the director of Applied Machine Learning had made him intimately familiar with the company's algorithms, especially the ones used for recommending posts, ads, and other content to users.

It also seemed that Facebook was ready to take these problems seriously. Whereas previous efforts to work on them had been scattered across the company, Quiñonero was now being granted a centralized team with leeway in his mandate to work on whatever he saw fit at the intersection of AI and society.

At the time, Quiñonero was engaging in his own reeducation about how to be a responsible technologist. The field of AI research was paying growing attention to problems of AI bias and accountability in the wake of high-profile studies showing that, for example, an algorithm was scoring Black defendants as more likely to be rearrested than white defendants who'd been arrested for the same or a more serious offense. Quiñonero began studying the scientific literature on algorithmic fairness, reading books on ethical engineering and the history of technology, and speaking with civil rights experts and moral philosophers.

Over the many hours I spent with him, I could tell he took this seriously. He had joined Facebook amid the Arab Spring, a series of revolutions against oppressive Middle Eastern regimes. Experts had lauded social media for spreading the information that fueled the uprisings and giving people tools to organize. Born in Spain but raised in Morocco, where he'd seen the suppression of free speech firsthand, Quiñonero felt an intense connection to Facebook's potential as a force for good.

Six years later, Cambridge Analytica had threatened to overturn this promise. The controversy forced him to confront his faith in the company and examine what staying would mean for his integrity. "I think what happens to most people who work at Facebook—and definitely has been my story—is that there's no boundary between Facebook and me," he says. "It's extremely personal." But he chose to stay, and to head SAIL, because he believed he could do more for the world by helping turn the company around than by leaving it behind.

"I think if you're at a company like Facebook, especially over the last few years, you really realize the impact that your products have on people's lives—on what they think, how they communicate, how they interact with each other," says Quiñonero's longtime friend Zoubin Ghahramani, who helps lead the Google Brain team. "I know Joaquin cares deeply about all aspects of this. As somebody who strives to achieve better and improve things, he sees the important role that he can have in shaping both the thinking and the policies around responsible AI."

At first, SAIL had only five people, who came from different parts of the company but were all interested in the societal impact of algorithms. One founding member, Isabel Kloumann, a research scientist who'd come from the company's core data science team, brought with her an initial version of a tool to measure the bias in AI models.

The team also brainstormed many other ideas for projects. The former leader in the AI org, who was present for some of the early meetings of SAIL, recalls one proposal for combating polarization. It involved using sentiment analysis, a form of machine learning that interprets opinion in bits of text, to better identify comments that expressed extreme points of view. These comments wouldn't be deleted, but they would be hidden by default with an option to reveal them, thus limiting the number of people who saw them.

And there were discussions about what role SAIL could play within Facebook and how it should evolve over time. The sentiment was that the team would first produce responsible-AI guidelines to tell the product teams what they should or should not do. But the hope was that it would ultimately serve as the company's central hub for evaluating AI projects and stopping those that didn't follow the guidelines.

Former employees described, however, how hard it could be to get buy-in or financial support when the work didn't directly improve Facebook's growth. By its nature, the team was *not* thinking about growth, and in some cases it was proposing ideas antithetical to growth. As a result, it received few resources and languished. Many of its ideas stayed largely academic.

On August 29, 2018, that suddenly changed. In the ramp-up to the US midterm elections, President Donald Trump and other Republican leaders ratcheted up accusations that Facebook, Twitter, and Google had anti-conservative bias. They claimed that Facebook's moderators in particular, in applying the community standards, were suppressing conservative voices more than liberal ones. This charge would later be debunked, but the hashtag#StopTheBias, fueled by a Trump tweet, was rapidly spreading on social media.

For Trump, it was the latest effort to sow distrust in the country's mainstream information distribution channels. For Zuckerberg, it threatened to alienate Facebook's conservative US users and make the company more vulnerable to regulation from a Republican-led government. In other words, it threatened the company's growth.

Facebook did not grant me an interview with Zuckerberg, but previous reporting has shown how he increasingly pandered to Trump and the Republican leadership. After Trump was elected, Joel Kaplan, Facebook's VP of global public policy and its highest-ranking Republican, advised Zuckerberg to tread carefully in the new political environment.

