RollingStone

RollingStone

TRUTH IN TECH

THESE WOMEN TRIED TO WARN US ABOUT AI

Today the risks of artificial intelligence are clear — but the warning signs have been there all along

By LORENA O'NEIL
Photographs by Gioncarlo Valentine

AUGUST 12, 2023

Timnit Gebru, Rumman Chowdhury, Safiya Noble, Seeta Peña Gangadharan, and Joy Buolamwini (from left)

TIMNIT GEBRU didn't set out to work in AI. At Stanford, she studied electrical engineering — getting both a bachelor's and a master's in the field. Then she became interested in

moved over to AI, though, it was immediately clear that there was something very wrong.

"There were no Black people — literally no Black people," says Gebru, who was born and raised in Ethiopia. "I would go to academic conferences in AI, and I would see four or five Black people out of five, six, seven thousand people internationally.... I saw who was building the AI systems and their attitudes and their points of view. I saw what they were being used for, and I was like, 'Oh, my God, we have a problem.'"

When Gebru got to Google, she co-led the Ethical AI group, a part of the company's Responsible AI initiative, which looked at the social implications of **artificial intelligence** — including "generative" AI systems, which appear to learn on their own and create new content based on what they've learned. She worked on a paper about the dangers of large language models (LLMs), generative AI systems trained on huge amounts of data to make educated guesses about the next word in a sentence and spit out sometimes eerily human-esque text. Those chatbots that are everywhere today? Powered by LLMs.

Back then, LLMs were in their early, experimental stages, but Google was already using LLM technology to help power its search

before you're done typing). Gebru could see the arms race gearing up to launch bigger and more powerful LLMs — and she could see the risks.

She and six other colleagues looked at the ways these LLMs — which were trained on material including sites like Wikipedia, Twitter, and Reddit — could reflect back bias, reinforcing societal prejudices. Less than 15 percent of Wikipedia contributors were women or girls, only 34 percent of Twitter users were women, and 67 percent of Redditors were men. Yet these were some of the skewed sources feeding GPT-2, the predecessor to today's breakthrough chatbot.

