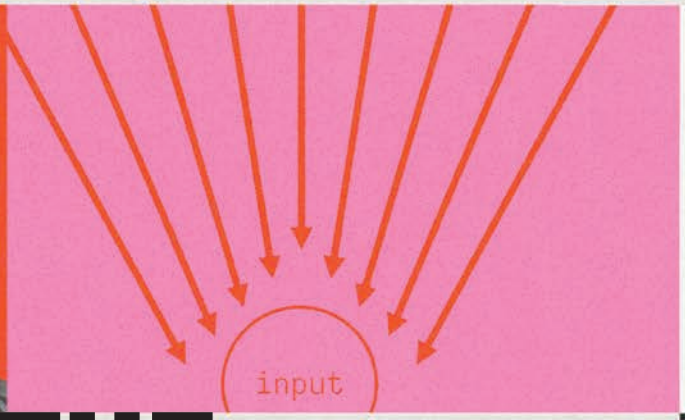
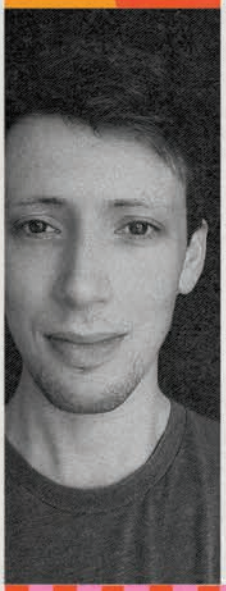


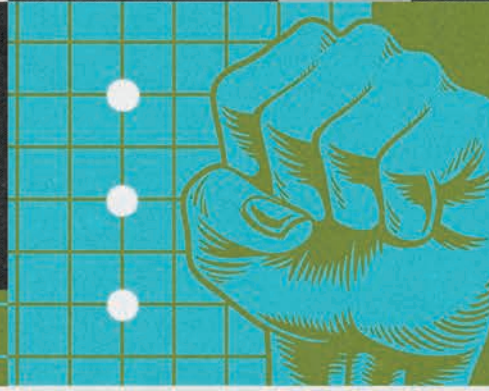
THE FIGHT TO RECLAIM AI

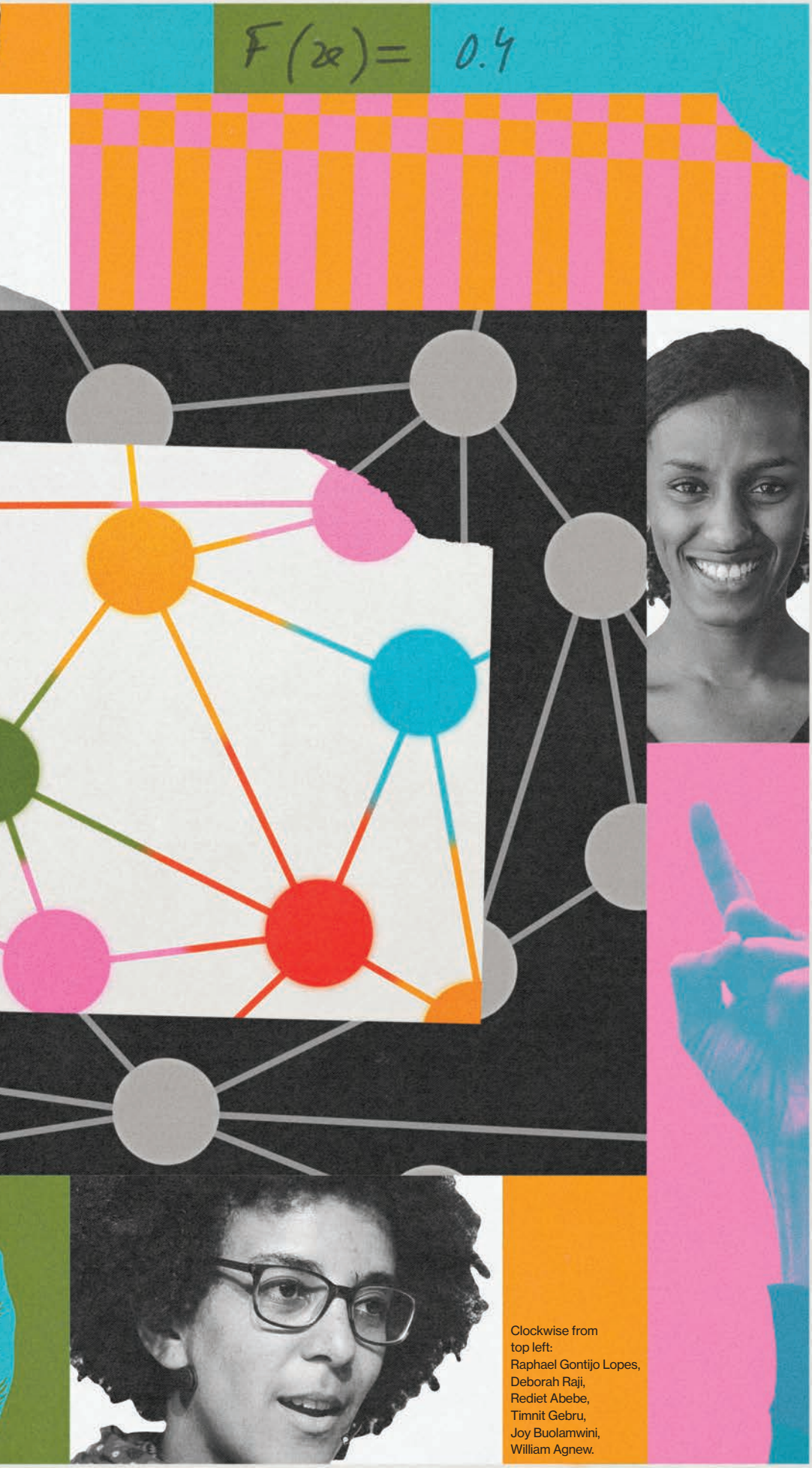
How a small group of researchers at the margins built a movement to wrest the world's most powerful emerging technology from Big Tech's control.