On September 20, 2018, three weeks after Trump's #StopTheBias tweet, Zuckerberg held a meeting with Quiñonero for the first time since SAIL's creation. He wanted to know everything Quiñonero had learned about AI bias and how to quash it in Facebook's content-moderation models. By the end of the meeting, one thing was clear: AI bias was now Quiñonero's top priority. "The leadership has been very, very pushy about making sure we scale this aggressively," says Rachad Alao, the engineering director of Responsible AI who joined in April 2019.

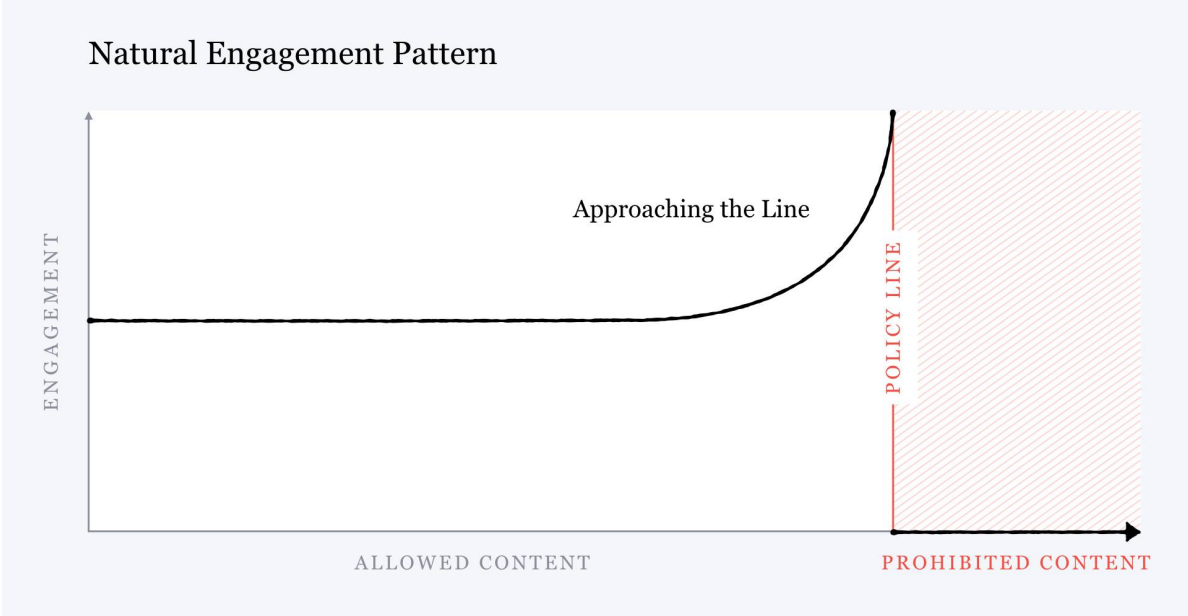
It was a win for everybody in the room. Zuckerberg got a way to ward off charges of anti-conservative bias. And Quiñonero now had more money and a bigger team to make the overall Facebook experience better for users. They could build upon Kloumann's existing tool in order to measure and correct the alleged anti-conservative bias in content-moderation models, as well as to correct other types of bias in the vast majority of models across the platform.

This could help prevent the platform from unintentionally discriminating against certain users. By then, Facebook already had thousands of models running concurrently, and almost none had been measured for bias. That would get it into legal trouble a few months later with the US Department of Housing and Urban Development (HUD), which alleged that the company's algorithms were inferring "protected" attributes like race from users' data and showing them ads for housing based on those attributes—an illegal form of discrimination. (The lawsuit is still pending.) Schroepfer also predicted that Congress would soon pass laws to regulate algorithmic discrimination , so Facebook needed to make headway on these efforts anyway.

(Facebook disputes the idea that it pursued its work on AI bias to protect growth or in anticipation of regulation. "We built the Responsible AI team because it was the right thing to do," a spokesperson said.)