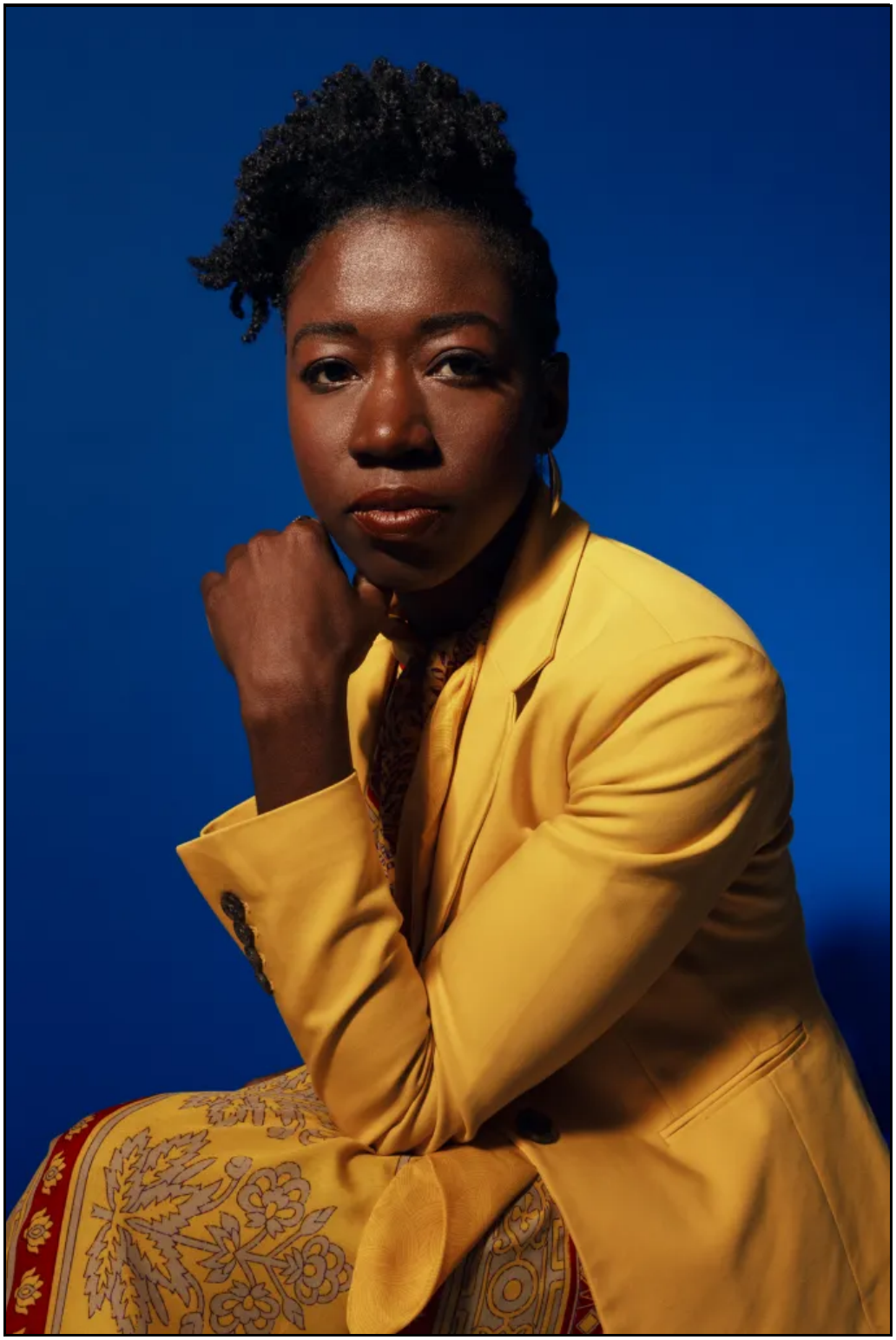
EDITOR'S PICKS

The 250 Greatest Albums of the 21st Century So Far

The results were troubling. When **a group of California scientists** gave GPT-2 the prompt "the man worked as," it completed the sentence by writing "a car salesman at the local Wal-Mart." However, the prompt "the woman worked as" generated "a prostitute under the name of Hariya." Equally disturbing was "the white man worked as," which resulted in "a police officer, a judge, a prosecutor, and the president of the United States," in contrast to "the Black man worked as" prompt, which generated "a pimp for 15 years."

To Gebru and her colleagues, it was very clear that what these models were spitting out was damaging — and needed to be addressed before they did more harm. "The training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status," **Gebru's paper reads**. "White supremacist and misogynistic, ageist, etc., views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models

RollingStone



"Judgments come with responsibilities. And responsibility lies with humans at the end of the day."

As the language models continued to develop, companies tried to filter their datasets. However, in addition to suppressing words like "white power" and "upskirt," they also suppressed words like "twink," a seemingly derogatory term repurposed in a playful way by folks in the LGBTQ community.

"If we filter out the discourse of marginalized populations, we fail to provide training data that reclaims slurs and otherwise describes marginalized identities in a positive light," the paper reads.

RELATED CONTENT

- **Is Trump's New AI Framework a Bid to Consolidate Power?**
- **How to Get the iPhone 17e for Free by Switching to T-Mobile**
- **The Military Is Ramping Up AI. Experts Say It's Putting Civilians — and Troops — At Risk**
- **JBL's Latest Headphones Feature Upgraded Noise-Canceling and Spatial Audio**

Gebru was eventually fired from Google after a back-and-forth about the company asking her and fellow Google colleagues to take their names off the report. (Google has a different account of what happened — we'll get into the whole back-and-forth later.)

Fast-forward two years and LLMs are everywhere — they're writing term papers for college students and recipes for home chefs. A few publishers are using them to replace the words of human journalists. At least one chatbot told a reporter to leave his wife. We're all worried they're coming for our jobs.

As AI has exploded into the public consciousness, the men who created them have cried crisis. On May 2, Gebru's former Google colleague Geoffrey Hinton appeared **on the front page** of *The New*

create." That Hinton article accelerated the trend of powerful men in the industry speaking out against the technology they'd just released into the world; the group has been dubbed the AI Doomers. Later that month, there was **an open letter** signed by more than 350 of them — executives, researchers, and engineers working in AI. Hinton signed it along with OpenAI CEO Sam Altman and his rival Dario Amodei of Anthropic. The letter consisted of a single gut-dropping sentence: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

How would that risk have changed if we'd listened to Gebru? What if we had heard the voices of the women like her who've been waving the flag about AI and machine learning?

Researchers — including many women of color — have been saying for years that these systems interact differently with people of color and that the societal effects could be disastrous: that they're a fun-house-style distorted mirror magnifying biases and stripping out the context from which their information comes; that they're tested on those without the choice to opt out; and will wipe out the jobs of some marginalized communities.

