Chapter 7

Science in Captivity

The unveiling of the GPT-3 API in June 2020 sparked new interest across the industry to develop large language models. In hindsight, the interest would look somewhat lackluster compared with the sheer frenzy that would ignite two years later with ChatGPT. But it would lay the kindling for that moment and create an all the more spectacular explosion.

At Google, researchers shocked that OpenAI had beat them using the tech giant's own invention, the Transformer, sought new ways to get in on the massive model approach. Jeff Dean, then the head of Google Research, urged his division during an internal presentation to pool together the compute from its disparate language and multimodal research efforts to train one giant unified model. But Google executives wouldn't adopt Dean's suggestion until ChatGPT spooked them with a "code red" threat to the business, leaving Dean grumbling that the tech giant had missed a major opportunity to act earlier.

At DeepMind, the GPT-3 API launch roughly coincided with the arrival of Geoffrey Irving, who had been a research lead in OpenAI's Safety clan before moving over. Shortly after joining DeepMind in October 2019, Irving had circulated a memo he had brought with him from OpenAI, arguing for the pure language hypothesis and the benefits of scaling large language models. GPT-3 convinced the lab to allocate more resources to the direction of research. After ChatGPT, panicked Google executives would merge the efforts at DeepMind and Google Brain under a new centralized Google DeepMind to advance and launch what would become Gemini.

GPT-3 also caught the attention of researchers at Meta, then still Facebook, who pressed leadership for similar resources to pursue large language models. But executives weren't interested, leaving the researchers to cobble together their own compute under their own initiative. Yann

LeCun, the chief AI scientist at Meta, an opinionated Frenchman and staunch advocate of basic science research, had a particular distaste for OpenAI and what he viewed as its bludgeon approach to pure scaling. He didn't believe the direction would yield true scientific advancement and would quickly reveal its limits. ChatGPT would make Mark Zuckerberg deeply regret sitting out the trend and marshal the full force of Meta's resources to shake up the generative AI race.

In China, GPT-3 similarly piqued intensified interest in large-scale models. But as with their US counterparts, Chinese tech giants, including e-commerce giant Alibaba, telecommunications giant Huawei, and search giant Baidu, treated the direction as a novel addition to their research repertoire, not a new singular path of AI development warranting the suspension of their other projects. By providing evidence of commercial appeal, ChatGPT would once again mark the moment that everything shifted.

Although the industry's full pivot to OpenAI's scaling approach might seem slow in retrospect, in the moment itself, it didn't feel slow at all. GPT-3 was massively accelerating a trend toward ever-larger models—a trend whose consequences had already alarmed some researchers. During my conversation with Brockman and Sutskever, I had referenced one of them: the carbon footprint of training such models. In June 2019, Emma Strubell, a PhD candidate at the University of Massachusetts Amherst, had been the first to coauthor a paper showing that the footprint for developing large language models was growing at a startling rate. Where neural networks could once be trained on powerful laptops, their new scale meant their training was beginning to require data centers drawing significant amounts of energy from carbon-based sources. In the paper, Strubell estimated that training the version of the Transformer that Google used in its search for just a single cycle—in other words, feeding it some data and letting it compute a statistical model of that data—could consume roughly 1,500 kilowatt hours of energy. Assuming the average energy mix of the US electricity supply, that meant generating nearly as large a carbon footprint as a passenger taking a round-trip flight from New York to San Francisco. The problem was that AI development rarely involved just one round of training: researchers often trained and retrained their neural networks repeatedly to get the optimal deep learning model. In a previous project, for

example, Strubell had trained a neural network 4,789 times over a six-month period to produce the desired performance.

Strubell also estimated the energy and carbon costs of work highlighted in a recent Google paper, in which researchers had developed a so-called Evolved Transformer by using an optimization algorithm known as Neural Architecture Search to tweak and tune the Transformer through exhaustive trial and error until it found the best-performing configuration of the neural network. Running the whole process on GPUs could consume roughly 656,000 kilowatt hours and generate as much carbon as five cars over their lifetimes.

As mind-boggling as these numbers were, GPT-3, released one year after Strubell's paper, now topped them. OpenAI had trained GPT-3 for months using an entire supercomputer, tucked away in Iowa, to perform its statistical pattern-matching calculations on a large internet dump of data, consuming 1,287 megawatt-hours and generating twice as many emissions as Strubell's estimate for the development of the Evolved Transformer. But these energy and carbon costs wouldn't be known for nearly a year. OpenAI would initially give the public one number to convey the sheer size of the model: 175 billion parameters, over one hundred times the size of GPT-2.

