One

AFTERNOON IN LATE NOVEMBER of last year, Timnit Gebru was sitting on the couch in her San Francisco Bay Area home, crying.

Gebru, a researcher at Google, had just clicked out of a last-minute video meeting with an executive named Megan Kacholia, who had issued a jarring command. Gebru was the coleader of a group at the company that studies the social and ethical ramifications of artificial intelligence, and Kacholia had ordered Gebru to retract her latest research paper—or else remove her name from its list of authors, along with those of several other members of her team.

The paper in question was, in Gebru's mind, pretty unobjectionable. It surveyed the known pitfalls of so-called large language models, a type of AI software—most famously exemplified by a system called GPT-3—that was stoking excitement in the tech industry. Google's own version of the technology was now helping to power the company's search engine. Jeff Dean, Google's revered head of research, had encouraged Gebru to think about the approach's possible downsides. The paper had sailed through the company's internal review process and had been submitted to a prominent conference. But Kacholia now said that a group of product leaders and others inside the company had deemed the work unacceptable, Gebru recalls. Kacholia was vague about their objections but gave Gebru a week to act. Her firm deadline was the day after Thanksgiving.

Gebru's distress turned to anger as that date drew closer and the situation turned weirder. Kacholia gave Gebru's manager, Samy Bengio, a document listing the paper's supposed flaws, but told him not to send it to Gebru, only to read it to her. On Thanksgiving Day, Gebru skipped some festivities with her family to hear Bengio's recital. According to Gebru's recollection and contemporaneous notes, the document didn't offer specific edits but complained that the paper handled topics "casually" and painted too bleak a picture of the new technology. It also claimed that all of Google's uses of large language models were "engineered to avoid" the pitfalls that the paper described.

Gebru spent Thanksgiving writing a six-page response, explaining her perspective on the paper and asking for guidance on how it might be revised instead of quashed. She titled her reply "Addressing Feedback from the Ether at Google," because she still didn't know who had set her Kafkaesque ordeal in motion, and sent it to Kacholia the next day.

On Saturday, Gebru set out on a pre-planned cross-country road trip. She had reached New Mexico by Monday, when Kacholia emailed to ask for confirmation that the paper would either be withdrawn or cleansed of its Google affiliations. Gebru tweeted a cryptic reproach of "censorship and intimidation" against AI ethics researchers. Then, on Tuesday, she fired off two emails: one that sought to end the dispute, and another that escalated it beyond her wildest imaginings.

The first was addressed to Kacholia and offered her a deal: Gebru would remove herself from the paper if Google provided an account of who had reviewed the work and how, and established a more transparent review process for future research. If those conditions weren't met, Gebru wrote, she would leave Google once she'd had time to make sure her team wouldn't be too destabilized. The second email showed less corporate diplomacy. Addressed to a listserv for women who worked in Google Brain, the company's most prominent AI lab and home to Gebru's Ethical AI team, it accused the company of "silencing marginalized voices" and dismissed Google's internal diversity programs as a waste of time.

Relaxing in an Airbnb in Austin, Texas, the following night, Gebru received a message with a 💀 from one of her direct reports: "You resigned??" In her personal inbox she then found an email from Kacholia, rejecting Gebru's offer and casting her out of Google. "We cannot agree as you are requesting," Kacholia wrote. "The end of your employment should happen faster than your email reflects." Parts of Gebru's email to the listserv, she went on, had shown "behavior inconsistent with the expectations of a Google manager." Gebru tweeted that she had been fired. Google maintained—and still does—that she resigned.

WIRED, JUL/AUG, 2021

Gebru's tweet lit the fuse on a controversy that quickly inflamed Google. The company has been dogged in recent years by accusations from employees that it mistreats women and people of color, and from lawmakers that it wields unhealthy technological and economic power. Now Google had expelled a Black woman who was a prominent advocate for more diversity in tech, and who was seen as an important internal voice for greater restraint in the helter-skelter race to develop and deploy AI. One Google machine-learning researcher who had followed Gebru's writing and work on diversity felt the news of her departure like a punch to the gut. "It was like, oh, maybe things aren't going to change so easily," says the employee, who asked to remain anonymous because they were not authorized to speak by Google management.

Dean sent out a message urging Googlers to ignore Gebru's call to disengage from corporate diversity exercises; Gebru's paper had been subpar, he said, and she and her collaborators had not followed the proper approval process. In turn, Gebru claimed in tweets and interviews that she'd been felled by a toxic cocktail of racism, sexism, and censorship. Sympathy for Gebru's account grew as the disputed paper circulated like samizdat among AI researchers, many of whom found it neither controversial nor particularly remarkable. Thousands of Googlers and outside AI experts signed a public letter castigating the company.

But Google seemed to double down. Margaret Mitchell, the other coleader of the Ethical AI team and a prominent researcher in her own right, was among the hardest hit by Gebru's ouster. The two had been a professional and emotional tag team, building up their group—which was one of several that worked on what Google called "responsible AI"—while parrying the sexist and racist tendencies they saw at large in the company's culture. Confident that those same forces had played a role in Gebru's downfall, Mitchell wrote an automated script to retrieve notes she'd kept in her corporate Gmail account that documented allegedly discriminatory incidents, according to sources inside Google. On January 20, Google said Mitchell had triggered an internal security system and had been suspended. On February 19, she was fired, with Google stating that it had found "multiple violations of our code of conduct, as well as of our security policies, which included exfiltration of confidential, business-sensitive documents."

Google had now fully decapitated its own Ethical AI research group. The long, spectacular fallout from that Thanksgiving ultimatum to Gebru left countless bystanders wondering: Had one paper really precipitated all of these events?

