AI

Behind the scenes: How Twitter decided to open up its image-cropping algorithm to the public

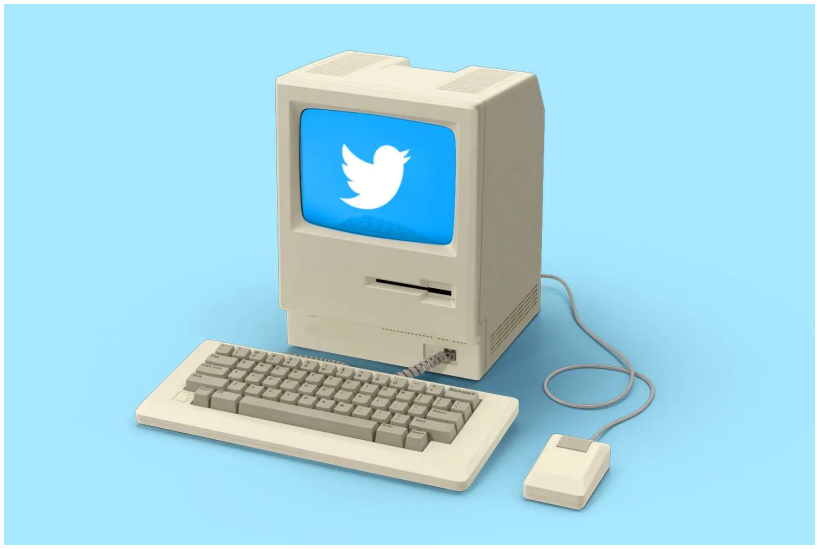
Its AI ethics team held a public competition to find bias in the algorithm, a rare act of transparency among major tech firms.

SEPTEMBER 27, 2021 · 12 MIN READ

HAYDEN FIELD
EMERGING TECH REPORTER

Follow

COPY



Francis Scialabba

On Monday, September 21, 2020, Kyra Yee, a research engineer on Twitter's Machine Learning Ethics Transparency and Accountability (META) team, rolled out of bed and opened her laptop to an influx of messages from her colleagues. When she logged onto Twitter, she immediately saw what had sparked the conversation.

Three days earlier, on Friday night, PhD student Colin Madland had tweeted about how Twitter's image-cropping algorithm favored his face—Madland is white—over that of his Black colleague. The thread had exploded over the weekend into a full-blown controversy and sparked a deluge of citizen science, as users tested how the algorithm cropped people of different races, genders, ages,

and more. To date, the most popular tweets investigating the controversy have hundreds of thousands of likes and retweets.

Ten months later, Twitter decided to open up the same image-cropping algorithm to more formal scrutiny.

In late July, the company announced it'd hold the first-ever "algorithmic bias bounty challenge," a spin on cybersecurity-focused bug bounty programs, a common way for tech companies to reward individuals who flag security flaws and site vulnerabilities. Between July 30 and August 6, anyone could peruse the code underpinning Twitter's image-cropping algorithm, then submit a bias assessment for the chance to win one of five cash prizes, ranging from $500 to $3,500.

It was a high-stakes—and unusual—move in the Big Tech stratosphere. It's rare for tech companies to allow direct, public engagement with the algorithms that govern their platforms: In the months before, Google fired the co-leads of its AI ethics team after their research on the dangers of large language models (though Google disputes the reasoning), for instance, and Facebook shut down research on the polarizing effects of its News Feed algorithm. And although the algorithm Twitter had put on public display had been partly decommissioned three months before the competition, it had been deployed since January 2018 to all ~200 million of Twitter's daily users. The open invitation for external audits was certain to unearth additional skeletons.

With bad press from Madland's initial exposé still fresh on the company's mind, how did Twitter's META team convince the company to greenlight the project?

"There's a moment, and then an idea happens, and you have to actually be ready—like the world has to be ready, in a sense," Rumman Chowdhury, director of the META team and a driving force behind the bias bounty, told us. "In our cases, the field had to be ready. I think three years ago…if someone launched a bias bounty, no one would've understood what to do with it. We needed to be at a place where it *needed* to happen."

Viral for the wrong reasons

After Madland's tweets in September 2020, one keyword kept coming up: saliency.

"I just remember being…half-asleep and being like, 'What the heck is saliency?'" Yee told us.

Saliency algorithms are…ideal…seen as a shortcut for identifying

Saliency algorithms are widely seen as a shortcut for identifying the most important or relevant aspects of a photo. It's also the type of model that Twitter used to crop images for nearly 3.5 years, until scaling down the practice in May.