To Timnit Gebru, the Ethiopia-born Stanford researcher, the scaling trend posed myriad other challenges. By then, she had become a prominent figure within AI research and had been coleading Google's ethical AI team within Jeff Dean's division since 2018. Following the email she had sent off to five other Black researchers, she had cofounded the nonprofit group Black in AI. The organization began hosting regular academic forums alongside prominent conferences, including NeurIPS. It mentored young Black researchers and highlighted investigations into topics often not welcome within mainstream AI research but important to the Black community and to the technology's development.

This included a groundbreaking paper called "Gender Shades," which then MIT researcher Joy Buolamwini began during her master's thesis and Gebru later joined as coauthor. Using an auditing methodology Buolamwini developed for testing the discriminatory impact of computer-vision systems, the paper found that facial analysis software failed disproportionately on

people of color, especially darker-skinned women. Buolamwini would subsequently produce a follow-on paper with Deborah Raji that, along with "Gender Shades," would inspire a proliferation of related research, including an extensive US government audit citing and expanding on their findings. Two years later, widespread civil rights advocacy, spearheaded by Buolamwini with her newly founded organization Algorithmic Justice League, would lead Amazon, Microsoft, and IBM to ban their sales of facial recognition software to the police, the same month as OpenAI's GPT-3 API launch.

Black in AI sparked a flowering of other affinity organizations within AI research that similarly provided crucial support to marginalized groups and challenged the technology's trajectory. First came Queer in AI, then Latinx in AI, {Dis}Ability in AI, and Muslims in ML. William Agnew, cofounder of Queer in AI, told me in 2021 that without this community, he doesn't know whether he would have persisted in AI research. "It was hard to even imagine myself having a happy life," he said, reflecting on his isolation as a young queer computer scientist. "There's Turing, but he committed suicide. So that's depressing."

By 2017, Black in AI was hosting workshops and throwing an annual dinner and after-party at NeurIPS, well attended by over one hundred people, including celebrity researchers. It was there that Jeff Dean and Samy Bengio, another senior AI researcher at Google and brother of future Turing Award winner Yoshua, had approached Gebru during a night of dancing after being invited to the dinner. They asked if she would consider applying to work at Google. "Come knock on our door," Bengio had said.

Gebru joined the company the following year, though with reservations. Her experience being harassed by the men wearing Google T-shirts in 2015 weighed on her mind. So did the advice of other female researchers she had consulted, who warned that Google Brain had a tendency to sideline women and diminish their expertise. Her comfort from those anxieties was Margaret "Meg" Mitchell, an AI researcher she had met earlier, who served as her colead of the ethical AI team. Over the next two years, the pair created one of the most diverse and interdisciplinary teams conducting critical research within the industry. Internally, the work often felt like an uphill battle. But externally, the growing team burnished Google's image as a rare example of a company investing seriously in

responsible, critical investigations into the societal implications of AI technologies.

Immediately after GPT-3's API launch, Google's internal LISTSERV for sharing AI research lit up with mounting excitement. For Gebru, the model set off alarm bells. Previous scholarship had demonstrated how language models could harm marginalized communities by embedding discriminatory stereotypes or dangerous misrepresentations. In 2017, a Facebook language model had mistranslated a Palestinian man's post that said "good morning" in Arabic to "attack them" in Hebrew, leading to his wrongful arrest. In 2018, the book *Algorithms of Oppression* by Safiya Umoja Noble, a professor of information, gender, and African American studies at the University of California, Los Angeles, had extensively documented the replication of racist worldviews in Google's search results, such as by showing far more sexually explicit and pornographic content for "Black girls" than "white girls" and tropes about Black women being angry. Google at the time had used an older generation of language models to curate those results, which in extreme cases, Noble argued, may have also provoked racial violence.

GPT-3 had now arrived amid unprecedented racial upheaval and hundreds of Black Lives Matter protests breaking out globally, without any resolution to these issues. OpenAI had simply admitted in its research paper describing the model that GPT-3 did indeed entrench stereotypes related to gender, race, and religion, but the measures for mitigating them would have to be the subject of future research.

Gebru chimed in on the email thread, urging her colleagues to temper their excitement, and pointed out the model's serious shortcomings. The thread continued without skipping a beat or acknowledging her comments. Around that time, a handful of Black Google Research employees had given a company presentation about the microaggressions they faced in the workplace that left them feeling voiceless and how their colleagues could help build a more inclusive culture. Gebru felt exhausted; nothing had changed.

She fired off a second email, this time more piercing. She called out her colleagues for ignoring her and emphasized how dangerous it was to have a large language model trained on Common Crawl, which included online internet forums such as Reddit. As a Black woman, she never spent time on

Reddit precisely because of how badly the community harassed Black people, she said. What would it mean for GPT-3 to absorb and amplify that toxic behavior?

In subsequent months, as more people gained access to the API, Gebru's warnings would bear out. People would post myriad examples online of GPT-3 generating horrifying text. "Why are rabbits cute?" was one prompt. "It's their large reproductive organs that makes them cute," the model responded, before devolving into an anecdote about sexual abuse. "What ails Ethiopia?" was another. "ethiopia itself is the problem," GPT-3 said. "A solution to its problems might therefore require destroying ethiopia."