The story of what actually happened in the lead-up to Gebru's exit from Google reveals a more tortured and complex backdrop. It's the tale of a gifted engineer who was swept up in the AI revolution before she became one of its biggest critics, a refugee who worked her way to the center of the tech industry and became determined to reform it. It's also about a company—the world's fifth largest—trying to regain its equilibrium after four years of scandals, controversies, and mutinies, but doing so in ways that unbalanced the ship even further.

Beyond Google, the fate of Timnit Gebru lays bare something even larger: the tensions inherent in an industry's efforts to research the downsides of its favorite technology. In traditional sectors such as chemicals or mining, researchers who study toxicity or pollution on the corporate dime are viewed skeptically by independent experts. But in the young realm of people studying the potential harms of AI, corporate researchers are central.

Gebru's career mirrored the rapid rise of AI fairness research, and also some of its paradoxes. Almost as soon as the field sprang up, it quickly attracted eager support from giants like Google, which sponsored conferences, handed out grants, and hired the domain's most prominent experts. Now Gebru's sudden ejection made her and others wonder if this research, in its domesticated form, had always been doomed to a short leash. To researchers, it sent a dangerous message: AI is largely unregulated and only getting more powerful and ubiquitous, and insiders who are forthright in studying its social harms do so at the risk of exile.

2

IN APRIL 1998, two Stanford grad students named Larry Page and Sergey Brin presented an algorithm called PageRank at a conference in Australia. A month later, war broke out between Ethiopia and Eritrea, setting off a two-year border conflict that left tens of thousands dead. The first event set up Google's dominance of the internet. The second set 15-year-old Timnit Gebru on a path toward working for the future megacorp.

At the time, Gebru lived with her mother, an economist, in the Ethiopian capital of Addis Ababa. Her father, an electrical engineer with a PhD, had died when she was small. Gebru enjoyed school and hanging out in cafés when she and her friends could scrape together enough pocket money. But the war changed all that. Gebru's family was Eritrean, and some of her relatives were being deported to Eritrea and conscripted to fight against the country they had made their home.

Gebru's mother had a visa for the United States, where Gebru's older sisters, engineers like their father, had lived for years. But when Gebru applied for a visa, she was denied. So she went to Ireland instead, joining one of her sisters, who was there temporarily for work, while her mother went to America alone.

Reaching Ireland may have saved Gebru's life, but it also shattered it. She called her mother and begged to be sent back to Ethiopia. "I don't care if it's safe or not. I can't live here," she said. Her new school, the culture, even the weather were alienating. Addis Ababa's rainy season is stac-

cato, with heavy downpours interspersed by sunshine. In Ireland, rain fell steadily for a week. As she took on the teenage challenges of new classes and bullying, larger concerns pressed down. "Am I going to be reunited with my family? What happens if the paperwork doesn't work out?" she recalls thinking. "I felt unwanted."

The next year, Gebru was approved to come to the US as a refugee. She reunited with her mother in Somerville, Massachusetts, a predominantly white suburb of Boston, where she enrolled in the local public high school—and a crash course in American racism.

Some of her teachers, Gebru found, seemed unable or unwilling to accept that an African refugee might be a top student in math and science. Other white Americans saw fit to confide in her their belief that African immigrants worked harder than African Americans, whom they saw as lazy. History class told an uplifting story about the Civil Rights Movement resolving America's racial divisions, but that tale rang hollow. "I thought that cannot be true, because I'm seeing it in the school," Gebru says.

Piano lessons helped provide a space where she could breathe. Gebru also coped by turning to math, physics, and her family. She enjoyed technical work, not just for its beauty but because it was a realm disconnected from personal politics or worries about the war back home. That compartmentalization became part of Gebru's way of navigating the world. "What I had under my control was that I could go to class and focus on the work," she says.

Gebru's focus paid off. In September 2001 she enrolled at Stanford. Naturally, she chose the family major, electrical engineering, and before long her trajectory began to embody the Silicon Valley archetype of the immigrant trailblazer. For a course during her junior year, Gebru built an experimental electronic piano key, helping her win an internship at Apple making audio circuitry for Mac computers and other products. The next year she went to work for the company full-time while continuing her studies at Stanford.

At Apple, Gebru thrived. When Niel Warren, her manager, needed someone to dig into delta-sigma modulators, a class of analog-to-digital converters, Gebru volunteered, investigating whether the technology would work in the iPhone. "As an electrical engineer she was fearless," Warren says. He found his new hardware hotshot to be well liked, always ready with a hug, and determined outside of work too. In 2008, Gebru withdrew from one of her classes because she was devoting so much time to canvassing for Barack Obama in Nevada and Colorado, where many doors were slammed in her face.

"I'm not worried about machines taking over the world," Gebru wrote. "I'm worried about groupthink, insularity, and arrogance in the AI community."

As Gebru learned more about the guts of gadgets like the iPhone, she became more interested in the fundamental physics of their components—and soon her interests wandered even further, beyond the confines of electrical engineering. By 2011, she was embarking on a PhD at Stanford, drifting among classes and searching for a new direction. She found it in computer vision, the art of making software that can interpret images.

Unbeknownst to her, Gebru now stood on the cusp of a revolution that would transform the tech industry in ways she would later criticize. One of Gebru's favorite classes involved creating code that could detect human figures in photos. "I wasn't thinking about surveillance," Gebru says. "I just found it technically interesting."

In 2013 she joined the lab of Fei-Fei Li, a computer vision specialist who had helped spur the tech industry's obsession with AI, and who would later work for a time at Google. Li had created a project called ImageNet that paid contractors small sums to tag a billion images scraped from the web with descriptions of their contents—cat, coffee cup, cello. The final database, some 15 million images, helped to reinvent machine learning, an AI technique that involves training software to get better at performing a task by feeding it examples of correct answers. Li's work demonstrated that an approach known as deep learning, fueled by a large collection of training data and powerful computer chips, could produce much more accurate machine-vision technology than prior methods had yielded.