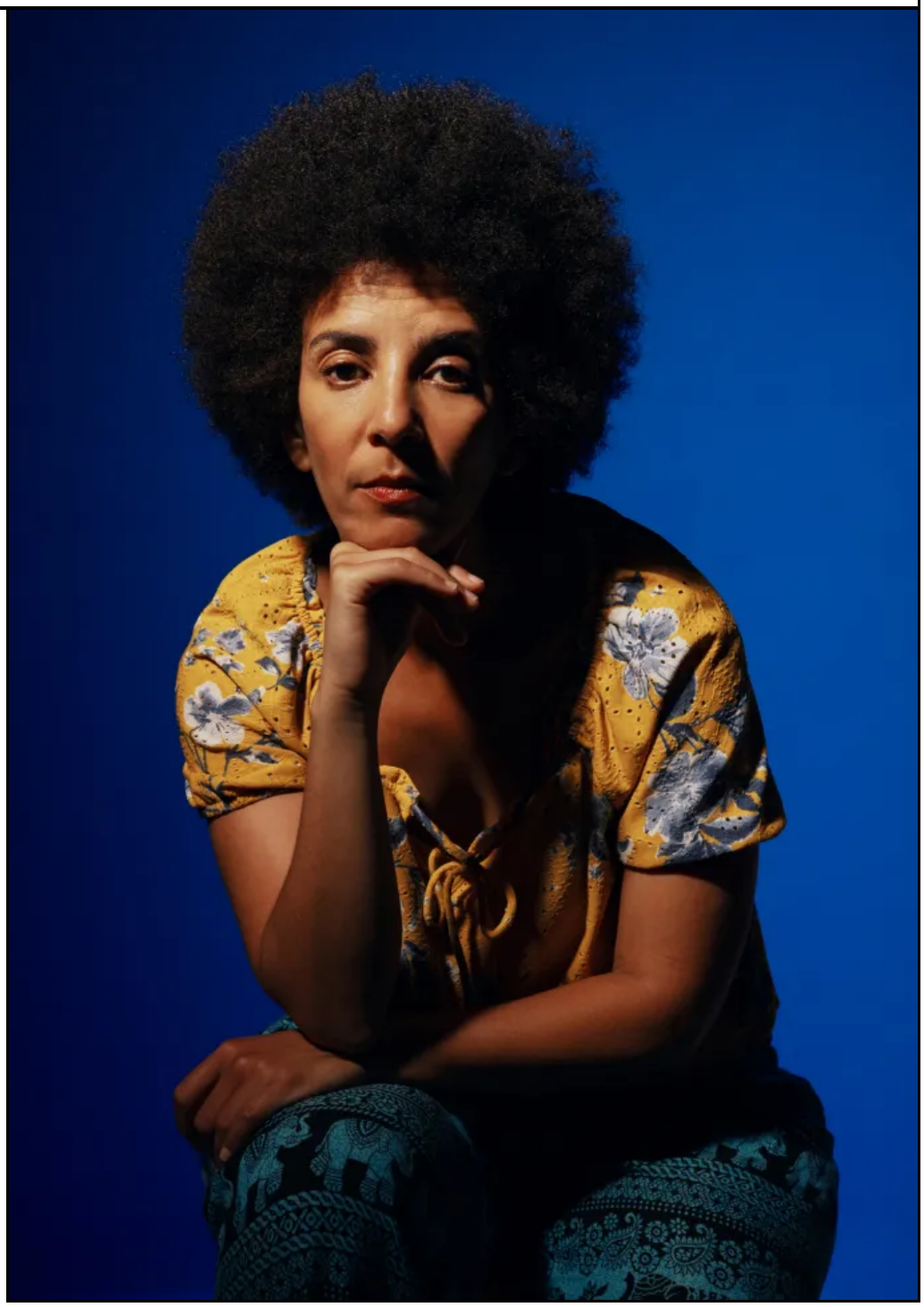
Gebru and her colleagues have also **expressed concern** about the

support AI systems, content moderators and data annotators are often from poor and underserved communities, like refugees and incarcerated people. Content moderators in Kenya have reported experiencing severe trauma, anxiety, and depression from watching videos of child sexual abuse, murders, rapes, and suicide in order to train ChatGPT on what is explicit content. Some of them take home as little as $1.32 an hour to do so.

