Who Is Making Sure the A.I. Machines Aren't Racist?

When Google forced out two well-known artificial intelligence experts, a long-simmering research controversy burst into the open.

By Cade Metz

Published March 15, 2021 Updated June 23, 2023

Hundreds of people gathered for the first lecture at what had become the world's most important conference on artificial intelligence — row after row of faces. Some were East Asian, a few were Indian, and a few were women. But the vast majority were white men. More than 5,500 people attended the meeting, five years ago in Barcelona, Spain.

Timnit Gebru, then a graduate student at Stanford University, remembers counting only six Black people other than herself, all of whom she knew, all of whom were men.

The homogeneous crowd crystallized for her a glaring issue. The big thinkers of tech say A.I. is the future. It will underpin everything from search engines and email to the software that drives our cars, directs the policing of our streets and helps create our vaccines.

But it is being built in a way that replicates the biases of the almost entirely male, predominantly white work force making it.

In the nearly 10 years I've written about artificial intelligence, two things have remained a constant: The technology relentlessly improves in fits and sudden, great leaps forward. And bias is a thread that subtly weaves through that work in a way that tech companies are reluctant to acknowledge.

On her first night home in Menlo Park, Calif., after the Barcelona conference, sitting cross-legged on the couch with her laptop, Dr. Gebru described the A.I. work force conundrum in a Facebook post.

"I'm not worried about machines taking over the world. I'm worried about groupthink, insularity and arrogance in the A.I. community — especially with the current hype and demand for people in the field," she wrote. "The people creating the technology are a big part of the system. If many are actively excluded from its creation, this technology will benefit a few while harming a great many."

The A.I. community buzzed about the mini-manifesto. Soon after, Dr. Gebru helped create a new organization, Black in A.I. After finishing her Ph.D., she was hired by Google.

She teamed with Margaret Mitchell, who was building a group inside Google dedicated to "ethical A.I." Dr. Mitchell had previously worked in the research lab at Microsoft. She had grabbed attention when she told Bloomberg News in 2016 that A.I. suffered from a "sea of dudes" problem. She estimated that she had worked with hundreds of men over the previous five years and about 10 women.

Their work was hailed as groundbreaking. The nascent A.I. industry, it had become clear, needed minders and people with different perspectives.

About six years ago, A.I. in a Google online photo service organized photos of Black people into a folder called "gorillas." Four years ago, a researcher at a New York start-up noticed that the A.I. system she was working on was egregiously biased against Black people. Not long after, a Black researcher in Boston discovered that an A.I. system couldn't identify her face — until she put on a white mask.

In 2018, when I told Google's public relations staff that I was working on a book about artificial intelligence, it arranged a long talk with Dr. Mitchell to discuss her work. As she described how she built the company's Ethical A.I. team — and brought Dr. Gebru into the fold — it was refreshing to hear from someone so closely focused on the bias problem.

But nearly three years later, Dr. Gebru was pushed out of the company without a clear explanation. She said she had been fired after criticizing Google's approach to minority hiring and, with a research paper, highlighting the harmful biases in the A.I. systems that underpin Google's search engine and other services.

"Your life starts getting worse when you start advocating for underrepresented people," Dr. Gebru said in an email before her firing. "You start making the other leaders upset."

As Dr. Mitchell defended Dr. Gebru, the company removed her, too. She had searched through her own Google email account for material that would support their position and forwarded emails to another account, which somehow got her into trouble. Google declined to comment for this article.

Their departure became a point of contention for A.I. researchers and other tech workers. Some saw a giant company no longer willing to listen, too eager to get technology out the door without considering its implications. I saw an old problem — part technological and part sociological — finally breaking into the open.

80 Mistagged Photos

Artificial intelligence technology will eventually find its way into almost everything Google does. Cody O'Loughlin for The New York Times

It should have been a wake-up call.