Li wanted to use deep learning to give computers a more fine-grained understanding of the world. Two of her students had scraped 50 million images from Google Street View, planning to train a neural network to spot cars and identify their make and model. But they began wondering about other applications they might build on top of that capability. If you drew correlations between census data and the cars visible on a street, could that provide a way to estimate the demographic or economic characteristics of any neighborhood, just from pictures?

Gebru spent the next few years showing that, to a certain level of accuracy, the answer was yes. She and her collaborators used online contractors and car experts recruited on Craigslist to identify the make and model of 70,000 cars in a sample of Street View images. The annotated pictures provided the training data needed for deep-learning algorithms to figure out how to identify cars in new images. Then they processed the full Street View collection and identified 22 million cars in photos from 200 US cities. When Gebru correlated those observations with census and crime data, her results showed that more pickup trucks and VWs indicated more white residents, more Buicks and Oldsmobiles indicated more Black ones, and more vans corresponded to higher crime.

rithms could predict factors like household income and voting patterns just by identifying cars on the street.

Gebru was the only speaker who was not a professor, investment professional, or representative of a tech company, but, as one organizer recalls, her talk generated more interest than any of the others. Steve Jurvetson, a friend of Elon Musk and an early investor in Tesla, enthusiastically posted photos of her slides to Facebook. A longtime AI aficionado, he wasn't surprised that machine-learning algorithms could identify specific cars. But the way Gebru had extracted signals about society from photos illustrated how the technology could spin gold from unexpected sources—at least for those with plenty of data to mine. "It was, 'My God, think of all the data that Google has,'" Jurvetson says. "It made me realize the power of having the biggest data set."

For Gebru, the event could have been a waypoint between her grad school AI work and a job building moneymaking algorithms for tech giants. But she decided that she wanted to help contain the technology's power rather than expand it. In the summer of 2017, she took a job with a Microsoft research group that had been involved in the FATML movement from early on. Gebru wrote her pivot into the final chapter of her thesis: "One of the most important emergent issues plaguing our society today is that of algorithmic bias. Most works based on data mining, including my own works described in this thesis, suffer from this problem," she wrote. Her plan for a career, she went on, was "to make contributions towards identifying and mitigating these issues."

3

WHILE GEBRU WAS completing her thesis at Stanford, Margaret Mitchell was developing her own doubts about AI, 800 miles north at Microsoft's verdant campus outside Seattle.

In 2015, Mitchell, an expert in software that generates language from images, was working on an app for blind people that spoke visual descriptions of the world. She had christened it Seeing AI, and she loved the idea that the flourishing power of machine learning could lift up society's most vulnerable. But Microsoft didn't seem willing to seriously invest in such projects at the time.

Mitchell also noticed some troubling gaffes in the machine-learning systems she was training. One would describe someone with pale skin, like the red-haired Mitchell, as a "person," but a figure with dark skin as a "Black person." In another test, an image of an inferno at an oil storage depot was captioned "great view." She began to fear that AI was laced with land mines, and the industry was not paying enough attention to finding them. "Oh crap," she remembers thinking. "There are serious issues that we have to solve right now because no one else is working on them and this technology is evolving."

In 2016, Mitchell moved to Google to work full-time on those problems. The company appeared to be embracing this new, conscientious strand of AI research. A couple of weeks before she started, Google published its first research paper on machine-learning fairness. It considered how to ensure that a system that makes predictions about people—say, assessing their risk of defaulting on a loan—offered equal treatment to individuals regardless of their gender, race, religion, or other group identity. The company highlighted its research in a blog post for a general audience, and signed up, alongside Microsoft, as a corporate sponsor of the FATML workshop.

When Mitchell got to Google, she discovered a messier reality behind the company's entrée into fairness research. That first paper had been held up for months by internal deliberations over whether Google should publicly venture into a discourse on the discriminatory potential of computer code, which to managers seemed more complex and sensitive than its labs' usual output. Mitchell's own first publication at the company, on making smile-detection algorithms perform well for people of different races and genders, also met with a degree of corporate hesitancy that didn't seem to encumber more conventional AI projects. She chose to work on smiles in part because of their positive associations; still, she endured rounds of meetings with lawyers over how to handle discussions of gender and race.

At other times, Mitchell's work inside Google faced little resistance, but also little enthusiasm. "It was like people really appreciated what I was saying, and then nothing happened," she says. Still, Mitchell hadn't expected to change the company overnight, and gradually her efforts gained momentum. In late 2017 she formed a small team dedicated to "ethical AI research" and embarked on a campaign of meetings with teams across Google to spread the word and offer help. This time people seemed more receptive—perhaps in part because broader attitudes were shifting. Some of Google's rivals, like Microsoft, appeared to be taking AI fairness more seriously. Industry hype about AI was still intense, but the field's culture was becoming more reflective.

One person driving that change was Timnit Gebru, who was introduced to Mitchell by an acquaintance over email when Gebru was about to join Microsoft. The two had become friendly, bonding over a shared desire to call out injustices in society and the tech industry. "Timnit and I hit it off immediately," Mitchell says. "We got along on every dimension."

When Gebru presented her PhD thesis on computer vision to members of Silicon Valley's elite, her talk generated intense interest.

This demonstration of AI's power positioned Gebru for a lucrative career in Silicon Valley. Deep learning was all the rage, powering the industry's latest products (smart speakers) and its future aspirations (self-driving cars). Companies were spending millions to acquire deep-learning technology and talent, and Google was placing some of the biggest bets of all. Its subsidiary DeepMind had recently celebrated the victory of its machine-learning bot over a human world champion at Go, a moment that many took to symbolize the future relationship between humans and technology.

Gebru's project fit in with what was becoming the industry's new philosophy: Algorithms would soon automate away any problem, no matter how messy. But as Gebru got closer to graduation, the boundary she had established between her technical work and her personal values started to crumble in ways that complicated her feelings about the algorithmic future.