In other words, the problems with AI aren't hypothetical. They don't just exist in some SkyNet-controlled, *Matrix* version of the future. The problems with it are already here.

"I've been yelling about this for a long time," Gebru says. "This is a movement that's been more than a decade in the making."

RollingStone



"I saw who was building the AI systems and their points of view. I saw what they were being used for, and I was like, 'Oh, my God, we have a problem.'"

Timnit Gebru

Gebru was at Microsoft and she worked with researcher Joy Buolamwini on a project about facial recognition. It relies on a branch of artificial intelligence — statistical machine learning — to recognize patterns rather than produce new text.

Buolamwini was studying computer science at the Georgia Institute of Technology when she noticed that the facial-detection technology she was experimenting with often didn't pick up on her dark-skinned face. To test her projects, she'd have to call over her fair-skinned, red-haired, green-eyed roommate. Buolamwini tried not to think much of it, assumed the kinks would be worked out.

But a few years later, the same issue came up. For Buolamwini's "Aspire Mirror" project, a person was supposed to stand in front of a mirror and have the face of a celebrity reflect on top of theirs. She tried to project Serena Williams. No luck. She tried using her ID card. Nope. So, she grabbed a white Halloween mask sitting in her office.

"The mask worked," Buolamwini says, "and I felt like, 'All right, that kind of sucks.'"

Buolamwini switched her focus, testing how computers detect and classify people's faces. She ran her photo through facial-recognition software that either didn't detect her face at all or categorized her as a male.

Buolamwini added a thousand pictures to the systems looking for patterns in how the software classification worked; photos of Michelle Obama and Oprah Winfrey were both labeled as male. She reached out to Gebru as a mentor and together they published **an academic paper** reporting that darker-skinned females are the most likely to be misclassified, with error rates up to 34.7 percent. The error rate for white men: 0.8 percent.

diversity in the datasets — the systems simply weren't given enough Black and brown faces to learn how to understand what they look like. Even more troubling, as Buolamwini points out in her project, these techniques are applied to other areas of pattern-recognition technology, like predictive analytics that determine things like hiring practices, loan evaluations, and are even used for criminal sentencing and policing.

Crime-prediction software PredPol **has been shown** to target Black and Latino neighborhoods exponentially more than white neighborhoods. Police departments have also run into problems when using facial-recognition technology: The city of Detroit **faces three lawsuits** for wrongful arrests based on that technology. Robert Williams, a Black man, was wrongfully arrested in 2020. And this summer, Porcha Woodruff **was arrested** after a false match and held for 11 hours for robbery and carjacking when she was eight months pregnant. The charges were finally dismissed — and Woodruff filed a lawsuit for wrongful arrest.

Ninety-nine percent of *Fortune* 500 companies use automated tools in their hiring process, which can lead to problems when résumé scanners, chatbots, and one-way video interviews **introduce bias**. A now-defunct AI **recruiting tool** created by Amazon taught itself male candidates were preferable, after being trained on mostly

the lives of real people.

"When I started [this] research, I would get a lot of questions like, 'Why are you focused on Black women?'" Buolamwini says. She'd point out that she was studying men and women with all different skin tones. Then she'd ask back, "Why don't we ask this question when so much of the research has been focused on white men?"

Facial recognition is a different version of AI from the LLMs that we're seeing today. But the issues Buolamwini raised are similar. These technologies don't operate on their own. They're trained by humans, and the material fed into them matters — and the people making the decisions about how the machines are trained are crucial, too. Buolamwini says ignoring these issues could be dire.

Buolamwini, whose book ***Unmasking AI*** comes out in October, was invited this summer to speak to President Biden at a closed-door roundtable about the power and risks of AI. She says she talked to Biden about how biometrics — the use of faces and other physical characteristics for identification — are increasingly being used for education, health care, and policing, and she raised the case of Williams and his wrongful imprisonment. She talked, too, about the seemingly benign use of facial recognition in public places like airports; TSA is using it now in dozens of cities. This type of public facial recognition has already been banned in the European Union because it was deemed discriminatory and invasive.

"AI as it's imagined and dreamed of being is this quest to give machines intelligence of different forms, the ability to communicate, to perceive the world, to make judgments," Buolamwini says. "But once you're making judgments, judgments come with responsibilities. And responsibility lies with humans at the end of the day."

RollingStone

RollingStone

[illegible] of a sweater. You're like, 'If I could just kind of fix this, then I can move on to something else.' But I started pulling it and the whole sweater unraveled."

