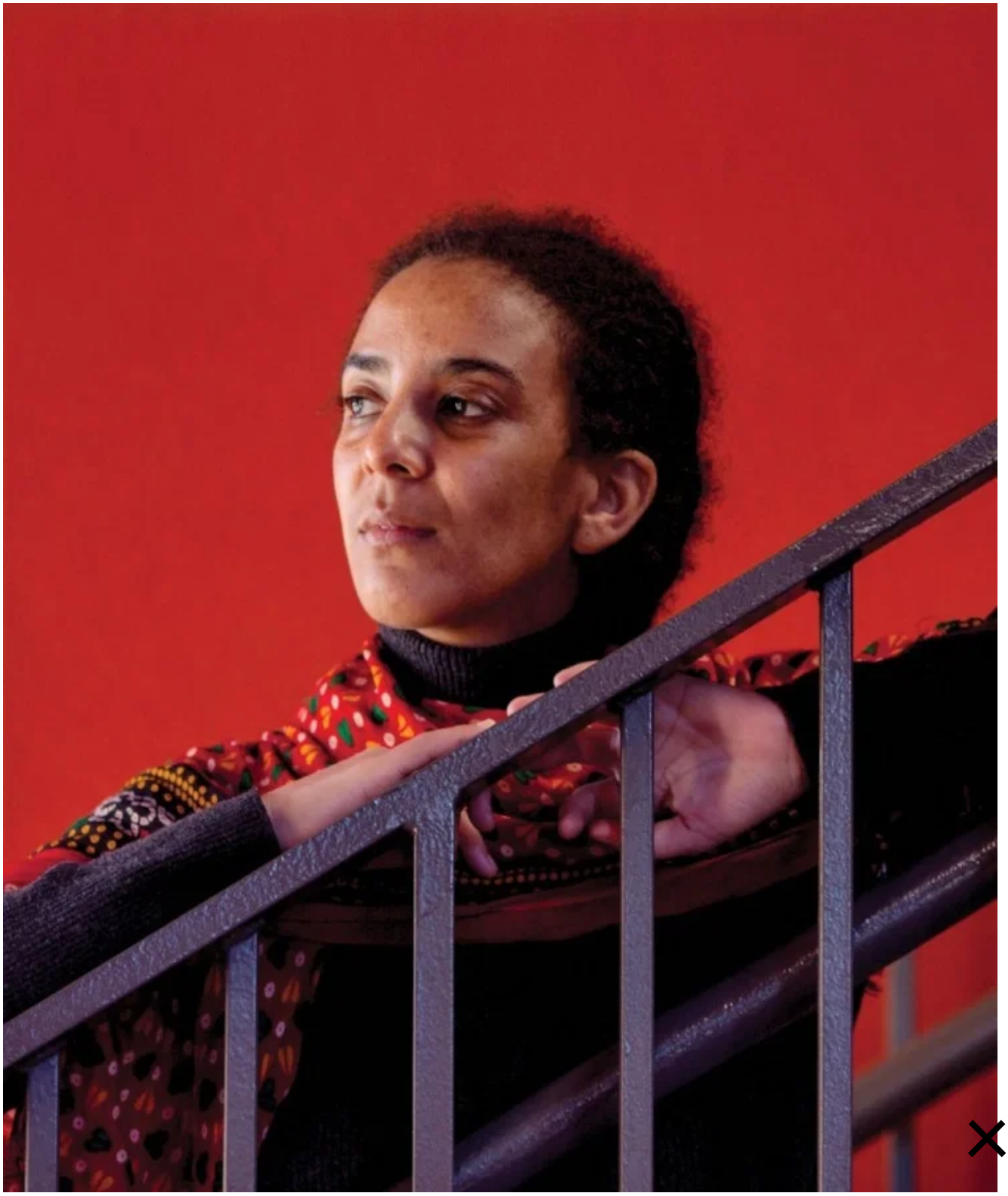
Artificial intelligence / Machine learning

We read the paper that forced Timnit Gebru out of Google. Here's what it says

The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business.

by **Karen Hao** December 4, 2020



COURTESY OF TIMNIT GEBRU

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 co-authored. 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 and 1,900 other supporters have also signed a letter of protest.

comment beyond their posts on social media. But MIT Technology Review obtained a copy of the research paper from one of the co-authors, 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.