By Karen Hao

ILLUSTRATION BY RICARDO SANTOS





Clockwise from top left: Raphael Gontijo Lopes, Deborah Raji, Rediet Abebe, Timnit Gebru, Joy Buolamwini, William Agnew.

Timnit Gebru never thought a scientific paper would cause her so much trouble.

In 2020, as the co-lead of Google's ethical AI team, Gebru had reached out to Emily Bender, a linguistics professor at the University of Washington, and asked to collaborate on research about the troubling direction of artificial intelligence. Gebru wanted to identify the risks posed by large language models, one of the most stunning recent breakthroughs in AI research. The models are algorithms trained on staggering amounts of text. Under the right conditions, they can compose what look like convincing passages of prose.

For a few years, tech companies had been racing to build bigger versions and integrate them into consumer products. Google, which invented the technique, was already using one to improve the relevance of search results. OpenAI announced the largest one, called GPT-3, in June 2020 and licensed it exclusively to Microsoft a few months later.

Gebru worried about how fast the technology was being deployed. In the paper she wound up writing with Bender and five others, she detailed the possible dangers. The models were enormously costly to create—both environmentally (they require huge amounts of computational power) and financially; they were often trained on the toxic and abusive language of the internet; and they'd come to dominate research in language AI, elbowing out promising alternatives.

Like other existing AI techniques, the models don't actually understand language. But because they can manipulate it to retrieve text-based information for users or generate natural conversation, they can be packaged into products and services that make tech companies lots of money.

That November, Gebru submitted the paper to a conference. Soon after, Google executives asked her to retract it, and when she refused, they fired her. Two months later, they also fired her coauthor Margaret Mitchell, the other leader of the ethical AI team.

The dismantling of that team sparked one of the largest controversies within the AI world in recent memory. Defenders of Google argued that the company has the right to supervise its own researchers. But for many others, it solidified fears about the degree of control that tech giants now have over the field. Big Tech is now the primary employer and funder of AI researchers, including, somewhat ironically, many of those who assess its social impacts.

Among the world's richest and most powerful companies, Google, Facebook, Amazon, Microsoft, and Apple have made AI core parts of their business. Advances over the last decade, particularly in an AI technique called deep learning, have allowed them to monitor users' behavior; recommend news, information, and products to them; and most of all, target them with ads. Last year Google's advertising apparatus generated over $140 billion in revenue. Facebook's generated $84 billion.

The companies have invested heavily in the technology that has brought them such vast wealth. Google's parent company, Alphabet, acquired the London-based AI lab DeepMind for $600 million in 2014 and spends hundreds of millions a year to support its research. Microsoft signed a $1 billion deal with OpenAI in 2019 for commercialization rights to its algorithms.