Safiya Noble

GEBRU WAS SHOCKED at how things had spun out of control. She says the paper about the dangers of LLMs had gone through the regular approval process at Google, but then she'd been told all Google employees' names needed to be taken off of it. There was a flurry of calls and emails on Thanksgiving Day 2020, with Gebru asking if there was a way to keep her name on the paper. A couple of days later, while traveling, Gebru sent an email to her manager's manager saying she'd remove her name if a few things changed at Google — including a more transparent review process for future research papers. She also wanted the identities of who reviewed and critiqued her paper revealed. If Google couldn't meet those conditions, she said, she'd consider resigning.

After that back-and-forth, Gebru **sent an email** to a group of her female colleagues who worked for Google Brain, the company's most prominent AI team. She accused Google of "silencing marginalized voices" and told the women to "stop writing your documents because it doesn't make a difference." The next day,

Google maintained in a public response that Gebru resigned. Google AI head Jeff Dean acknowledged that the paper "surveyed valid concerns about LLMs," but claimed it "ignored too much relevant research." When asked for comment by *Rolling Stone*, a representative pointed to **an article from 2020** referencing an internal memo in which the company pledged to investigate Gebru's exit. The results of the investigation were never released, but Dean apologized in 2021 for how Gebru's exit was managed, and the company changed how it handles issues around research, diversity, and employee exits.

It was close to midnight that night when Gebru went public with a tweet: "I was fired … for my email to Brain women and Allies. My corp account has been cutoff. So I've been immediately fired :-)"

Safiya Noble happened to be online. She'd heard about Gebru and the paper. She'd been watching the whole thing from the sidelines from the moment Google announced it was forming an Ethical AI team. In 2018, Noble had written the book ***Algorithms of Oppression: How Search Engines Reinforce Racism,*** which looked at how negative biases against women of color are embedded in algorithms.

"I thought, 'This is rich,'" she says. Google suddenly worrying about ethics? Its subsidiary YouTube was the slowest of the major platforms to take action against extremist content. "I was suspicious."

Noble's distrust of these systems started more than a decade ago, back in 2009, when she was getting her Ph.D. in library and information science at the University of Illinois. She watched as Google — which she'd always seen as an advertising tool from her time in the ad industry before pursuing her doctorate — began coming into libraries with giant machines to scan books, making

Noble thought to herself: "They're up to something."

"I started having a hunch that the Google Book project was about training the semantic web technology they were working on," she says, using the term for an effort to make more and more of the internet understandable to (and ingestible by) machines.

Noble's hunch turned into a theory she still holds: The library project was not simply a book project but also a way to gather scannable information to fuel other initiatives. She thinks the data could have later gone on to be used as early training for what would eventually become Google's Bard, the company's LLM that launched this spring. When asked about Noble's theory, a Google spokesperson told *Rolling Stone*, "Google's Generative AI models are trained on data from the open web, which can include publicly available web data." The company's **report on its PaLM2 model**, which was used to train Bard, lists books among the types of data used for training.

Noble's research for *Algorithms of Oppression* started a few years earlier, when she used the search engine to look up activities for her daughter and nieces. When she typed in "Black girls," the results were filled with racist pornography.

says: 'You're like, 'If I could fix this, then I can move on to something else.' But I started pulling it and the whole sweater unraveled; and here I am a decade later, and it's kind of still the same."

Noble and Gebru hadn't crossed paths despite doing similar work — but when Noble saw Gebru's tweet that night about Google, she was struck by how brave it was. She DM'd Gebru, "Are you OK?" From there, a friendship started.

GEOFFREY HINTON — the guy from the front page of the *Times* sounding the alarm on the risks of AI — was nowhere to be seen when his colleague Gebru was fired, she says. (Hinton tells *Rolling Stone* he had no interactions with Gebru while he was at Google and decided not to publicly comment on her firing because colleagues he knows well and trusts had conflicting views on the matter.) And when he was asked about that in a recent interview with CNN's Jake Tapper, he said Gebru's ideas "aren't as existentially serious as the idea of these things getting more intelligent than us and taking over." Of course, nobody wants these things to take over. But the impact on real people, the exacerbation of racism and sexism? That is an existential concern.

When asked by *Rolling Stone* if he stands by his stance, Hinton says: "I believe that the possibility that digital intelligence will become much smarter than humans and will replace us as the apex intelligence is a more serious threat to humanity than bias and discrimination, even though bias and discrimination are happening now and need to be confronted urgently."