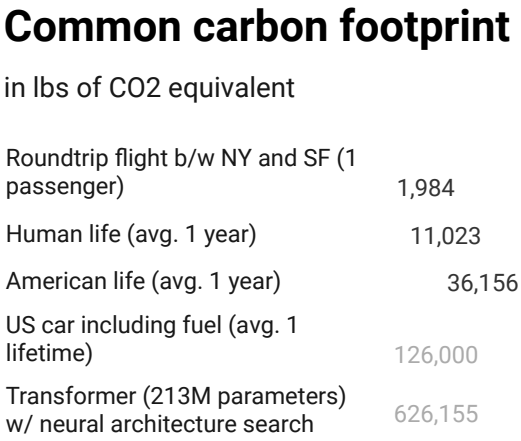
Titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" the paper lays out the risks of large language models—AIs trained on staggering amounts of text data. These have grown increasingly popular—and increasingly large—in the last three years. They are now extraordinarily good, 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 off 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



Strubell's study found that 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. 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 roundtrip flight between New York City and San Francisco.

The estimated costs of training a model

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud comput (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,57
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,2
GPT-2	Feb, 2019	-	-	$12,902-$43,0

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

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

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 datasets 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 datasets (and thus also use less energy).

Illusions of meaning

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 co-authors, 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.)

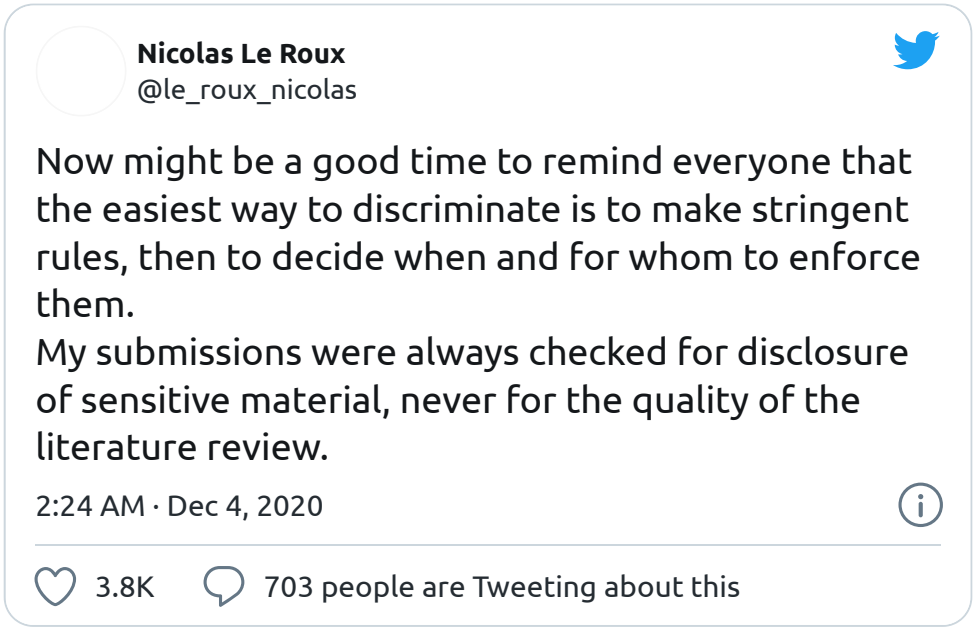
The paper's goal, Bender says, was to take stock of the landscape of current research in natural-language processing. "We are working at a scale where the people building the things can't actually get their arms around the data," she said. "And because the upsides are so obvious, it's particularly important 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.

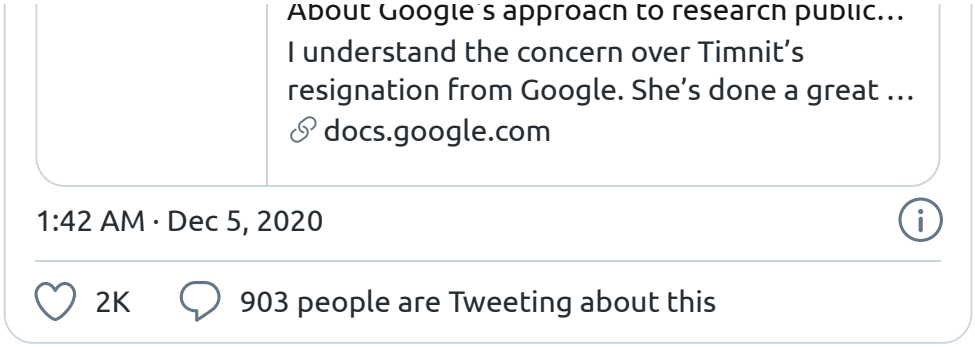
However, the six collaborators drew on a wide breadth of scholarship. The

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."

An AI ethics uproar Dec 4

We read the paper that forced Timnit Gebru out of Google. Here's what it says

The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business.

01.
The two-year fight to stop Amazon from selling face recognition to the police
Jun 12

02.
“We’re in a diversity crisis”: cofounder o Black in AI on what’s poisoning algorith in our lives
Feb 2018
Feb 14

out vaccines

Many countries are making important decisions about who gets vaccinated and when—but the answers vary from nation to nation.

