Publication: The New York Times, March 15, 2021 Author: Cade Metz Title: Who Is Making Sure the A.I. Machines Aren’t Racist? URL: https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html

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.

PHOTO CAPTION: Artificial intelligence technology will eventually find its way into almost everything Google does. Credit...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.

PHOTO CAPTION: Deborah Raji realized that a company’s technology wasn’t getting the input it needed to properly recognize people of color. Credit...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.

Publication: MIT Technology Review, December 4, 2020 Author: Karen Hao Title: We read the paper that forced Timnit Gebru out of Google. Here’s what it says. Subheader: The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business. URL: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

On the evening of Wednesday, December 2, Timnit Gebru, the co-lead of Google’s ethical AI team, announced via Twitter that the company had forced her out.

Gebru, a widely respected leader in AI ethics research, is known for coauthoring a groundbreaking paper that showed facial recognition to be less accurate at identifying women and people of color, which means its use can end up discriminating against them. She also cofounded the Black in AI affinity group, and champions diversity in the tech industry. The team she helped build at Google is one of the most diverse in AI and includes many leading experts in their own right. Peers in the field envied it for producing critical work that often challenged mainstream AI practices.

A series of tweets, leaked emails, and media articles showed that Gebru’s exit was the culmination of a conflict over another paper she coauthored. Jeff Dean, the head of Google AI, told colleagues in an internal email (which he has since put online) that the paper “didn’t meet our bar for publication” and that Gebru had said she would resign unless Google met a number of conditions, which it was unwilling to meet. Gebru tweeted that she had asked to negotiate “a last date” for her employment after she got back from vacation. She was cut off from her corporate email account before her return.

Online, many other leaders in the field of AI ethics are arguing that the company pushed her out because of the inconvenient truths that she was uncovering about a core line of its research—and perhaps its bottom line. More than 1,400 Google staff members and 1,900 other supporters have also signed a letter of protest.

Many details of the exact sequence of events that led up to Gebru’s departure are not yet clear; both she and Google have declined to comment beyond their posts on social media. But MIT Technology Review obtained a copy of the research paper from one of the coauthors, Emily M. Bender, a professor of computational linguistics at the University of Washington. Though Bender asked us not to publish the paper itself because the authors didn’t want such an early draft circulating online, it gives some insight into the questions Gebru and her colleagues were raising about AI that might be causing Google concern.

“On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” lays out the risks of large language models—AIs trained on staggering amounts of text data. These have grown increasingly popular—and increasingly large—in the last three years. They are now extraordinarily good, under the right conditions, at producing what looks like convincing, meaningful new text—and sometimes at estimating meaning from language. But, says the introduction to the paper, “we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.”

**The paper**

The paper, which builds on the work of other researchers, presents the history of natural-language processing, an overview of four main risks of large language models, and suggestions for further research. Since the conflict with Google seems to be over the risks, we’ve focused on summarizing those here.

**Environmental and financial costs**

Training large AI models consumes a lot of computer processing power, and hence a lot of electricity. Gebru and her coauthors refer to a 2019 paper from Emma Strubell and her collaborators on the carbon emissions and financial costs of large language models. It found that their energy consumption and carbon footprint have been exploding since 2017, as models have been fed more and more data.

Strubell’s study found that training one language model with a particular type of “neural architecture search” (NAS) method would have produced the equivalent of 626,155 pounds (284 metric tons) of carbon dioxide—about the lifetime output of five average American cars. Training a version of Google’s language model, BERT, which underpins the company’s search engine, produced 1,438 pounds of CO2 equivalent in Strubell’s estimate—nearly the same as a round-trip flight between New York City and San Francisco. These numbers should be viewed as minimums, the cost of training a model one time through. In practice, models are trained and retrained many times over during research and development.

Gebru’s draft paper points out that the sheer resources required to build and sustain such large AI models means they tend to benefit wealthy organizations, while climate change hits marginalized communities hardest. “It is past time for researchers to prioritize energy efficiency and cost to reduce negative environmental impact and inequitable access to resources,” they write.

**Massive data, inscrutable models**

Large language models are also trained on exponentially increasing amounts of text. This means researchers have sought to collect all the data they can from the internet, so there's a risk that racist, sexist, and otherwise abusive language ends up in the training data.