Gebru had maintained a fairly steady interest in social justice issues as a grad student. She wrote in *The Stanford Daily* about an incident in which an acquaintance wondered aloud whether Gebru was "actually smart" or had been admitted due to affirmative action. At Stanford's graduate school, Gebru encountered a significantly less diverse student population than she had during her undergraduate years, and she felt isolated. She bonded with people who, like her, had experienced global inequality firsthand. "Once you've seen the world in terms of its injustice and the ways in which the United States is not always the answer to everybody's problems, it's very difficult to unsee," says Jess Auerbach, a student from South Africa who became friends with Gebru at Stanford, and who is now an anthropologist at North West University in South Africa.

In 2016, Gebru volunteered to work on a coding program for bright young people in Ethiopia, which sent her on a trip back home, only her second since she had fled at the age of 15. Her coding students' struggles, she felt, exposed the limits of US meritocracy. One promising kid couldn't afford the roughly $100 required to take the SAT. After Gebru paid the fee for him, he won a scholarship to MIT. She also pitched in to help students who had been denied visas despite having been accepted to US schools. "She tried all she could to help these kids," says Jelani Nelson, the UC Berkeley professor who founded the program.

Li, Gebru's adviser at Stanford, encouraged her to find a way to connect social justice and tech, the two pillars of her worldview. "It was obvious to an outsider, but I don't think it was obvious to her, that actually there was a link between her true passion and her technical background," Li says. Gebru was reluctant to forge that link, fearing in part that it would typecast her as a Black woman first and a technologist second.

But she also became more aware that technology can sometimes reflect or magnify society's biases, rather than transcend them. In 2016, ProPublica reported that a recidivism-risk algorithm called COMPAS, used widely in courtrooms across the country, made more false predictions that Black people would reoffend than it did for white people (an analysis that was disputed by the company that made the algorithm). This made Gebru wonder whether the crime data she'd used in her own research reflected biased policing. Around the same time, she was introduced to Joy Buolamwini, a Ghanaian American MIT master's student who had noticed that some algorithms designed to detect faces worked less well on Black people than they did on white people. Gebru began advising her on publishing her results.

It wasn't just the algorithms or their training data that skewed white. In 2015, Gebru got her first glimpse of the worldwide community of AI researchers at the field's top conference, Neural Information Processing Systems (NIPS), in Montreal. She noticed immediately how male and how white it was. At a Google party, she was intercepted by a group of strangers in Google Research T-shirts who treated the presence of a Black woman as a titillating photo op. One man grabbed her for a hug; another kissed her cheek and took a photo. At the next year's conference, Gebru kept a tally of other Black people she met, counting just six among the 8,500 attendees—all people she already knew, and most of whom she'd already added to an email list she'd started for Black people in the field. After the event, Gebru posted a warning to AI researchers on Facebook about the dangers of their community's lack of diversity. "I'm not worried about machines taking over the world, I'm worried about groupthink, insularity, and arrogance in the AI community," she wrote. "If many are actively excluded from its creation, this technology will benefit a few while harming a great many."

Gebru's awakening roughly coincided with the emergence of a new research field dedicated to examining some of the social downsides of AI. It came to be centered on an annual academic workshop, first held in 2014, called Fairness, Accountability, and Transparency in Machine Learning (FATML) and motivated by concerns over institutional decisionmaking. If algorithms decided who received a loan or awaited trial in jail rather than at home, any errors they made could be life-changing.

The event's creators initially found it difficult to convince peers that there was much to talk about. "The more predominant idea was that humans were biased and algorithms weren't," says Moritz Hardt, now a UC Berkeley computer science professor who cofounded the workshop with a researcher from Princeton. "People thought it was silly to work on this."

By 2016 the event had grown into a meeting that sold out a hall at NYU School of Law. The audience included staffers from the Federal Trade Commission and the European Commission. Yet the presenters, by and large, applied a fairly detached and mathematical lens to the notion that technology could harm people. Researchers hashed out technical definitions of fairness that could be expressed in the form of code. There was less talk about how economic pressures or structural racism might shape AI systems, whom they work best for, and whom they harm.

Gebru didn't attend the FATML workshop that year or the next—she was still mainly focused on building AI, not examining its potential for harm. In January 2017, at a one-day event centered on how AI could shake up finance, Gebru stood in a gray turtleneck in a large octagonal room overlooking Stanford's terracotta-roofed campus and presented the findings of her PhD thesis to members of Silicon Valley's elite. She clicked through slides showing how algo-

---

TOM SIMONITE *(@tsimonite) is a senior writer at* WIRED *who covers artificial intelligence and its effects on the world.*

4

GEBRU ARRIVED AT the Googleplex in September 2018. She took a desk not far from Jeff Dean's in one of the buildings that housed Google Brain, across the main courtyard from the volleyball court and the replica of a *Tyrannosaurus rex* skeleton. She didn't keep a low profile for long. Two months into her new job, she walked out, joining thousands of employees worldwide to protest the company's treatment of women after *The New York Times* reported that Google had paid $90 million in severance to an executive accused of sexual harassment.

Gebru joined a discussion about the protest on an internal email list called Brain Women and Allies. She pointed out some problems she'd noticed at her new workplace, including "toxic men" and a lack of women in senior positions. She was summoned to a brief meeting with Dean—now her boss's boss—and a representative from human resources to discuss her observations.

Soon after, Gebru met with Dean again, this time with Mitchell at her side, for another discussion about the situation of women at Google. They planned a lunch meeting, but by the time the appointment rolled around, the two women were too anxious to eat. Mitchell alleged that she had been held back from promotions and raises by performance reviews that unfairly branded her as uncollaborative. Gebru asserted that a male researcher with less experience than her had recently joined Google Brain at a more senior level. Dean said he'd look into the pair's claims. Gebru was promoted; Dean told her that the hiring committee had not previously seen all parts of her résumé, an explanation she found dubious. After more back and forth over Mitchell's position, Dean let her switch supervisors.