At the same time, tech giants have become large investors in university-based AI research, heavily influencing its scientific priorities. Over the years, more and more ambitious scientists have transitioned to working for tech giants full time or adopted a dual affiliation. From 2018 to 2019, 58% of the most cited papers at the top two AI conferences had at least one author affiliated with a tech giant, compared with only 11% a decade earlier, according to a study by researchers in the Radical AI Network, a group that seeks to challenge power dynamics in AI.

The problem is that the corporate agenda for AI has focused on techniques with commercial potential, largely ignoring research that could help address challenges like economic inequality and climate change. In fact, it has made these challenges worse. The drive to automate tasks has cost jobs and led to the rise of tedious labor like data cleaning and content moderation. The push to create ever larger models has caused AI's energy consumption to explode. Deep learning has also created a culture in which our data is constantly scraped, often without consent, to train products like facial recognition systems. And recommendation algorithms have exacerbated political polarization, while large language models have failed to clean up misinformation.

It's this situation that Gebru and a growing movement of like-minded scholars want to change. Over the last five years, they've sought to shift the field's priorities away from simply enriching tech companies, by expanding who gets to participate in developing the technology. Their goal is not only to mitigate the harms caused by existing systems but to create a new, more equitable and democratic AI.

"HELLO FROM TIMNIT"

In December 2015, Gebru sat down to pen an open letter. Halfway through her PhD at Stanford, she'd attended the Neural Information Processing Systems conference, the largest annual AI research gathering. Of the more than 3,700 researchers there, Gebru counted only five who were Black.

Once a small meeting about a niche academic subject, NeurIPS (as it's now known) was quickly becoming the biggest annual AI job bonanza. The world's wealthiest companies were coming to show off demos, throw extravagant parties, and write hefty checks for the rarest people in Silicon Valley: skillful AI researchers.

That year Elon Musk arrived to announce the nonprofit venture OpenAI. He, Y Combinator's then president Sam Altman, and PayPal cofounder Peter Thiel had put up $1 billion to solve what they believed to be an existential problem: the prospect that a superintelligence could one day take over the world. Their solution: build an even better superintelligence. Of the 14 advisors or technical team members he anointed, 11 were white men.

While Musk was being lionized, Gebru was dealing with humiliation and harassment. At a conference party, a group of drunk guys in Google Research T-shirts circled her and subjected her to unwanted hugs, a kiss on the cheek, and a photo.

Gebru typed out a scathing critique of what she had observed: the spectacle, the cult-like worship of AI celebrities, and most of all, the overwhelming homogeneity. This boy's club culture, she wrote, had already pushed talented women out of the field. It was also leading the entire community toward a dangerously narrow conception of artificial intelligence and its impact on the world.

Google had already deployed a computer-vision algorithm that classified Black people as gorillas, she noted. And the increasing sophistication of unmanned drones was putting the US military on a path toward lethal autonomous weapons. But there was no mention of these issues in Musk's grand plan to stop AI from taking over the world in some theoretical future scenario. "We don't have to project into the future to see AI's potential adverse effects," Gebru wrote. "It is already happening."

Gebru never published her reflection. But she realized that something needed to change. On January 28, 2016, she sent an email with the subject line "Hello from Timnit" to five other Black AI researchers. "I've always been sad by the lack of color in AI," she wrote. "But now I have seen 5 of you :) and thought that it would be cool if we started a black in AI group or at least know of each other."

The email prompted a discussion. What was it about being Black that informed their research? For Gebru, her work was very much a product of her identity; for others, it was not. But after meeting they agreed: If AI was going to play a bigger role in society, they needed more Black researchers. Otherwise, the field would produce weaker science—and its adverse consequences could get far worse.

A PROFIT-DRIVEN AGENDA

As Black in AI was just beginning to coalesce, AI was hitting its commercial stride. That year, 2016, tech giants spent an estimated $20 to $30 billion

"WE DON'T HAVE TO PROJECT INTO THE FUTURE TO SEE AI'S POTENTIAL ADVERSE EFFECTS."

on developing the technology, according to the McKinsey Global Institute.

Heated by corporate investment, the field warped. Thousands more researchers began studying AI, but they mostly wanted to work on deep-learning algorithms, such as the ones behind large language models. "As a young PhD student who wants to get a job at a tech company, you realize that tech companies are all about deep learning," says Suresh Venkatasubramanian, a computer science professor who now serves at the White House Office of Science and Technology Policy. "So you shift all your research to deep learning. Then the next PhD student coming in looks around and says, 'Everyone's doing deep learning. I should probably do it too.'"