In June 2015, a friend sent Jacky Alciné, a 22-year-old software engineer living in Brooklyn, an internet link for snapshots the friend had posted to the new Google Photos service. Google Photos could analyze snapshots and automatically sort them into digital folders based on what was pictured. One folder might be "dogs," another "birthday party."

When Mr. Alciné clicked on the link, he noticed one of the folders was labeled "gorillas." That made no sense to him, so he opened the folder. He found more than 80 photos he had taken nearly a year earlier of a friend during a concert in nearby Prospect Park. That friend was Black.

He might have let it go if Google had mistakenly tagged just one photo. But 80? He posted a screenshot on Twitter. "Google Photos, y'all," messed up, he wrote, using much saltier language. "My friend is not a gorilla."

Like facial recognition services, talking digital assistants and conversational "chatbots," Google Photos relied on an A.I. system that learned its skills by analyzing enormous amounts of digital data.

Called a "neural network," this mathematical system could learn tasks that engineers could never code into a machine on their own. By analyzing thousands of photos of gorillas, it could learn to recognize a gorilla. It was also capable of egregious mistakes. The onus was on engineers to choose the right data when training these mathematical systems. (In this case, the easiest fix was to eliminate "gorilla" as a photo category.)

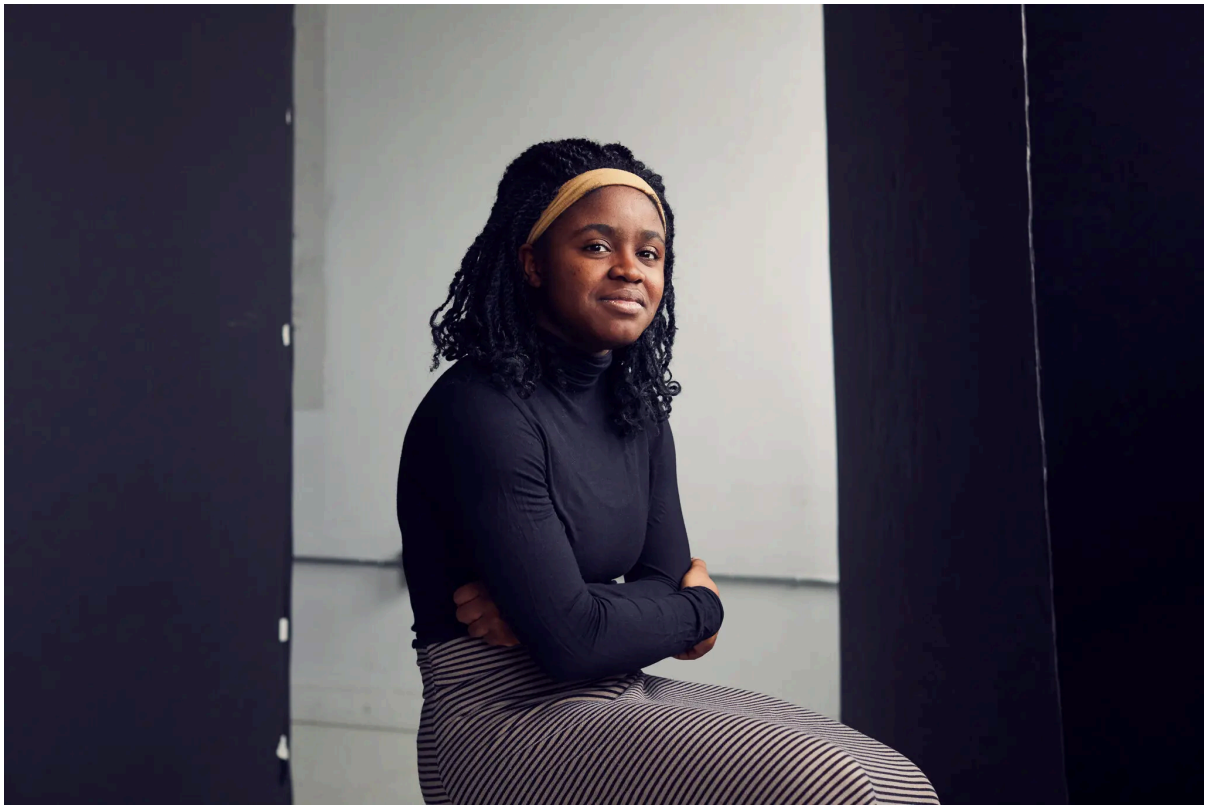
As a software engineer, Mr. Alciné understood the problem. He compared it to making lasagna. "If you mess up the lasagna ingredients early, the whole thing is ruined," he said. "It is the same thing with A.I. You have to be very intentional about what you put into it. Otherwise, it is very difficult to undo."