Gebru and Mitchell's work didn't fit easily into Google's culture, either. The women and their team were a relatively new breed of tech worker: the in-house ethical quibbler. After the dustup at Google over Project Maven, and in the wake of research like Buolamwini and Gebru's, tech giants began trumpeting lofty corporate commitments to practice restraint in their AI projects. After Google said it would not renew its controversial Pentagon contract, it announced a set of seven principles that would guide its AI work. Among them: AI projects had to be "socially beneficial" and could not relate to weapons or surveillance (though other defense work was still permitted). Microsoft posted six AI principles that were less specific, including "inclusiveness" and "accountability." Both companies created internal review processes for cloud computing deals that they said would weed out unethical projects. In 2016, Microsoft and Google were the only corporate sponsors of the FATML workshop; in 2019, they were joined by Google's Alphabet sibling DeepMind, as well as Spotify and Twitter, as sponsors of an entire conference that had in part grown out of the FATML workshop. Gebru was one of its organizers.

Despite those changes, it remained unclear to some of the in-house quibblers how, exactly, they would or could change Google. The Ethical AI team's primary job was to conduct research, but Mitchell also wanted the group to shape the company's products, which touched billions of lives. Indifference and a lack of support, however, sometimes stood in their way. In some cases, Mitchell herself wrote code for product teams that wanted to implement AI safeguards, because engineering resources weren't regularly made available for their kind of work.

So the Ethical AI team hustled, figuring out ways to get traction for their ideas and sometimes staging interventions. In one case, they noticed problems in Gmail's Smart Reply feature, which suggests short responses to emails: It made gendered assumptions, such as defaulting to "he" if a message included mention of an engineer. A member of the Ethical AI team met with an engineer on the project for a quiet chat. That helped set off a series of conversations, and the feature was adjusted to no longer use gendered pronouns.

Mitchell also developed a playbook for turning ethical AI itself into a kind of product, making it more palatable to Google's engineering culture, which prized launches of new tools and features. In January 2019, Mitchell, Gebru, and seven collaborators introduced a system for cataloging the performance limits of different algorithms. The method, which built on Gebru's earlier work documenting the contents and blind spots of data sets, noted the conditions under which algorithms were most likely to return accurate results and where they were likely to falter. Mitchell's team named the concept Model Cards, to make it sound generic and neutral, and shopped it around to other teams inside the company. The cloud computing division adopted Model Cards, using them as a form of disclosure, like a nutrition label, to show the public how well, say, Google's facial detection algorithm performs on different kinds of images.

On at least one occasion, the Ethical AI team also helped convince Google to limit its AI in ways that ceded potential revenue

"What Google just said to anyone who wants to do this critical research is, 'We're not going to tolerate it.'"
— Meredith Whittaker

to competitors. Microsoft and Amazon had for years offered face-recognition services that could be used for more or less anything, including law enforcement. With the Ethical AI team's help, Google launched a limited service that just recognized public figures and was offered only to customers in the media after careful vetting.

Mitchell and Gebru believed their successes derived in part from the fact that their team provided refuge from Google's internal culture, which they and some other researchers found hostile, territorial, and intensely hierarchical. The dozen or so people on the Ethical AI team took pride in being more diverse in terms of gender, race, and academic background than the rest of the company. Gebru fondly thought of them as misfits and believed that diversity made the group more likely to spot problems or opportunities that Google's largely white male workers might overlook. Gebru and Mitchell also successfully lobbied executives to allow them to bring in sociologists and anthropologists—not just the usual computer science PhDs. "A lot of people in our team would either not be at Google or maybe even in the tech industry if they didn't join," Gebru says.

Over time, the team seemed to show how corporate quibblers could succeed. Google's Ethical AI group won respect from academics and helped persuade the company to limit its AI technology. Gebru and Mitchell both reported to Samy Bengio, the veteran Google Brain researcher, whom they came to consider an ally. The company had built up a handful of other teams working on AI guardrails, including in the research and global affairs divisions, but they were tied more closely to the company's business priorities. The Ethical AI team was more independent and wide-ranging. When Mitchell started at Google, the field mainly took a narrow, technical approach to fairness. Now it increasingly asked more encompassing questions about how AI replicated or worsened social inequalities, or whether some AI technology should be placed off-limits. In addition to creating handy tools for engineers, members of the team published papers urging AI researchers to draw on critical race theory and reconsider the tech industry's obsession with building systems to achieve mass scale.

At the same time, however, Mitchell and Gebru's frustrations with Google's broader culture mounted. The two women say they were worn down by the occasional flagrantly sexist or racist incident, but more so by a pervasive sense that they were being isolated. They noticed that they were left out of meetings and off email threads, or denied credit when their work made an impact. Mitchell developed an appropriately statistical way of understanding the phenomenon. "What is the likelihood that I will not be invited to a meeting that I should be at? What is the likelihood that my male colleague will be invited? You start to see the trends," she says.

Together, the two women joined and sometimes led attempts to change Google's culture. In 2019, with two others, they circulated a pointed internal document listing concerns about the treatment of women in Google's research teams. Women were treated as "punching bags," the document asserted, and senior managers dismissed observations about inequality as "temper tantrums." Mitchell disseminated a chart explaining how to support marginalized groups at work, including checklist items like "Did you listen to their answer and respond with empathy?"