An AI model taught to view racist language as normal is obviously bad. The researchers, though, point out a couple of more subtle problems. One is that shifts in language play an important role in social change; the MeToo and Black Lives Matter movements, for example, have tried to establish a new anti-sexist and anti-racist vocabulary. An AI model trained on vast swaths of the internet won’t be attuned to the nuances of this vocabulary and won’t produce or interpret language in line with these new cultural norms.

It will also fail to capture the language and the norms of countries and peoples that have less access to the internet and thus a smaller linguistic footprint online. The result is that AI-generated language will be homogenized, reflecting the practices of the richest countries and communities.

Moreover, because the training data sets are so large, it’s hard to audit them to check for these embedded biases. “A methodology that relies on datasets too large to document is therefore inherently risky,” the researchers conclude. “While documentation allows for potential accountability, [...] undocumented training data perpetuates harm without recourse.”

**Research opportunity costs**

The researchers summarize the third challenge as the risk of “misdirected research effort.” Though most AI researchers acknowledge that large language models don’t actually understand language and are merely excellent at manipulating it, Big Tech can make money from models that manipulate language more accurately, so it keeps investing in them. “This research effort brings with it an opportunity cost,” Gebru and her colleagues write. Not as much effort goes into working on AI models that might achieve understanding, or that achieve good results with smaller, more carefully curated data sets (and thus also use less energy).

**Illusions of meaning**

The final problem with large language models, the researchers say, is that because they’re so good at mimicking real human language, it’s easy to use them to fool people. There have been a few high-profile cases, such as the college student who churned out AI-generated self-help and productivity advice on a blog, which went viral.

The dangers are obvious: AI models could be used to generate misinformation about an election or the covid-19 pandemic, for instance. They can also go wrong inadvertently when used for machine translation. The researchers bring up an example: In 2017, Facebook mistranslated a Palestinian man’s post, which said “good morning” in Arabic, as “attack them” in Hebrew, leading to his arrest.

**Why it matters**

Gebru and Bender’s paper has six coauthors, four of whom are Google researchers. Bender asked to avoid disclosing their names for fear of repercussions. (Bender, by contrast, is a tenured professor: “I think this is underscoring the value of academic freedom,” she says.)

The paper’s goal, Bender says, was to take stock of the landscape of current research in natural-language processing. “We are working at a scale where the people building the things can’t actually get their arms around the data,” she said. “And because the upsides are so obvious, it’s particularly important to step back and ask ourselves, what are the possible downsides? … How do we get the benefits of this while mitigating the risk?”

In his internal email, Dean, the Google AI head, said one reason the paper “didn’t meet our bar” was that it “ignored too much relevant research.” Specifically, he said it didn’t mention more recent work on how to make large language models more energy efficient and mitigate problems of bias.

However, the six collaborators drew on a wide breadth of scholarship. The paper’s citation list, with 128 references, is notably long. “It’s the sort of work that no individual or even pair of authors can pull off,” Bender said. “It really required this collaboration.”

The version of the paper we saw does also nod to several research efforts on reducing the size and computational costs of large language models, and on measuring the embedded bias of models. It argues, however, that these efforts have not been enough. “I’m very open to seeing what other references we ought to be including,” Bender said.

Nicolas Le Roux, a Google AI researcher in the Montreal office, later noted on Twitter that the reasoning in Dean’s email was unusual. “My submissions were always checked for disclosure of sensitive material, never for the quality of the literature review,” he said.

Dean’s email also says that Gebru and her colleagues gave Google AI only a day for an internal review of the paper before they submitted it to a conference for publication. He wrote that “our aim is to rival peer-reviewed journals in terms of the rigor and thoughtfulness in how we review research before publication.”

Bender noted that even so, the conference would still put the paper through a substantial review process: “Scholarship is always a conversation and always a work in progress,” she said.

Others, including William Fitzgerald, a former Google PR manager, have further cast doubt on Dean’s claim.

Google pioneered much of the foundational research that has since led to the recent explosion in large language models. Google AI was the first to invent the Transformer language model in 2017 that serves as the basis for the company’s later model BERT, and OpenAI’s GPT-2 and GPT-3. BERT, as noted above, now also powers Google search, the company’s cash cow.

Bender worries that Google’s actions could create “a chilling effect” on future AI ethics research. Many of the top experts in AI ethics work at large tech companies because that is where the money is. “That has been beneficial in many ways,” she says. “But we end up with an ecosystem that maybe has incentives that are not the very best ones for the progress of science for the world.”