"We have taken steps to decommission the algorithm, but it's a work in progress," Lindsay McCallum, a senior communications manager at Twitter, told us. "Standard aspect ratio photos now appear uncropped on the Home Timeline on mobile and we are working on further improvements, including bringing uncropped photos to Twitter. com, to decrease our dependence on the saliency algorithm."

Saliency algorithms are trained via human eye-tracking—e.g., where someone's eyes go first when shown an image—and the subjects are often students on college campuses.

"It's…multiplying, 100 million times, an unconscious bias," META product lead Jutta Williams told us. "Then we all use it as transferred learning, and we all apply it in all these crazy ways. This is the worst kind of algorithm, because there is no intention to make it fair or equitable or not objectifying or non-racist or non-beauty-biased. It's just human eye-tracking."

Sure enough, when the team investigated over a four-month period, it found clear evidence of the biases flagged by Madland and other Twitter users: The algorithm tended to favor women over men and white people over Black people, including comparisons by gender demographic. While this research was being conducted and refined, Chowdhury—as well as Williams—joined the META team.

Chowdhury had interviewed with Twitter around October 2020, and she had been impressed by the company's reaction to citizen data science.

"People go out there and they do this of their own volition and they don't get rewarded for this work, and companies often benefit because they're learning about the problems that exist in their systems—and they're essentially getting free labor," Chowdhury said. "I saw that happening on Twitter, and [you] saw the company respond really positively and say, 'Oh, we're going to go…do something about it, here's how we're correcting it.'"

After eight months, the META team finally published their investigation results and partly shut down the saliency algorithm. But there was still more to do.

"[With] that paper, I think we did a decent job, like a good job, but

it's pretty limited in scope," Yee said. "The problems it covers are limited, like the discussion of race is basically just very overly simplified to Black and white, and there's a bunch of other things that could potentially be going on."

In an ideal world, Yee hoped they could create something that would allow the community to formally submit algorithm concerns that weren't on META's radar.

Building a bias bounty

Chowdhury got the idea for the contest while attending DEF CON—known as the hacking world's oldest, largest conference—three years prior. Twitter was the first major tech company where she felt leadership might buy into the idea, she told us—and after meeting with CTO Parag Agrawal, she got the green light to proceed. The next step was running the idea by the rest of the META team, which would have to do the heavy lifting. Fast.

Chowdhury had received an email about submitting ideas to AI Village, a community of hackers and data scientists that would be part of DEF CON 2021. It was June 16, they needed a yes or no answer by July 16, and the conference was scheduled for the first week of August—giving the team about four weeks to plan and execute the first-ever contest of its kind.

Once Williams, Yee, and Irene Font Peradejordi—a researcher who had joined the team a few weeks before—were on board, they could move forward.

"When Rumman and Jutta came to us, well, I was very confused," Peradejordi recalled. "I was like, 'All right, I've never seen this before. I've never heard this before'....But then I started researching a little bit, and I was like, 'This idea is brilliant.' So it's really amazing because it's like a new way that no one [has] tried before to get feedback from users, and as a user researcher, that's my passion."

And after Chowdhury and Williams met with CEO Jack Dorsey in between his meetings to seal the multimillion-dollar purchase of Revue, he tweeted his own version of a green light. But turning the idea into reality still meant cutting through a lot of red tape—and according to multiple META team members, Williams "made…administrative walls disappear," as Yee put it.

Those walls included navigating potential issues with the legal team, information security team, marketing teams, and even budget. META had to make sure the participation agreement's legal language allowed anyone anywhere to participate; that the bias bounty's partner vendor, HackerOne, had a viable contract; and that the challenge was marketed differently than Twitter's InfoSec bug bounties, so as not to confuse anyone. Sharing IP and trade secrets was a concern, too. Williams said reminding people that Agrawal was on board helped dissolve bottlenecks, and since Twitter's image-cropping algorithm was already open source, that helped assuage concerns about trade secrets.

But even with the official go-ahead, the team had a complicated problem to solve: How do you create a grading rubric for a competition focused on the real-world harms caused by an algorithm?

"For traditional bug bounties, right, it's pretty simple to decide who wins money and stuff because it's pretty black-and-white," Yee said. "This is either a vulnerability or it's not; you can either hack into something or you can't. But for bias and harms…different people think different things are harmful, and different people have different definitions of fairness. "

To get external opinions on the rubric, the META team enlisted Twitter's own legal and risk teams and the four judges who would eventually name the contest winners: Ariel Herbert-Voss, a research scientist at OpenAI and cofounder of AI Village; Matt Mitchell, a tech fellow at the Ford Foundation and founder of cryptography nonprofit CryptoHarlem; Peiter "Mudge" Zatko, a veteran hacker who joined Twitter as head of security in 2020; and Patrick Hall, principal scientist at AI-focused law firm Bnh.ai.