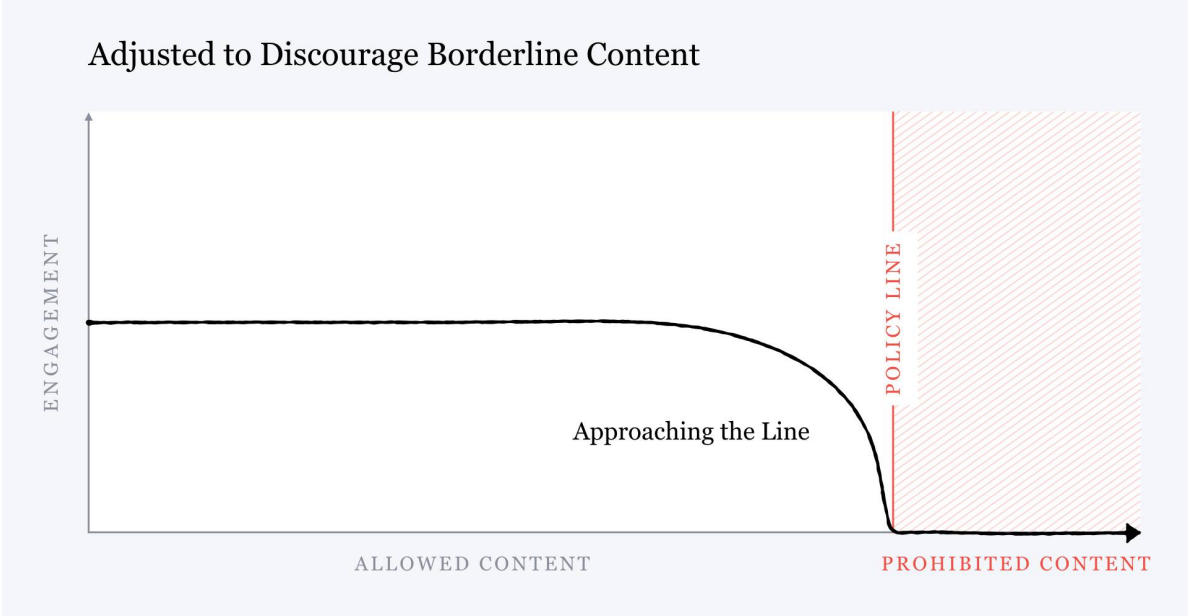
But narrowing SAIL's focus to algorithmic fairness would sideline all Facebook's other long-standing algorithmic problems. Its content-recommendation models would continue pushing posts, news, and groups to users in an effort to maximize engagement, rewarding extremist content and contributing to increasingly fractured political discourse.

Zuckerberg even admitted this. Two months after the meeting with Quiñonero, in a public note outlining Facebook's plans for content moderation, he illustrated the harmful effects of the company's engagement strategy with a simplified chart. It showed that the more likely a post is to violate Facebook's community standards, the more user engagement it receives, because the algorithms that maximize engagement reward inflammatory content.



A chart titled "natural engagement pattern" that shows allowed content on the X axis, engagement on the Y axis, and an exponential increase in engagement as content nears the policy line for prohibited content.

But then he showed another chart with the inverse relationship. Rather than rewarding content that came close to violating the community standards, Zuckerberg wrote, Facebook could choose to start "penalizing" it, giving it "less distribution and engagement" rather than more. How would this be done? With more AI. Facebook would develop better content-moderation models to detect this "borderline content" so it could be retroactively pushed lower in the news feed to snuff out its virality, he said.



A chart titled "adjusted to discourage borderline content" that shows the same chart but the curve inverted to reach no engagement when it reaches the policy line.

The problem is that for all Zuckerberg's promises, this strategy is tenuous at best.

Misinformation and hate speech constantly evolve. New falsehoods spring up; new people and groups become targets. To catch things before they go viral, content-moderation models must be able to identify new unwanted content with high accuracy. But machine-learning models do not work that way. An algorithm that has learned to recognize Holocaust denial can't immediately spot, say, Rohingya genocide denial. It must be trained on thousands, often even millions, of examples of a new type of content before learning to filter it out. Even then, users can quickly learn to outwit the model by doing things like changing the wording of a post or replacing incendiary phrases with euphemisms, making their message illegible to the AI while still obvious to a human. This is why new conspiracy theories can rapidly spiral out of control, and partly why, even after such content is banned, forms of it can persist on the platform.

In his New York Times profile, Schroepfer named these limitations of the company's content-moderation strategy. "Every time Mr. Schroepfer and his more than 150 engineering specialists create A.I. solutions that flag and squelch noxious material, new and dubious posts that the A.I. systems have never seen before pop up—and are thus not caught," wrote the Times. "It's never going to go to zero," Schroepfer told the publication.

Meanwhile, the algorithms that recommend this content still work to maximize engagement. This means every toxic post that escapes the content-moderation filters will continue to be pushed higher up the news feed and promoted to reach a larger audience. Indeed, a study from New York University recently found that among partisan publishers' Facebook pages, those that regularly posted political misinformation received the most engagement in the lead-up to the 2020 US presidential election and the Capitol riots. "That just kind of got me," says a former employee who worked on integrity issues from 2018 to 2019. "We fully acknowledged [this], and yet we're still increasing engagement."