Publication: The Washington Post, December 23, 2020 Author: Nitasha Tiku Title: Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it. Subheader: Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility. URL: https://www.washingtonpost.com/technology/2020/12/23/google-timnit-gebru-ai-ethics/

PHOTO CAPTION: Google AI research scientist Timnit Gebru speaks on Sept. 7, 2018, at TechCrunch Disrupt SF 2018 at the Moscone Center in San Francisco. (Kimberly White/Getty Images/TechCrunch)

Two months ago, Google promoted Timnit Gebru, co-lead of a group focused on ethical artificial intelligence, after she earned a high score on her annual employee appraisal. Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility.

In his peer review of Gebru, Jeff Dean, the head of Google Artificial Intelligence, left only one comment when asked what she could do to have a greater impact, according to documents viewed by The Washington Post: Ensure that her team helps make a promising new software tool for processing human language “consistent with our AI Principles.”

In an email thanking Dean for his review, Gebru let him know that her team was already working on a paper about the ethical risks around the same language models, which are essential to understanding the complexity of language in search queries. On Oct. 20, Dean wrote that he wanted to see a draft, adding, “definitely not my area of expertise, but would definitely learn from reading it.”

Six weeks later, Google fired Gebru while she was on vacation.

“I can’t imagine anybody else who would be safer than me,” Gebru, 37, said. “I was super visible. I’m well known in the research community, but also the regulatory space. I have a lot of grass-roots support — and this is what happened.”

In an internal memo that he later posted online explaining Gebru’s departure, Dean told employees that the paper “didn’t meet our bar for publication” and “ignored too much relevant research” on recent positive improvements to the technology. Gebru’s superiors had insisted that she and the other Google co-authors either retract the paper or remove their names. Employees in Google Research, the department that houses the ethical AI team, say authors who make claims about the benefits of large language models have not received the same scrutiny during the approval process as those who highlight the shortcomings.

Her abrupt firing shows that Google is pushing back on the kind of scrutiny that it claims to welcome, according to interviews with Gebru, current Google employees, and emails and documents viewed by The Post.

It raises doubts about Silicon Valley’s ability to self-police, especially when it comes to advanced technology that is largely unregulated and being deployed in the real world despite demonstrable bias toward marginalized groups. Already, AI systems shape decision-making in law enforcement, employment opportunity and access to health care worldwide.

That made Gebru’s perspective essential in a field that is predominantly White, Asian and male. Women made up only 15 percent of the AI research staff at Facebook and 10 percent at Google, according to a 2018 report in Wired magazine. At Google, Black women make up 1.6 percent of the workforce.

Although Google publicly celebrated Gebru’s work identifying problems with AI, it disenfranchised the work internally by keeping it hierarchically distinct from other AI initiatives, not heeding the group’s advice, and not creating an incentive structure to put in practice the ethical findings, Gebru and other employees said.

Google declined to comment, but noted that in addition to the dozen or so staff members on Gebru’s team, 200 employees are focused on responsible AI.

Google has said that it did not fire Gebru, but accepted her “resignation,” citing her request to explain who at Google demanded that the paper be retracted, according to Dean’s memo. The company also blamed an email Gebru wrote to an employee resource group for women and allies at Google working in AI as inappropriate for a manager. The message warned the group that pushing for diversity was no use until Google leadership took accountability.

Rumman Chowdhury, a former global lead for responsible AI at Accenture and chief executive of Parity, a start-up that helps companies figure out how to audit algorithms, said there is a fundamental lack of respect within the industry for work on AI ethics compared with equivalent roles in other industries, such as model risk managers in quantitative hedge funds or threat analysts in cybersecurity.

“It’s being framed as the AI optimists and the people really building the stuff [versus] the rest of us negative Nellies, raining on their parade,” Chowdhury said. “You can’t help but notice, it’s like the boys will make the toys and then the girls will have to clean up.”

Google, which for decades evangelized an office culture that embraced employee dissent, has fired outspoken workers in recent years and shut down forums for exchange and questioning.