But deep learning isn't the only technique in the field. Before its boom, there was a different AI approach known as symbolic reasoning. Whereas deep learning uses massive amounts of data to teach algorithms about meaningful relationships in information, symbolic reasoning focuses on explicitly encoding knowledge and logic based on human expertise.

Some researchers now believe those techniques should be combined. The hybrid approach would make AI more efficient in its use of data and energy, and give it the knowledge and reasoning abilities of an expert as well as the capacity to update itself with new information. But companies have little incentive to explore alternative approaches when the surest way to maximize their profits is to build ever bigger models.

In their paper, Gebru and Bender alluded to a basic cost of this tendency to stick with deep learning: the more advanced AI systems we need are not being developed, and similar problems keep recurring. Facebook, for example, relies heavily on large language models for automated content moderation. But without really understanding the meaning behind text, those models often fail. They regularly take down innocuous posts while giving hate speech and misinformation a pass.

AI-based facial recognition systems suffer from the same issue. They're trained on massive amounts of data but see only pixel patterns—they do not have a grasp of visual concepts like eyes, mouths, and noses. That can trip these systems up when they're used on individuals with a different skin tone from the people they were shown during training. Nonetheless, Amazon and other companies have sold these systems to law enforcement. In the US, they have caused three known cases of police jailing the wrong person—all Black men—in the last year.

For years, many in the AI community largely acquiesced to Big Tech's role in shaping the development and impact of these technologies. While some expressed discomfort with the corporate takeover, many more welcomed the industry's deep well of funding.

But as the shortcomings of today's AI have become more evident—both its failure to solve social problems and the mounting examples that it can exacerbate them—faith in Big Tech has weakened. Google's ousting of Gebru and Mitchell further stoked the discussion by revealing just how much companies will prioritize profit over self-policing.

In the immediate aftermath, over 2,600 Google employees and 4,300 others signed a petition denouncing Gebru's dismissal as "unprecedented research censorship." Half a year later, research groups are still rejecting the company's funding, researchers refuse to participate in its conference workshops, and employees are leaving in protest.

Unlike five years ago, when Gebru began raising these questions, there's now a well-established movement questioning what AI should be and who it should serve. This isn't a coincidence. It's very much a product of Gebru's own initiative, which began with the simple act of inviting more Black researchers into the field.

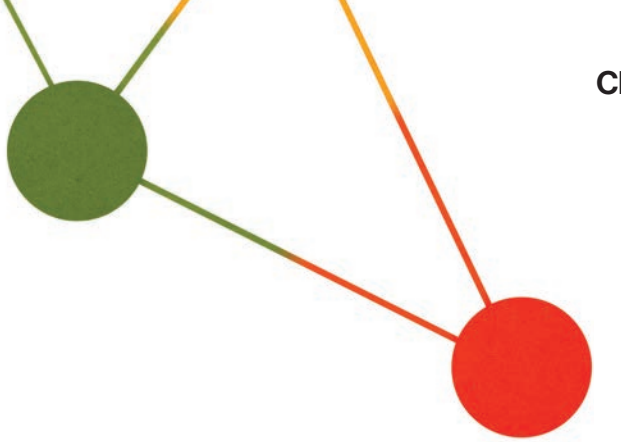
IT TAKES A CONFERENCE

In December 2017, the new Black in AI group hosted its first workshop at NeurIPS. While organizing the workshop, Gebru approached Joy Buolamwini, an MIT Media Lab researcher who was studying commercial facial recognition systems for possible bias. Buolamwini had begun testing these systems after one failed to detect her own face unless she donned a white mask. She submitted her preliminary results to the workshop.

Deborah Raji, then an undergraduate researcher, was another early participant. Raji was appalled by the culture she'd observed at NeurIPS. The workshop became her respite. "To go from four or five days of that to a full day of people that look like me talking about succeeding in this space—it was such important encouragement for me," she says.

Buolamwini, Raji, and Gebru would go on to work together on a pair of groundbreaking studies about discriminatory computer-vision systems. Buolamwini and Gebru coauthored Gender Shades, which showed that the facial recognition systems sold by Microsoft, IBM, and Chinese tech giant Megvii had remarkably high failure rates on Black women despite near-perfect performance on white men. Raji and Buolamwini then collaborated on a follow-up called Actionable Auditing, which found the same to be true for Amazon's Rekognition. In 2020, Amazon would agree to a one-year moratorium on police sales of its product, in part because of that work.