But Quiñonero's SAIL team wasn't working on this problem. Because of Kaplan's and Zuckerberg's worries about alienating conservatives, the team stayed focused on bias. And even after it merged into the bigger Responsible AI team, it was never mandated to work on content-recommendation systems that might limit the spread of misinformation. Nor has any other team, as I confirmed after Entin and another spokesperson gave me a full list of all Facebook's other initiatives on integrity issues—the company's umbrella term for problems including misinformation, hate speech, and polarization.

A Facebook spokesperson said, "The work isn't done by one specific team because that's not how the company operates." It is instead distributed among the teams that have the specific expertise to tackle how content ranking affects misinformation for their part of the platform, she said. But Schroepfer told me precisely the opposite in an earlier interview. I had asked him why he had created a centralized Responsible AI team instead of directing existing teams to make progress on the issue. He said it was "best practice" at the company.

"[If] it's an important area, we need to move fast on it, it's not well-defined, [we create] a dedicated team and get the right leadership," he said. "As an area grows and matures, you'll see the product teams take on more work, but the central team is still needed because you need to stay up with state-of-the-art work."

When I described the Responsible AI team's work to other experts on AI ethics and human rights, they noted the incongruity between the problems it was tackling and those, like misinformation, for which Facebook is most notorious. "This seems to be so oddly removed from Facebook as a product—the things Facebook builds and the questions about impact on the world that Facebook

faces," said Rumman Chowdhury, whose startup, Parity, advises firms on the responsible use of AI, and was acquired by Twitter after our interview. I had shown Chowdhury the Quiñonero team's documentation detailing its work. "I find it surprising that we're going to talk about inclusivity, fairness, equity, and not talk about the very real issues happening today," she said.

"It seems like the 'responsible AI' framing is completely subjective to what a company decides it wants to care about. It's like, 'We'll make up the terms and then we'll follow them,'" says Ellery Roberts Biddle, the editorial director of Ranking Digital Rights, a nonprofit that studies the impact of tech companies on human rights. "I don't even understand what they mean when they talk about fairness. Do they think it's fair to recommend that people join extremist groups, like the ones that stormed the Capitol? If everyone gets the recommendation, does that mean it was fair?"

"We're at a place where there's one genocide [Myanmar] that the UN has, with a lot of evidence, been able to specifically point to Facebook and to the way that the platform promotes content," Biddle adds. "How much higher can the stakes get?"

Over the last two years, Quiñonero's team has built out Kloumann's original tool, called Fairness Flow. It allows engineers to measure the accuracy of machine-learning models for different user groups. They can compare a face-detection model's accuracy across different ages, genders, and skin tones, or a speech-recognition algorithm's accuracy across different languages, dialects, and accents.

Fairness Flow also comes with a set of guidelines to help engineers understand what it means to train a "fair" model. One of the thornier problems with making algorithms fair is that there are different definitions of fairness, which can be mutually incompatible. Fairness Flow lists four definitions that engineers can use according to which suits their purpose best, such as whether a speech-recognition model recognizes all accents with equal accuracy or with a minimum threshold of accuracy.

But testing algorithms for fairness is still largely optional at Facebook. None of the teams that work directly on Facebook's news feed, ad service, or other products are required to do it. Pay incentives are still tied to engagement and growth metrics. And while there are guidelines about which fairness definition to use in any given situation, they aren't enforced.

This last problem came to the fore when the company had to deal with allegations of anti-conservative bias.

In 2014, Kaplan was promoted from US policy head to global vice president for policy, and he began playing a more heavy-handed role in content moderation and decisions about how to rank posts in users' news feeds. After Republicans started voicing claims of anti-conservative bias in 2016, his team began manually reviewing the impact of misinformation-detection models on users to ensure—among other things—that they didn't disproportionately penalize conservatives.

All Facebook users have some 200 "traits" attached to their profile. These include various dimensions submitted by users or estimated by machine-learning models, such as race, political and religious leanings, socioeconomic class, and level of education. Kaplan's team began using the traits to assemble custom user segments that reflected largely conservative interests: users who engaged with conservative content, groups, and pages, for example. Then they'd run special analyses to see how content-moderation decisions would affect posts from those segments, according to a former researcher whose work was subject to those reviews.