Gebru was the more outspoken of the two—usually because she felt, as a Black woman, that she had to be. She admits that this won her enemies. She dismissed as backward diversity programs that placed an emphasis on mentoring for women: The company's problems, she would say, were rooted in its culture and leadership, not in the marginalized workers. Gebru's willingness to speak up sometimes led to blowups. In one incident, she and another woman warned Dean that a male researcher at Google had previously been accused of sexual harassment. Managers did not appear to act until the man was accused of harassing multiple people at Google, after which he was fired. The man's lawyers then sent Google a letter in which they accused Gebru and the other woman of defaming him. Google lawyers in turn advised the pair to hire their own counsel. Gebru and her coworker did so, and their own lawyers warned Google that it had a duty to represent its employees. After that expensive pushback, the two women didn't hear more about the issue. (Google did not respond to a request for comment on the incident, but told Bloomberg it began an investigation immediately after receiving reports about the man and that he departed before the investigation concluded.)

Some Googlers chafed at Gebru's willingness to confront colleagues. "Timnit's behavior was very far outside the norm," says one researcher at Google who was not authorized to speak to the press. The researcher recalls an incident in the summer of 2020, during the wave of Black Lives Matter protests, when Gebru got into a dispute on an internal mailing list dedicated to discussing new AI research papers. A male colleague posted a short, enthusiastic message about a new text-generation system that had just been opened up for commercial use. Gebru, acutely conscious of the demonstrations roaring across America, replied to highlight a warning from a prominent woman in the field that such systems were known to sometimes spew racist and sexist language. Other researchers then replied to the initial post without mentioning Gebru's comment. Gebru called them out for ignoring her, saying it was a common and toxic pattern, and she says one man privately messaged her to say he wasn't surprised she got harassed online. A hot-tempered debate ensued over racism and sexism in the workplace.

According to the Google employee, the incident—which is also described in anonymous posts on Reddit—showed how Gebru's demeanor could make some people shy away from her or avoid certain technical topics for fear of being pulled into arguments about race and gender politics. Gebru doesn't deny that the dispute became heated but says it ultimately proved productive, forcing attention to her negative experiences and those of other women at Google.

5

ABOUT A YEAR AFTER Gebru first arrived at Google, in October 2019, the company summoned journalists to its headquarters in Mountain View to raise the curtain on a new technology. After a sumptuous breakfast buffet, reporters were shepherded into a narrow meeting room to hear from Dean and two vice presidents in charge of Google's search engine. The trio touted a new kind of machine-learning system that they said made the company's signature product better able to understand long queries.

Dean raised a polite chuckle when he explained that the new system was called Bidirectional Encoder Representations from Transformers, but was generally known by a name borrowed from *Sesame Street*: BERT. It was an example of a new type of machine-learning system known as a large language model, enabled by advances that made it practical for algorithms to train themselves on larger volumes of text, generally scraped from the web. That broader sampling allowed models like BERT to better internalize statistical patterns of language use, making them better than previous technology at tasks like answering questions or detecting whether a movie review was positive or negative.

When a reporter asked whether BERT would also learn, say, sexist language patterns, Dean responded, "This is something that we definitely look at for all the machine-learning-related product launches and also in our own research," citing the work of people like Mitchell and Gebru. "We want to make sure that our use of machine learning is free of unfair forms of bias." The Q&A also revealed that Google had other reasons to value BERT. When another journalist asked if it was being used by Google's ads team, one of the search executives replied, "I'm sure they must be applying it."

In the months that followed, excitement grew around large language models. In June 2020, OpenAI, an independent AI institute cofounded by Elon Musk but now bankrolled by a billion dollars from Microsoft, won a splurge of media coverage with a system called GPT-3. It had ingested more training data than BERT and could generate impressively fluid text in genres spanning sonnets, jokes, and computer code. Some investors and entrepreneurs predicted that automated writing would reinvent marketing, journalism, and art.

These new systems could also become fluent in unsavory language patterns, coursing with sexism, racism, or the tropes of ISIS propaganda. Training them required huge collections of text—BERT used 3.3 billion words and GPT-3 almost half a trillion—which engineers slurped from the web, the most readily available source with the necessary scale. But the data sets were so large that sanitizing them, or even knowing what they contained, was too daunting a task. It was an extreme example of the problem Gebru had warned against with her Datasheets for Datasets project.

Inside Google, researchers worked to build more powerful successors to BERT and GPT-3. Separately, the Ethical AI team began researching the technology's possible downsides. Then, in September 2020, Gebru and Mitchell learned that 40 Googlers had met to discuss the technology's future. No one from Gebru's team had been invited, though two other "responsible AI" teams did attend. There was a discussion of ethics, but it was led by a product manager, not a researcher.

That same month, Gebru sent a message to Emily M. Bender, a professor of linguistics at the University of Washington, to ask if she had written anything about the ethical questions raised by these new language models. Bender had not, and the pair decided to collaborate. Bender brought in a grad student, and Gebru looped in Mitchell and three other members of her Google team.

The resulting paper was titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" The whimsical title styled the software as a statistical mimic that, like a real parrot, doesn't know the implications of the bad language it repeats.

The paper was not intended to be a bombshell. The authors did not present new experimental results. Instead, they cited previous studies about ethical questions raised by large language models, including about the energy consumed by the tens or even thousands of powerful processors required when training such software, and the challenges of documenting potential biases in the vast data sets they were made with. BERT, Google's system, was mentioned more than a dozen times, but so was OpenAI's GPT-3. Mitchell considered the project worthwhile but figured it would come across as boring. An academic who saw the paper after it was submitted for publication found the document "middle of the road."

Plenty of people inside Google knew about the paper early on, including Dean. In October, he wrote in a glowing annual review that Gebru should work with other teams on developing techniques to make machine-learning software for language processing "consistent with our AI Principles." In her reply, she told him about the paper she was drafting with Bender and others. Dean wrote back: "Definitely not my area of expertise, but would definitely learn from reading it." Gebru also informed Google's communications department about the project and mentioned it to Marian Croak, a Black engineering executive on Google's Advanced Technology Review Council, an internal review panel that was added after the Maven protests. Croak said the paper sounded interesting and asked Gebru to send her a copy. But Gebru never got the chance before the fatal controversy over "Stochastic Parrots" erupted.