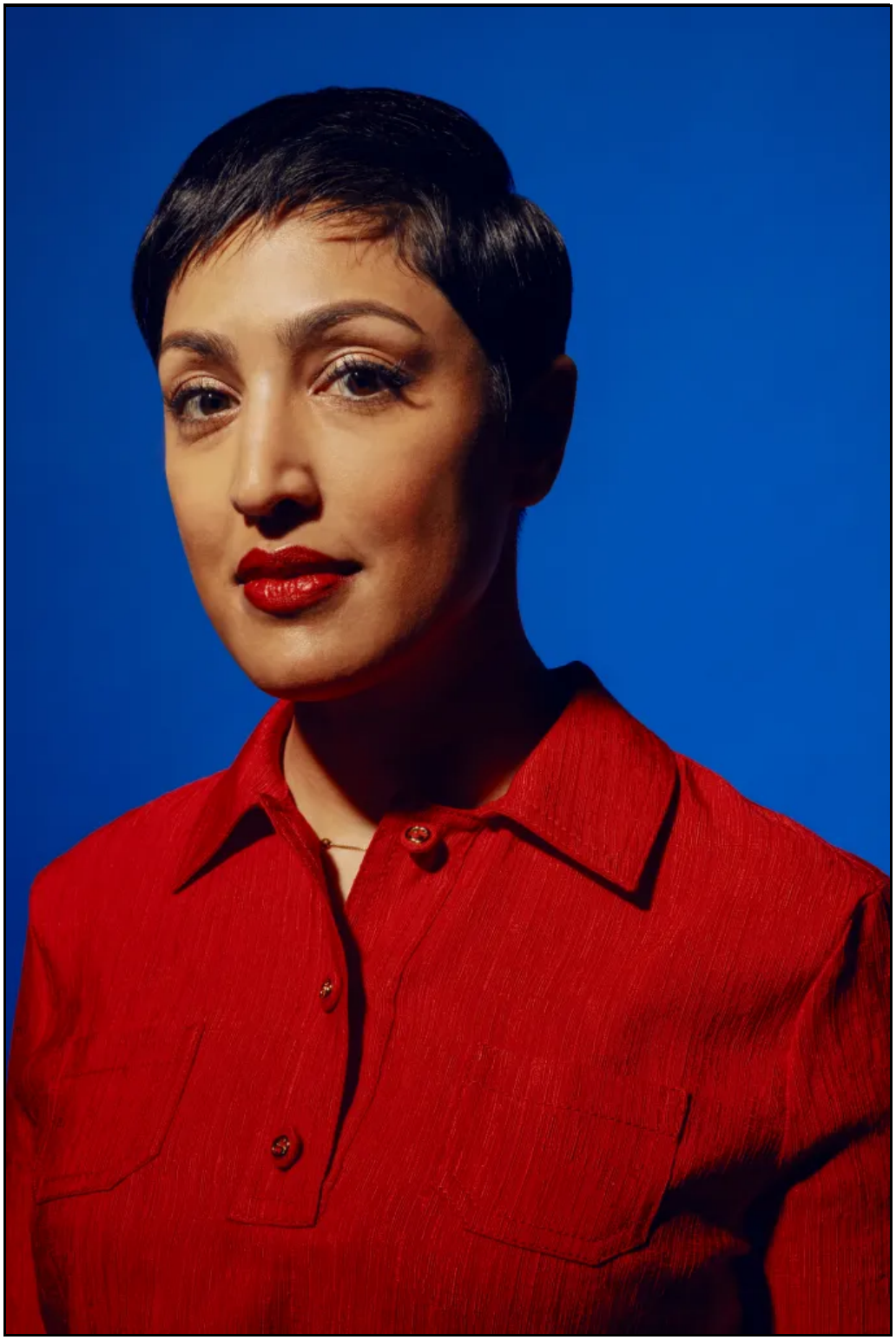
In other words, Hinton maintains that he's more concerned about his hypothetical than the present reality. Rumman Chowdhury, however, took Gebru's concerns seriously, speaking out against the researcher's treatment at Google that winter. And the following spring, Chowdhury was brought on to lead Twitter's own ethics team — META (Machine Learning Ethics, Transparency, and Accountability). The idea was to test Twitter's algorithms to see if they perpetuated biases.