The Fairness Flow documentation, which the Responsible AI team wrote later, includes a case study on how to use the tool in such a situation. When deciding whether a misinformation model is fair with respect to political ideology, the team wrote, "fairness" does *not* mean the model should affect conservative and liberal users equally. If conservatives are posting a greater fraction

of misinformation, as judged by public consensus, then the model should flag a greater fraction of conservative content. If liberals are posting more misinformation, it should flag their content more often too.

But members of Kaplan's team followed exactly the opposite approach: they took "fairness" to mean that these models should not affect conservatives more than liberals. When a model did so, they would stop its deployment and demand a change. Once, they blocked a medical-misinformation detector that had noticeably reduced the reach of anti-vaccine campaigns, the former researcher told me. They told the researchers that the model could not be deployed until the team fixed this discrepancy. But that effectively made the model meaningless. "There's no point, then," the researcher says. A model modified in that way "would have literally no impact on the actual problem" of misinformation.

> "I don't even understand what they mean when they talk about fairness. Do they think it's fair to recommend that people join extremist groups, like the ones that stormed the Capitol? If everyone gets the recommendation, does that mean it was fair?"
>
> *Ellery Roberts Biddle, editorial director of Ranking Digital Rights*

This happened countless other times—and not just for content moderation. In 2020, the Washington Post reported that Kaplan's team had undermined efforts to mitigate election interference and polarization within Facebook, saying they could contribute to anti-conservative bias. In 2018, it used the same argument to shelve a project to edit Facebook's recommendation models even though researchers believed it would reduce divisiveness on the platform, according to the Wall Street Journal. His claims about political bias also weakened a proposal to edit the ranking models for the news feed that Facebook's data scientists believed would strengthen the platform against the manipulation tactics Russia had used during the 2016 US election.

And ahead of the 2020 election, Facebook policy executives used this excuse, according to the New York Times, to veto or weaken several proposals that would have reduced the spread of hateful and damaging content.

Facebook disputed the Wall Street Journal's reporting in a follow-up blog post, and challenged the New York Times's characterization in an interview with the publication. A spokesperson for Kaplan's team also denied to me that this was a pattern of behavior, saying the cases reported by the Post, the Journal, and the Times were "all individual instances that we believe are then mischaracterized." He declined to comment about the retraining of misinformation models on the record.

Many of these incidents happened before Fairness Flow was adopted. But they show how Facebook's pursuit of fairness in the service of growth had already come at a steep cost to progress on the platform's other challenges. And if engineers used the definition of fairness that Kaplan's team had adopted, Fairness Flow could simply systematize behavior that rewarded misinformation instead of helping to combat it.

Often "the whole fairness thing" came into play only as a convenient way to maintain the status quo, the former researcher says: "It seems to fly in the face of the things that Mark was saying publicly in terms of being fair and equitable."

The last time I spoke with Quiñonero was a month after the US Capitol riots. I wanted to know how the storming of Congress had affected his thinking and the direction of his work.

In the video call, it was as it always was: Quiñonero dialing in from his home office in one window and Entin, his PR handler, in another. I asked Quiñonero what role he felt Facebook had played in the riots and whether it changed the task he saw for Responsible AI. After a long pause, he sidestepped the question, launching into a description of recent work he'd done to promote greater diversity and inclusion among the AI teams.

I asked him the question again. His Facebook Portal camera, which uses computer-vision algorithms to track the speaker, began to slowly zoom in on his face as he grew still. "I don't know that I have an easy answer to that question, Karen," he said. "It's an extremely difficult question to ask me."

Entin, who'd been rapidly pacing with a stoic poker face, grabbed a red stress ball.

I asked Quiñonero why his team hadn't previously looked at ways to edit Facebook's content-ranking models to tamp down misinformation and extremism. He told me it was the job of other teams (though none, as I confirmed, have been mandated to work on that task). "It's not feasible for the Responsible AI team to study all those things ourselves," he said. When I asked whether he would consider having his team tackle those issues in the future, he vaguely admitted, "I would agree with you that that is going to be the scope of these types of conversations."

Near the end of our hour-long interview, he began to emphasize that AI was often unfairly painted as "the culprit." Regardless of whether Facebook used AI or not, he said, people would still spew lies and hate speech, and that content would still spread across the platform.

I pressed him one more time. Certainly he couldn't believe that algorithms had done absolutely nothing to change the nature of these issues, I said.

"I don't know," he said with a halting stutter. Then he repeated, with more conviction: "That's my honest answer. Honest to God. I don't know."