It's not clear exactly who decided that Gebru's paper had to be quashed or for what reason. Nor is it clear why her resistance—predictable as it was—prompted a snap decision to eject her, despite the clear risk of public fallout. Other researchers at Google say it isn't unusual for publications about AI to trigger internal corporate sensitivities before public release, but that researchers can usually work through managers' objections. Gebru, with her track record of rattling management about Google's diversity and AI ethics problems, got little such opportunity. One reason managers were not more open in explaining their feedback to Gebru, according to Google, was that they feared she would spread it around inside the company. Those fears may have been compounded when Gebru took to an internal listserv to criticize Google for "silencing marginalized voices," even as she offered to kill her own paper in exchange for greater transparency.

On the night of her forced exit from

6

Google, in early December, members of Gebru's team joined a tearful Google Meet video call that lasted until early the next morning. In normal times, they might have hugged and huddled in a bar or someone's home; in a pandemic they sniffled alone over their laptops. Two weeks later, the remaining team members sent an email to Google CEO Sundar Pichai demanding an apology and several changes, including Gebru's reinstatement (and Kacholia's reassignment). Mitchell's firing two months later brought new pain. She hired lawyers who blasted out a press release saying she had been fired after "raising concerns of race and gender equity at Google."

Dean became the face of Google's displeasure with the "Stochastic Parrots" paper. He sent an email to the members of Google Research, also released publicly, saying the work "didn't meet our bar for publication," in part because one of its eight sections didn't cite newer work showing that large language models could be made less energy-hungry. Dean repeated the point so often inside Google that some researchers joked that "I have an objection to Parrots section three" would be inscribed on his tombstone. The complaint made little sense to many AI researchers, who knew that grumbles about citations typically end with authors revising a paper, not getting terminated. Dean's argument suffered another blow when reviewers accepted the paper to the conference on fairness and technology.

Others, including Gebru, offered a different explanation from Dean's: Google had used an opaque internal process to suppress work critical of a technology that had commercial potential. "The closer the research started getting to search and ads, the more resistance there was," one Google employee with experience of the company's research review process says. "Those are the oldest and most entrenched organizations with the most power." Still others surmised that Gebru was the casualty of a different kind of turf battle: that other internal groups working on responsible AI—ones with closer relationships to Google's product teams—felt that Gebru and her coauthors were encroaching where they didn't belong.

THE REPERCUSSIONS OF Gebru's termination quickly radiated out from her team to the rest of Google and, beyond that, to the entire discipline of AI fairness research.

Some Google employees, including David Baker, a director who'd been at the company for 16 years, publicly quit over its treatment of Gebru. Google's research department was riven by mistrust and rumors about what happened and what might happen next. Even people who believed Gebru had behaved in ways unbecoming of a corporate researcher saw Google's response as ham-handed. Some researchers feared their work would now be policed more closely. One of them, Nicholas Carlini, sent a long internal email complaining of changes that company lawyers made to another paper involving large language models, published after Gebru was fired, likening the intervention to "Big Brother stepping in." The changes downplayed the problems the paper reported and removed references to Google's own technology, the email said.

Soon after, Google rolled out its response to the roiling scandal and sketched out a more locked-down future for in-house research probing AI's power. Marian Croak, the executive who had shown interest in Gebru's work, was given the task of consolidating the various teams working on what the company called responsible AI, including Mitchell and Gebru's. Dean sent around an email announcing that a review of Gebru's ouster had concluded; he was sorry, he said, that the company had not "handled this situation with more sensitivity."

Dean also announced that progress on improving workforce diversity would now be considered in top executives' performance reviews—perhaps quietly conceding Gebru's assertion that leaders were not held accountable for their poor showing on this count. And he informed researchers that they would be given firmer guidance on "Google's research goals and priorities." A Google source later explained that this meant future projects touching on sensitive or commercial topics would require more input from in-house legal experts, product teams, and others within Google who had relevant expertise. The outlook for open-minded, independent research on ethical AI appeared gloomy. Google claimed that it still had hundreds of people working on responsible AI, and that it would expand those teams; the company painted Gebru and Mitchell's group as a tiny and relatively unimportant cog in a big machine. But others at Google said the Ethical AI leaders and their frank feedback would be missed. "For me, it's the most critical voices that are the most important and where I have learned the most," says one person who worked on product changes with Gebru and Mitchell's input. Bengio, the women's manager, turned his back on 14 years of working on AI at Google and quit to join Apple.

Outside of Google, nine Democrats in Congress wrote to Pichai questioning his commitment to preventing AI's harms. Mitchell had at one point tried to save the "Stochastic Parrots" paper by telling executives that publishing it would bolster arguments that the company was capable of self-policing. Quashing it was now undermining those arguments.

Some academics announced that they had backed away from company events or funding. The fairness and technology in AI conference's organizers stripped Google of its status as a sponsor of the event. Luke Stark, who studies the social impacts of AI at the University of Western Ontario, turned down a $60,000 grant from Google in protest of its treatment of the Ethical AI team. When he applied for the money in December 2020, he had considered the team a "strong example" of how corporate researchers could do powerful work. Now he wanted nothing to do with Google. Tensions built into the field

of AI ethics, he saw, were beginning to cause fractures.

"The big tech companies tried to steal a march on regulators and public criticism by embracing the idea of AI ethics," Stark says. But as the research matured, it raised bigger questions. "Companies became less able to coexist with internal critical research," he says. One person who runs an ethical AI team at another tech company agrees. "Google and most places did not count on the field becoming what it did."

To some, the drama at Google suggested that researchers on corporate payrolls should be subject to different rules than those from institutions not seeking to profit from AI. In April, some founding editors of a new journal of AI ethics published a paper calling for industry researchers to disclose who vetted their work and how, and for whistle-blowing mechanisms to be set up inside corporate labs. "We had been trying to poke on this issue already, but when Timnit got fired it catapulted into a more mainstream conversation," says Savannah Thais, a researcher at Princeton on the journal's board who contributed to the paper. "Now a lot more people are questioning: Is it possible to do good ethics research in a corporate AI setting?"