Tech policy 3 days

A leading AI ethics researcher says she's been fired from Google

Timnit Gebru says she's facing retaliation for conducting research that was critical of Google and sending an email "inconsistent with the expectations of a Google manager."

Tech policy 3 days

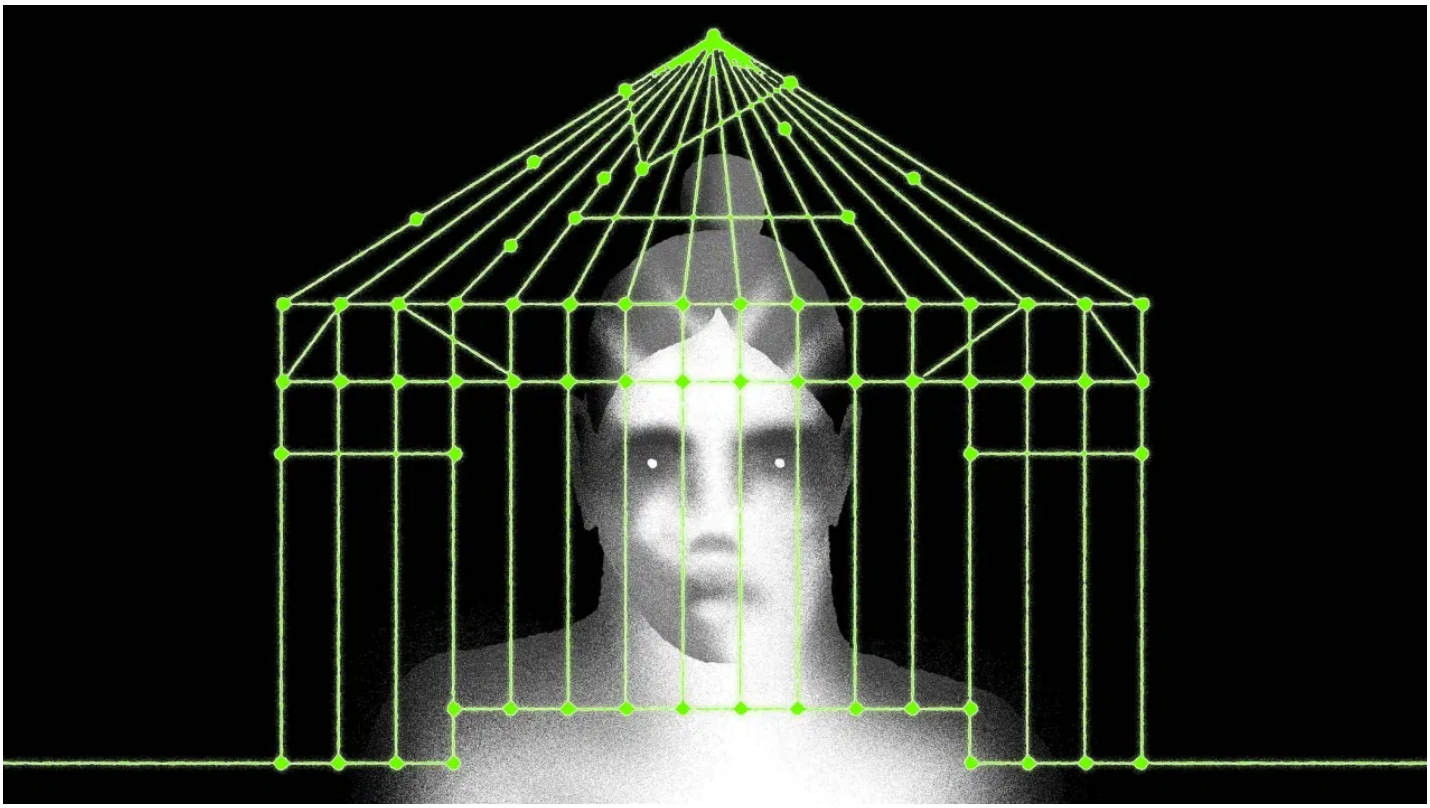
Space 4 days

This is the most precise 3D map of the Milky Way ever made

Through data collected by ESA's Gaia observatory, astronomers have just released a new 3D atlas of the galaxy and its stars moving through space

Tech policy Dec 4

The coming war on the hidden algorithms that trap people in



01.
The UK exam debacle reminds us that algorithms can't fix broken systems
Aug 20

02.
Predictive policing algorithms are racis They need to be dismantled.
Jul 17

asteroid 180 million miles away

Hayabusa2 fired bullets into an asteroid to collect some space dust, and scientists will soon get a chance to study that material in the lab.

Election 2020 Dec 02

The election is over, but voter fraud conspiracies aren't going away

A month after Election Day, the volume of political disinformation has dropped—but experts say the problems are far from over.

How to make the next election even more secure

How election results get certified

Load more