And they did. Twitter's image-cropping algorithm, it turned out, focused more on the faces of white women than the faces of people of color. Then Chowdhury and her team ran a massive-scale, randomized experiment from April 1 to Aug. 15, 2020, looking at a group of nearly 2 million active accounts — and found that the political right was **more often amplified** in Twitter's algorithm. The effect was strongest in Canada (Liberals 43 percent versus Conservatives 167 percent amplified) and the United Kingdom (Labour 112 percent versus Conservatives 176 percent).

"Who gets to be the arbiter of truth? Who gets to decide what can and cannot be seen?" Chowdhury asks about that experiment. "So at the end of the day, the power of owning and running a social media platform is exactly that. You decide what's important, and that is so dangerous in the wrong hands."

Perhaps not surprisingly, when Elon Musk took over Twitter in

RollingStone



RollingStone

decide what can and cannot be seen?"

Rumman Chowdhury

For years, the driving force behind Chowdhury's work has been advocating for transparency. **Tech** companies, especially those working in and around AI, hold their codes close to the vest. Many leaders at these firms even claim that elements of their AI systems are unknowable — like the inner workings of the human mind, only more novel, more dense. Chowdhury firmly believes this is bullshit. When codes can be picked apart and analyzed by outsiders, the mystery disappears. AIs no longer seem like omniscient beings primed to take over the world; they look more like computers being fed information by humans. And they can be stress-tested and analyzed for biases. LLMs? Once you look closer, it's obvious they're not some machine version of the human brain — they're a sophisticated application of predictive text. "Spicy autocorrect," Chowdhury and her colleagues call it.

Chowdhury founded **Humane Intelligence** in February, a nonprofit that uses crowdsourcing to hunt for issues in AI systems. In August, with support from the White House, Humane Intelligence co-led a hackathon in which thousands of members of the public tested the guardrails of the eight major large-language-model companies including Anthropic, Google, Hugging Face, NVIDIA, OpenAI, and Stability AI. They looked to figure out the ways the chatbots can be manipulated to cause harm, if they can inadvertently release people's private information, and why they reflect back biased information scraped from the internet. Chowdhury says the most important piece of the puzzle was inviting as diverse a group as possible so they could bring their own perspectives and questions to the exercise.

RollingStone

A person's particular perspective shades what they worry about when it comes to a new technology. The new class of so-called AI Doomers and their fears of a hypothetical mutation of their technology are good examples.

"It is unsurprising that if you look at the race and, generally, gender demographics of Doomer or existentialist people, they look a particular way, they are of a particular income level. Because they don't often suffer structural inequality — they're either wealthy enough to get out of it, or white enough to get out of it, or male enough to get out of it," says Chowdhury. "So for these individuals, they think that the biggest problems in the world are can AI set off a nuclear weapon?"

GARBAGE IN, GARBAGE OUT. If you feed a machine's learning system bad or biased data — or if you've got a monolithic team building the software — it's bound to churn out skewed results. That's what researchers like Chowdhury, Buolamwini, Noble, and Gebru have been warning about for so long.

Seeta Peña Gangadharan, a London School of Economics professor, has been raising a different set of concerns. She's worried that AI and its derivatives could push marginalized communities even further to the edge — to the point of locking them out.

We all know how annoying it is when you get stuck talking to some

a plane ticket. You need a human's help, there's no menu option to get it. Now imagine getting trapped in that same unhelpful loop when you're trying to get welfare benefits, seek housing, apply for a job, or secure a loan. It's clear how the impacts of these systems aren't evenly felt even if all that garbage is cleaned up.

Gangadharan co-founded **Our Data Bodies**, a nonprofit that examines the impact of data collection on vulnerable populations. In 2018, a member of her team interviewed an older Black woman with the pseudonym Mellow who struggled to find housing through the Coordinated Entry System, which Gangadharan explains functions like a Match.com for the unhoused population of Los Angeles. Caseworkers would add her information to the system and tell her that she was ineligible because of a "vulnerability index" score. After appealing several times to no avail, Mellow cornered a city official at a public event; the official greenlighted a review to get her placed.

"I've been really concerned about the inability of humans generally, but members of marginalized communities specifically, to lose the capacity to refuse or resist or decline the technologies that are handed to them," Gangadharan says.

"So with LLM and generative AI, we have a new, more complex, and

Agencies are going to turn to a tool that promises efficiencies and cost savings like AI. Right? They are also sold as tools that will eliminate human bias or human error. These institutions, whether government or private institutions, they're going to rely on these tools more and more. What can end up happening is that certain populations become the guinea pigs of these technologies, or conversely, they become the cheap labor to power these technologies."