The Porn Problem

In 2017, Deborah Raji, a 21-year-old Black woman from Ottawa, sat at a desk inside the New York offices of Clarifai, the start-up where she was working. The company built technology that could automatically recognize objects in digital images and planned to sell it to businesses, police departments and government agencies.

She stared at a screen filled with faces — images the company used to train its facial recognition software.

As she scrolled through page after page of these faces, she realized that most — more than 80 percent — were of white people. More than 70 percent of those white people were male. When Clarifai trained its system on this data, it might do a decent job of recognizing white people, Ms. Raji thought, but it would fail miserably with people of color, and probably women, too.



Deborah Raji realized that a company's technology wasn't getting the input it needed to properly recognize people of color. Jaime Hogge for The New York Times

Clarifai was also building a "content moderation system," a tool that could automatically identify and remove pornography from images people posted to social networks. The company trained this system on two sets of data: thousands of photos pulled from online pornography sites, and thousands of G-rated images bought from stock photo services.

The system was supposed to learn the difference between the pornographic and the anodyne. The problem was that the G-rated images were dominated by white people, and the pornography was not. The system was learning to identify Black people as pornographic.

"The data we use to train these systems matters," Ms. Raji said. "We can't just blindly pick our sources."

This was obvious to her, but to the rest of the company it was not. Because the people choosing the training data were mostly white men, they didn't realize their data was biased.

"The issue of bias in facial recognition technologies is an evolving and important topic," Clarifai's chief executive, Matt Zeiler, said in a statement. Measuring bias, he said, "is an important step."

'Black Skin, White Masks'

Before joining Google, Dr. Gebru collaborated on a study with a young computer scientist, Joy Buolamwini. A graduate student at the Massachusetts Institute of Technology, Ms. Buolamwini, who is Black, came from a family of academics. Her grandfather specialized in medicinal chemistry, and so did her father.

She gravitated toward facial recognition technology. Other researchers believed it was reaching maturity, but when she used it, she knew it wasn't.

In October 2016, a friend invited her for a night out in Boston with several other women. "We'll do masks," the friend said. Her friend meant skin care masks at a spa, but Ms. Buolamwini assumed Halloween masks. So she carried a white plastic Halloween mask to her office that morning.

It was still sitting on her desk a few days later as she struggled to finish a project for one of her classes. She was trying to get a detection system to track her face. No matter what she did, she couldn't quite get it to work.

In her frustration, she picked up the white mask from her desk and pulled it over her head. Before it was all the way on, the system recognized her face — or, at least, it recognized the mask.

"Black Skin, White Masks," she said in an interview, nodding to the 1952 critique of historical racism from the psychiatrist Frantz Fanon. "The metaphor becomes the truth. You have to fit a norm, and that norm is not you."

Ms. Buolamwini started exploring commercial services designed to analyze faces and identify characteristics like age and sex, including tools from Microsoft and IBM.

She found that when the services read photos of lighter-skinned men, they misidentified sex about 1 percent of the time. But the darker the skin in the photo, the larger the error rate. It rose particularly high with images of women with dark skin. Microsoft's error rate was about 21 percent. IBM's was 35.