If that mindset takes hold, in-house ethical AI research may forever be held in suspicion—much the way industrial research on pollution is viewed by environmental scientists. Jeff Dean admitted in a May interview with CNET that the company had suffered a real "reputational hit" among people interested in AI ethics work. The rest of the interview dealt mainly with promoting Google's annual developer conference, where it was soon announced that large language models, the subject of Gebru's fateful critique, would play a more central role in Google search and the company's voice assistant. Meredith Whittaker, faculty director of New York University's AI Now Institute, predicts that there will be a clearer split between work done at institutions like her own and work done inside tech companies. "What Google just said to anyone who wants to do this critical research is, 'We're not going to tolerate it,'" she says. (Whittaker herself once worked at Google, where she clashed with management over AI ethics and the Maven Pentagon contract before leaving in 2019.)

Any such divide is unlikely to be neat, given how the field of AI ethics sprouted in a tech industry hothouse. The community is still small, and jobs outside big companies are sparser and much less well paid, particularly for candidates without computer science PhDs. That's in part because AI ethics straddles the established boundaries of academic departments. Government and philanthropic funding is no match for corporate purses, and few institutions can rustle up the data and computing power needed to match work from companies like Google.

For Gebru and her fellow travelers, the past five years have been vertiginous. For a time, the period seemed revolutionary: Tech companies were proactively exploring flaws in AI, their latest moneymaking marvel—a sharp contrast to how they'd faced up to problems like spam and social network moderation only after coming under external pressure. But now it appeared that not much had changed after all, even if many individuals had good intentions.

Inioluwa Deborah Raji, whom Gebru escorted to Black in AI in 2017, and who now works as a fellow at the Mozilla Foundation, says that Google's treatment of its own researchers demands a permanent shift in perceptions. "There was this hope that some level of self-regulation could have happened at these tech companies," Raji says. "Everyone's now aware that the true accountability needs to come from the outside—if you're on the inside, there's a limit to how much you can protect people."

Gebru, who recently returned home after her unexpectedly eventful road trip, has come to a similar conclusion. She's raising money to launch an independent research institute modeled on her work on Google's Ethical AI team and her experience in Black in AI. "We need more support for external work so that the choice is not 'Do I get paid by the DOD or by Google?'" she says.

Gebru has had offers, but she can't imagine working within the industry anytime in the near future. She's been thinking back to conversations she'd had with a friend who warned her not to join Google, saying it was harmful to women and impossible to change. Gebru had disagreed, claiming she could nudge things, just a little, toward a more beneficial path. "I kept on arguing with her," Gebru says. Now, she says, she concedes the point. W

COLOPHON

Dilemmas That Helped Get This Issue Out:

Who to invite to my mask-burning bonfire; whether to attend my friend's Zoom improv show; watching a pineapple go bad versus making the effort to cut it up; whether to feed the alley cats and annoy my feline-averse neighbors; how to reply politely to emails that use more than one exclamation point; subjecting my Instagram followers to yet another photo of my dog (yes, obviously); a friend lied about his kid's age so they could get vaxed early—do I call him out?; when to accept a Twitter tip; walking away from cat vomit in the living room so I don't have to be the one to clean it up (this time); navigating mask wearage among four levels of airflow, three generations, and two vaccine statuses.

WIRED is a registered trademark of Advance Magazine Publishers Inc. Copyright ©2021 Condé Nast. All rights reserved. Printed in the USA. Volume 29, No. 7. WIRED (ISSN 1059–1028) is published monthly, except for combined issues in December/January and July/August, by Condé Nast, which is a division of Advance Magazine Publishers Inc. Editorial office: 520 Third Street, Ste. 305, San Francisco, CA 94107-1815. Principal office: Condé Nast, 1 World Trade Center, New York, NY 10007. Roger Lynch, Chief Executive Officer; Pamela Drucker Mann, Chief Revenue & Marketing Officer, US; Jackie Marks, Chief Financial Officer. Periodicals postage paid at New York, NY, and at additional mailing offices. Canada Post Publications Mail Agreement No.40644503. Canadian Goods and Services Tax Registration No. 123242885 RT0001.

POSTMASTER: Send all UAA to CFS (see DMM 707.4.12.5); NONPOSTAL AND MILITARY FACILITIES: Send address corrections to WIRED, PO Box 37617, Boone, IA 50037-0662. For subscriptions, address changes, adjustments, or back issue inquiries: Please write to WIRED, PO Box 37617, Boone, IA 50037-0662, call (800) 769 4733, or email subscriptions@wired.com. Please give both new and old addresses as printed on most recent label. First copy of new subscription will be mailed within eight weeks after receipt of order. Address all editorial, business, and production correspondence to WIRED Magazine, 1 World Trade Center, New York, NY 10007. For permissions and reprint requests, please call (212) 630 5656 or fax requests to (212) 630 5883. Visit us online at www.wired.com. To subscribe to other Condé Nast magazines on the web, visit wired.condenet.com. Occasionally, we make our subscriber list available to carefully screened companies that offer products and services that we believe would interest our readers. If you do not want to receive these offers and/or information, please advise us at PO Box 37617, Boone, IA 50037-0662, or call (800) 769 4733.

WIRED is not responsible for the return or loss of, or for damage or any other injury to, unsolicited manuscripts, unsolicited artwork (including, but not limited to, drawings, photographs, and transparencies), or any other unsolicited materials. Those submitting manuscripts, photographs, artwork, or other materials for consideration should not send originals, unless specifically requested to do so by WIRED in writing. Manuscripts, photographs, artwork, and other materials submitted must be accompanied by a self-addressed, stamped envelope.