Nearly 3,000 Google employees and more than 4,000 academics, engineers and industry colleagues have signed a petition calling Gebru’s termination an act of retaliation by Google. Last week, nine Democratic lawmakers, including Sens. Elizabeth Warren (Mass.) and Cory Booker (N.J.) and Rep. Yvette D. Clarke (N.Y.), sponsor of the Algorithmic Accountability Act, a bill that would require companies to audit and correct race and gender bias in its algorithms, sent a letter to Google chief executive Sundar Pichai asking the company to affirm its commitment to research freedom and diversity.

Like any good researcher, Gebru is comfortable in the gray areas. And she has been using her ouster as an opportunity to shed light on the black box of algorithmic accountability inside Google — annotating the company’s claims with contradictory data, drawing connections to larger systemic issues, and illuminating the way internal AI ethics efforts can break down without oversight or a change in incentives to corporate practices and power structures.

Big Tech dominates AI research around advancements in machine learning, image recognition, language translation — poaching talent from top universities, sponsoring conferences and publishing influential papers. In response to concerns about the way those technologies could be abused or compound bias, the industry ramped up funding and promotion of AI ethics initiatives, beginning around 2016.

Tech giants have made similar investments in shaping policy debate around antitrust and online privacy, as a way to ward off lawmakers. Pichai invoked Google’s AI principles in an interview in 2018 with The Post, arguing for self-regulation around AI.

Google created its Ethical AI group in 2018 as an outgrowth of an employee-led push to prioritize fairness in the company’s machine learning applications. Margaret Mitchell, Gebru’s co-lead, pitched the idea of a team of researchers investigating the long-term effects of AI and translating those findings into action to mitigate harm and risk.

The same year, Pichai released a broadly worded set of principles governing Google’s AI work after thousands of employees protested the company’s contract with the Pentagon to analyze surveillance imagery from drones. But Google, which requires privacy and security tests before any product launch, has not mandated an equivalent process for vetting AI ethics, employees say.

Gebru, whose family’s ethnic origins are in Eritrea, was born and raised in Ethiopia and came to Massachusetts as 16-year-old after receiving political asylum from the war between the two African countries. She began her career as an electrical engineer at Apple and received her PhD from the Stanford Artificial Intelligence Lab, studying computer vision under renowned computer scientist Fei-Fei Li, a former Google executive and now co-director of Stanford’s Human-Centered AI Institute, which receives funding from Google.

Gebru did her postdoctoral research at Microsoft Research as part of a group focused on accountability and ethics in AI. There, she and Joy Buolamwini, then a masters student at MIT Media Lab, co-wrote a groundbreaking 2018 study that found that commercial facial recognition tools sold by companies such as IBM and Microsoft were 99 percent accurate at identifying White males, but only 35 percent effective with Black women.

In June, IBM, Microsoft and Amazon announced that they would stop selling the software to law enforcement, which Dean credited to Gebru’s work. She also co-founded Black in AI, a nonprofit organization that increased the number Black attendees at the largest annual AI conference.

VIDEO https://www.washingtonpost.com/video/technology/googles-search-evolution-from-oracle-to-advertiser/2020/10/19/236f4f59-6f35-4f04-a2fd-a3065c96929b\_video.html CAPTION: Compare Google search engine results over nearly two decades and a trend emerges: Results are filled with advertising and non-Google results are lower down. (The Washington Post)

Gebru said that in 2018, Google recruited her with the promise of total academic freedom. She was unconvinced, but the company was opening its first artificial intelligence lab on the African continent in Accra, the capital of Ghana, and she wanted to be involved. When she joined Google, Gebru said, she was the first Black female researcher in the company. (When she left, there were still only a handful of Black women working in research, out of hundreds.)

Gebru said she was also drawn to working with Mitchell. Both women prioritized foresight and building practical solutions to prevent AI risk, whereas the operating mind-set in tech is biased toward benefits and “rapid hindsight,” in response to harm, Mitchell said.

Gebru’s approach to ethical AI was shaped by her experiences. Hardware, for instance, came with datasheets that documented whether components were safe to use in certain situations. “When you look at this field as a whole, that doesn’t exist,” said Gebru, an electrical engineer. “It’s just super behind in terms of documentation and standards of safety.”

She also leaned on her industry experience when collaborating with other teams. Engineers live on a product cycle, consumed with putting out fires and fixing bugs. A vague requirement to “make things fair” would only cause more work and frustration, she thought. So she tried to build institutional structures and documentation tools for “when people want to do the right thing.”