The Algorithmic Justice League (AJL), an advocacy organization founded in 2016, watched the contest's rollout closely.

"We think that Twitter has been very thoughtful about the criteria," Sasha Costanza-Chock, AJL's director of research and design, told us. "Some really interesting innovations in the scoring system include attention to the sociotechnical context…basically,

you can get more points if you clearly explain why the bias that you identified is…going to harm communities. I think that they also included extra points if your analysis crosses more than just one marginalized community, which was nice to see."

But there's still room for improvement, according to the AJL—especially in the decision to grant additional points based on the number of people the issue may affect.

"Thinking about the way that bias and harms in algorithmic systems can play out, it's often—not always, but often—smaller groups of people, maybe people who are multiply marginalized by systems of race, class, gender, disability, and so on, who sometimes can suffer the worst harms from algorithmic systems," Costanza-Chock said. "As a trans person, I'm from a community that's a very small percentage of the population, and so the scoring mechanism that rewards you for focusing on bigger segments of the population would de-incentivize researchers from focusing on bias problems that might affect trans folks. And…that's just one example."

Pictured: A screenshot of META team members and judges of the bias bounty competition; Source: Twitter

An around-the-clock weekend

When Williams fell asleep on Friday, August 6—hours before the bias bounty portal closed at midnight—they had received eight submissions. She awoke the next morning to four times as many.

The team had less than 48 hours to read, analyze, and grade them all.

Yee had torn her Achilles tendon days before, so she graded submissions from bed. Peradejordi concentrated so hard on reading entries that she forgot to eat. Chowdhury was juggling her frazzled cousin's questions about how to handle his hyper-energetic new Bengal kitten. And while Williams helped input everything into an Excel spreadsheet, her husband regularly deposited snacks by her side to stave off her hangry side.

Finally, the entries were passed along to the judges, who hashed things out. First prize went to Bogdan Kulynych, a researcher at Switzerland's EPFL technical university, who found that the algorithm prioritized young, slim, and lighter-skinned faces. Other winners found that the algorithm rarely chose to feature people pictured in wheelchairs, that it preferred emojis with lighter skin tones, and that for memes, it was more likely to showcase English text than Arabic script.

There were points of conflict, Williams recalled; not only did each judge have their own point of view, but members of the META team were also passionate about different entries that weren't slated to win. So they created a spur-of-the-moment "honorable mentions" category—for certain submissions that didn't score high enough via the rubric, but highlighted important findings.

In the future, Chowdhury told us, she sees this transforming from a competition into an ongoing vulnerability rewards program at Twitter. But, she acknowledged, that would require a lot of investment. It would also require a strong corporate stomach: The competition did indeed unearth more skeletons, and headlines tended to focus on that.

One, from NBC News, read: "Twitter's racist algorithm is also ageist, ableist and Islamophobic, researchers find." And another,

from The Guardian: “Student proves Twitter algorithm ‘bias’ toward lighter, slimmer, younger faces.”

The AJL said it’s enthusiastic about any sort of initiative to shine a light on bias and harms in AI systems, especially ones that encourage researchers to explore these consequences instead of shutting down their efforts. But members also emphasized that bounties are a small step forward.

In order to create a “robust, equitable, and accountable life cycle for AI development,” Costanza-Chock said, companies also need internal bias and harms teams to vet every step of an AI system’s development, from scoping to deployment. Also necessary, according to Costanza-Chock: significant regulatory oversight, external researchers, and third-party algorithmic auditors. It’s already been demonstrated in cybersecurity that on their own, bounties aren’t enough to tip the scales, and “the same holds true in the algorithmic bias and harms space,” they said.

After all, tech giants aren’t currently under any obligation in the US to draw back the curtains on their algorithms—Twitter’s bias bounty hinged on executives choosing to move forward with it.

You might also like...

WORK LIFE

Slack: its pros, cons, and moral reckoning

MALIAH WEST / 10.14.2021

AI

Report: US AI development is concentrated in 15 metro areas

HAYDEN FIELD / 09.13.2021

CONNECTIVITY

Smart grids could soften the blow of cyberattacks, but make them more common

JORDAN MCDONALD / 09.20.2021