Published in the winter of 2018, the study drove a backlash against facial recognition technology and, particularly, its use in law enforcement. Microsoft's chief legal officer said the company had turned down sales to law enforcement when there was concern the technology could unreasonably infringe on people's rights, and he made a public call for government regulation.

Twelve months later, Microsoft backed a bill in Washington State that would require notices to be posted in public places using facial recognition and ensure that government agencies obtained a court order when looking for specific people. The bill passed, and it takes effect later this year. The company, which did not respond to a request for comment for this article, did not back other legislation that would have provided stronger protections.

Ms. Buolamwini began to collaborate with Ms. Raji, who moved to M.I.T. They started testing facial recognition technology from a third American tech giant: Amazon. The company had started to market its technology to police departments and government agencies under the name Amazon Rekognition.

Ms. Buolamwini and Ms. Raji published a study showing that an Amazon face service also had trouble identifying the sex of female and darker-skinned faces. According to the study, the service mistook women for men 19 percent of the time and misidentified darker-skinned women for men 31 percent of the time. For lighter-skinned males, the error rate was zero.

Amazon called for government regulation of facial recognition. It also attacked the researchers in private emails and public blog posts.

"The answer to anxieties over new technology is not to run 'tests' inconsistent with how the service is designed to be used, and to amplify the test's false and misleading conclusions through the news media," an Amazon executive, Matt

Wood, wrote in a blog post that disputed the study and a New York Times article that described it.

In an open letter, Dr. Mitchell and Dr. Gebru rejected Amazon's argument and called on it to stop selling to law enforcement. The letter was signed by 25 artificial intelligence researchers from Google, Microsoft and academia.

Last June, Amazon backed down. It announced that it would not let the police use its technology for at least a year, saying it wanted to give Congress time to create rules for the ethical use of the technology. Congress has yet to take up the issue. Amazon declined to comment for this article.

The End at Google

Dr. Gebru and Dr. Mitchell had less success fighting for change inside their own company. Corporate gatekeepers at Google were heading them off with a new review system that had lawyers and even communications staff vetting research papers.

Dr. Gebru's dismissal in December stemmed, she said, from the company's treatment of a research paper she wrote alongside six other researchers, including Dr. Mitchell and three others at Google. The paper discussed ways that a new type of language technology, including a system built by Google that underpins its search engine, can show bias against women and people of color.

After she submitted the paper to an academic conference, Dr. Gebru said, a Google manager demanded that she either retract the paper or remove the names of Google employees. She said she would resign if the company could not tell her why it wanted her to retract the paper and answer other concerns.

The response: Her resignation was accepted immediately, and Google revoked her access to company email and other services. A month later, it removed Dr. Mitchell's access after she searched through her own email in an effort to defend Dr. Gebru.

In a Google staff meeting last month, just after the company fired Dr. Mitchell, the head of the Google A.I. lab, Jeff Dean, said the company would create strict rules meant to limit its review of sensitive research papers. He also defended the reviews. He declined to discuss the details of Dr. Mitchell's dismissal but said she had violated the company's code of conduct and security policies.

One of Mr. Dean's new lieutenants, Zoubin Ghahramani, said the company must be willing to tackle hard issues. There are "uncomfortable things that responsible A.I. will inevitably bring up," he said. "We need to be comfortable with that discomfort."

But it will be difficult for Google to regain trust — both inside the company and out.

"They think they can get away with firing these people and it will not hurt them in the end, but they are absolutely shooting themselves in the foot," said Alex Hanna, a longtime part of Google's 10-member Ethical A.I. team. "What they have done is incredibly myopic."

Cade Metz is a technology correspondent at The Times and the author of "Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and the World," from which this article is adapted.

Cade Metz is a technology reporter and the author of "Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and The World." He covers artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas. More about Cade Metz

A version of this article appears in print on , Section BU, Page 5 of the New York edition with the headline: Can Artificial Intelligence Be Bias-Free?