Despite their expertise, the Ethical AI group fought to be taken seriously and included in Google’s other AI efforts, employees said.

Within the company, Gebru and her former colleagues said, there is little transparency or accountability regarding how decisions around AI ethics or diversity initiatives get made. Work on AI principles, for instance, falls under Kent Walker, the senior vice president of global affairs, whose vast purview includes lobbying, public policy and legal work. Walker also runs an internal ethics board of top executives, including Dean, called the Advanced Technology Review Council, which is responsible for yes or no decisions when issues escalate, Gebru said. The Ethical AI team had to fight to be consulted on Walker’s initiatives, she said.

“Here’s the guy tasked with covering Google’s a--, lobbying and also … working on AI principles,” Gebru said. “Shouldn’t you have a different entity that pushes back a little bit internally — some sort of push and pull?” What’s more, members of Walker’s council are predominantly vice presidents or higher, constricting diversity, Gebru said.

In her conversations with product teams, such as a group working on fairness in Machine Learning Infrastructure, Gebru said she kept getting questions about what tools and features they could build to protect against the ethical risks involved with large language models. Google had credited it with the biggest breakthrough in improving search results in the past five years. The models can process words in relation to the other words that come before and after them, which is useful for understanding the intent behind conversational search queries.

But despite the increasing use of these models, there was limited research investigating groups that might be negatively impacted. Gebru says she wanted to help develop those safeguards, one of the reasons she agreed to collaborate with the research paper proposed by Emily M. Bender, a linguist at the University of Washington.

Mitchell, who developed the idea of model cards, like nutrition labels for machine learning models, described the paper as “due diligence.” Her model card idea is being adopted more widely across the industry, and engineers needed to know how to fill out the section on harm.

Gebru said her biggest contribution to both her team and the paper has been identifying researchers who study directly-affected communities.

That diversity was reflected in the authors of the paper, including Mark Diaz, a Black and Latino Google researcher whose previous work looked at how platforms leave out the elderly, who talk about ageism in blog posts, but don’t share as much on sites such as Twitter. For the paper, he identified the possibility that large data sets from the Internet, particularly if they are from a single moment in time, will not reflect cultural shifts from social movements, such as the #MeToo movement or Black Lives Matter, which seek to shift power through changes in language.

The paper identified four overarching categories of harm, according to a recent draft viewed by The Post. It delved into the environmental effect of the computing power, the inscrutability of massive data sets used to train these models, the opportunity costs of the “hype” around claims that these models can understand language, as opposed to identifying patterns, and the danger that the real-sounding text generated by such models could be used to spread misinformation.

Because Google depends on large language models, Gebru and Mitchell expected that the company might push back against certain sections or attempt to water down their findings. So they looped in PR & Policy representatives in mid-September, with plenty of time before the deadline for changes at the end of January 2021.

Before making a pre-publication draft available online, Gebru first wanted to vet the paper with a variety of experts, including those who have built large language models. She asked for feedback from two top people at OpenAI, an AI research lab co-founded by Elon Musk, in addition to her manager at Google, and about 30 others. They suggested additions or revisions, Gebru said. “I really wanted to send it to people who would disagree with our view and be defensive,” Gebru said.

Given all their upfront effort, Gebru was baffled when she received a notification for a meeting with Google Research Vice President Megan Kacholia at 4.30 p.m. on the Thursday before Thanksgiving.

At the meeting, Kacholia informed Gebru and her co-authors that Google wanted the paper retracted.

On Thanksgiving, a week after the meeting, Gebru composed a six-page email to Kacholia and Dean outlining how disrespectful and oddly secretive the process had been.

“Specific individuals should not be empowered to unilaterally shut down work in such a disrespectful manner,” she wrote, adding that researchers from underrepresented groups were mistreated.

Mitchell, who is White, said she shared Gebru’s concerns but did not receive the same treatment from the company. “Google is very hierarchical, and it’s been a battle to have any sort of recognition,” she said. “We tried to explain that Timnit, and to a lesser extent me, are respected voices publicly, but we could not communicate upwards.”

“How can you still ask why there aren’t Black women in this industry?” Gebru said.

Gebru said she found out this week that the paper was accepted to the Conference on Fairness, Accountability and Transparency, as part of its anonymous review process. “It’s sad, the scientific community respects us a lot more than anybody inside Google,” she said.