At the very first Black in AI workshop, though, these successes were distant possibilities. There was no agenda other than to build community and produce research based on their sorely lacking perspectives. Many onlookers didn't understand why such a group needed to exist.



"THESE CHANGES THAT WE'RE FIGHTING FOR—IT'S NOT JUST FOR MARGINALIZED GROUPS."

Gebru remembers dismissive comments from some in the AI community. But for others, Black in AI pointed a new way forward.

This was true for William Agnew and Raphael Gontijo Lopes, both queer men conducting research in computer science, who realized they could form a Queer in AI group. (Other groups that took shape include Latinx in AI, {Dis}Ability in AI, and Muslim in ML.) For Agnew, in particular, having such a community felt like an urgent need. "It was hard to even imagine myself having a happy life," he says, reflecting on the lack of queer role models in the field. "There's Turing, but he committed suicide. So that's depressing. And the queer part of him is just ignored."

Not all affinity group members see a connection between their identity and their research. Still, each group has established particular expertise. Black in AI has become the intellectual center for exposing algorithmic discrimination, critiquing surveillance, and developing data-efficient AI techniques. Queer in AI has become a center for contesting the ways algorithms infringe on people's privacy and classify them into bounded categories by default.

Venkatasubramanian and Gebru also helped create the Fairness, Accountability, and Transparency (FAccT) conference to create a forum for research on the social and political implications of AI. Ideas and draft papers discussed at NeurIPS affinity group workshops often become the basis for papers published at FAccT, which then showcases that research to broader audiences.

It was after Buolamwini presented at the first Black in AI workshop, for example, that FAccT published Gender Shades. Along with Actionable Auditing, it then fueled several major education and advocacy campaigns to limit government use of facial recognition. When Amazon attempted to undermine the legitimacy of Buolamwini's and Raji's research, dozens of AI researchers and civil society organizations banded together to defend them, foreshadowing what they would later do for Gebru. Those efforts eventually contributed to Amazon's moratorium, which in May the company announced it would extend indefinitely.

The research also set off a cascade of regulation. More than a dozen cities have banned police use of facial recognition, and Massachusetts now requires police to get a judge's permission to use it. Both the US and the European Commission have proposed additional regulation.

"First we had to just be there," says Gebru. "And at some point, what Black in AI says starts to become important. And what all of these groups together say becomes important. You have to listen to us now."

FOLLOW THE MONEY

After Gebru and Mitchell's firing, the field is grappling anew with an age-old question: Is it possible to change the status quo while working from within? Gebru still believes working with tech giants is the best way to identify the problems. But she also believes that corporate researchers need stronger legal protections. If they see risky practices, they should be able to publicly share their observations without jeopardizing their careers.

Then there's the question of funding. Many researchers want more investment from the US government to support work that is critical of commercial AI development and advances the public welfare. Last year, it committed a measly $1 billion to non-defense-related AI research. The Biden administration is now asking Congress to invest an additional $180 billion in emerging technologies, with AI as a top priority.

Such funding could help people like Rediet Abebe, an assistant professor of computer science at the University of California, Berkeley. Abebe came into AI with ideas of using it to advance social equity. But when she started her PhD at Cornell, no one was focused on doing such research.

In the fall of 2016, as a PhD student, she began a small Cornell reading group with a fellow graduate student to study topics like housing instability, health-care access, and inequality. She then embarked on a new project to see whether her computational skills could support efforts to alleviate poverty.

Eventually, she found the Poverty Tracker study, a detailed data set on the financial shocks—unexpected expenses like medical bills or parking tickets—experienced by more than 2,000 New York families. Over many conversations with the study's authors, social workers, and nonprofits serving marginalized communities, she learned about their needs and told them how she could help. Abebe then developed a model that showed how the frequency and type of shocks affected a family's economic status.

Five years later, the project is still ongoing. She's now collaborating with nonprofits to improve her model and working with policymakers through the California Policy Lab to use it as a tool for preventing homelessness. Her reading group has also since grown into a 2,000-person community and is holding its inaugural conference later this year.

Abebe sees it as a way to incentivize more researchers to flip the norms of AI. While traditional computer science conferences emphasize advancing computational techniques for the sake of doing so, the new one will publish work that first seeks to deeply understand a social issue. The work is no less technical, but it builds the foundation for more socially meaningful AI to emerge.

"These changes that we're fighting for—it's not just for marginalized groups," she says. "It's actually for everyone." ■

Karen Hao is a senior editor for AI at MIT Technology Review.