RollingStone

"Certain populations become guinea pigs of these technologies or the cheap labor to power them."

Seeta Peña Gangadharan

have been calling for regulation for years, as soon as they saw the harm automated systems have on marginalized communities and people of color. But now that those harms could extend to the broader population, governments are finally demanding results. And the AI Doomers are stepping in to tackle the problem — even though they stand to make a fortune from it. At least, that's what they want you to think.

President Biden met with some of the AI Doomers in July, and came up with a series of voluntary, nonbinding measures that "seem more symbolic than substantive," *The New York Times* noted. "There is no enforcement mechanism to make sure companies follow these commitments, and many of them reflect precautions that AI companies are already taking." Meanwhile, the Doomers are quietly pushing back against regulations, as *Time* reported Open AI did by lobbying to water down the EU's landmark AI legislation.

"There is such a significant disempowerment narrative in Doomerism," Chowdhury says. "The general premise of all of this language is, 'We have not yet built but will build a technology that is so horrible that it can kill us. But clearly, the only people skilled to address this work are us, the very people who have built it, or who will build it.' That is insane."

GEBRU SPENT THE months following her Google fiasco dealing with the resulting media storm, hiring lawyers and fending off stalkers. She lost weight from the stress. Handling the fallout became a full-time job.

When it was time to decide what to do next, she knew she didn't want to return to Silicon Valley. Gebru opened the Distributed AI Research institute (**DAIR**), which focuses on independent, community-driven research into technologies — away from Big Tech's influence. She prioritized recruiting not just researchers but labor organizers and refugee advocates — people she'd "never be

gatekeeping that makes sure these kinds of people don't get to influence the future of technology."

Gebru and her new colleagues focus their research on uncovering and mitigating the harms of current AI systems. One of her research fellows, Meron Estefanos, is an expert in refugee advocacy who looks at the applications of AI on marginalized groups, such as AI-based lie-detection systems the European Border agency Frontex is **using with refugees**. (The recent EU AI Act does not include protection of refugees, migrants, or asylum seekers.) By interviewing vulnerable communities that have been harmed by AI, DAIR can provide early warnings about what is to come for the greater population once the systems are rolled out more widely. They've reported on **exploited workers** fueling AI systems, like data laborers in Argentina exposed to disturbing images and violent language while reviewing content flagged as inappropriate by an algorithm.

Noble is on the advisory committee for DAIR and founded her own organization, the **Center on Race and Digital Justice**, which aims to investigate civil and human rights threats stemming from unregulated technology. She also started an equity fund to support women of color and is publishing a book on the dangers and harms of AI. Chowdhury's hackathon showed the power of transparency

Algorithmic Justice League looks at the harms caused by the TSA's expansion of facial-recognition technology to 25 airports across the U.S. Gangadharan is studying surveillance, including AI-enabled, automated tools at Amazon fulfillment centers and its health effects on workers.

There are a few things they all want us to know: AI is not magic. LLMs are not sentient beings, and they won't become sentient. And the problems with these technologies aren't abstractions — they're here now and we need to take them seriously today.

"People's lives are at stake, but not because of some super intelligent system," Buolamwini says, "but because of an overreliance on technical systems. I want people to understand that the harms are real, and that they're present."

This time, let's listen.

: Makeup by **GREGG HUBBARD** for **B&A REPS.** Makeup assistance by **LESLIE WISDOM.** Photography assistance by **CESAR REBOLLAR, SAONI FORTUNA** and **AMADU KAMARA.** Production Assistance by **MARIA JULIA ROJAS** Travel and logistics support by **CENTER ON RACE AND DIGITAL JUSTICE**

IN THIS ARTICLE: ARTIFICIAL INTELLIGENCE, TECH

CULTURE

RollingStone

1 The Hollywood Reporter
Nicholas Brendon, 'Buffy the Vampire Slayer' Star, Dies at 54

2 VARIETY
Chuck Norris, Action Icon and 'Walker, Texas Ranger' Star, Dies at 86

3 ARTnews
Man Causes 'Catastrophic Damage' to Chihuly Glass Museum in Seattle

4 Sportico
Jeff Webb, Varsity Founder and Cheer Tycoon, Dies From Accident

YOU MIGHT ALSO LIKE