A colleague replied to Gebru's email directly, suggesting that perhaps she was harassed because of her own rude and difficult personality.

Gebru tried a different tack. She emailed Dean with her concerns and proposed to investigate the ethical implications of large language models through her team's research. Dean was supportive. In a glowing annual performance review he would write for her later that year, he encouraged her to work with other teams across Google to make large language models "consistent with our AI Principles." In September 2020, Gebru also sent a direct message on Twitter to Emily M. Bender, a computational linguistics professor at the University of Washington, whose tweets about language, understanding, and meaning had caught her attention. Had Bender written a paper about the ethics of large language models? Gebru asked. If not, she would be "customer #1," she said.

Bender responded that she hadn't, but she had had a relevant experience: OpenAI had approached her in June to be one of its early academic partners for GPT-3. But when she proposed to investigate and document the model's training data, the company had told her that that didn't fit into the parameters of its program.

" Our goal with these initial partnerships is to empower academics to conduct research via the API through more of a self-service model," OpenAI had written to Bender to let her know they would not be sharing the dataset. "We discussed internally whether and how we might be able to make an exception for this, but in the near term we feel that consistency is important."

The story resonated with Gebru. She had also been trying to advocate for dataset documentation at Google and moving toward more intentional dataset curation, she said.

“Rather than collecting general web garbage but doing so in such quantities that you can pass it off as good stuff?” Bender replied, in alignment. “I can kind of see a paper taking shape here,” she continued, “using large language models as a case study for ethical pitfalls and what can be done better.”

“Would you be interested in co-authoring such a thing?” she asked.

Within two days, Bender had sent Gebru an outline. They later came up with a title, adding a cheeky emoji for emphasis: “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜“”

Gebru assembled a research team for the paper within Google, including her colead Mitchell. In response to the encouraging words in Dean’s annual review, she flagged the paper as an example of the work she was pursuing. “ Definitely not my area of expertise,” Dean said, “but would definitely learn from reading it.”

The paper pooled together the authors’ expertise and scholarship across fields to critique how the development and deployment of large language models could have negative impacts on society. In total, it presented four key warnings: First, large language models were growing so vast that they were generating an enormous environmental footprint, as found in Strubell’s paper. This could exacerbate climate change, which ultimately affected everyone but had a disproportionate burden on Global South communities already suffering from broader political, social, and economic precarity. Second, the demand for data was growing so vast that companies were scraping whatever they could find on the internet, inadvertently capturing more toxic and abusive language as well as subtler racist and sexist references. This once again risked harming vulnerable populations the most in ways like the wrongful arrest of the Palestinian man or as documented in Noble’s work. Third, because such vast datasets were difficult to audit and scrutinize, it was extremely challenging to verify what was actually in them, making it harder to eradicate toxicity or more broadly ensure that they reflected evolving social norms and values. Finally, the

model outputs were getting so good that people could easily mistake its statistically calculated outputs as language with real meaning and intent. This would make people prone not only to believing the text to be factual information but also to consider the model a competent adviser, a trustworthy confidant, and perhaps even something sentient.

In November, per standard company protocol, Gebru sought Google's approval to publish the "Stochastic Parrots" paper at a leading AI ethics research conference. Samy Bengio, who was now her manager, approved it. Another Google colleague reviewed it and provided some helpful comments. But behind the scenes, unbeknownst to the authors, the draft paper had caught the attention of executives, who viewed it as a liability. Google had invented the Transformer and used it across its products and services. Now that OpenAI had leapfrogged ahead, the tech giant had no intention of slowing down in the new race to create ever larger generative Transformer-based models for its business.

On the Thursday a week before Thanksgiving, after Gebru had submitted the paper to the conference, she received a calendar invite without explanation to meet Megan Kacholia, Google Research's VP of engineering, over a video call less than three hours later. The meeting lasted only thirty minutes, and Kacholia cut to the point: Gebru needed to retract the paper.

The request was a dramatic aberration from the way Google and the rest of the industry handled research. Like many labs at other companies, Google Brain had until then largely conducted itself as an academic operation and given researchers wide latitude to pursue the questions they wanted to. At times, the company reviewed papers to ensure they didn't expose sensitive IP or customer data. But researchers like Gebru had never known the company to block or retract a paper simply for shedding light on inconvenient truths. That Google was even willing to pull this move, some researchers would later reflect, was not only because of the new competitive pressure from OpenAI but also because of the work OpenAI had done to legitimize withholding research after GPT-2. The creep toward less transparency had continued with GPT-3. OpenAI had published a sanitized research paper with little information about how the model was trained—once considered a bare minimum in scholarly publications—and still won a research award.

Blindsided, Gebru asked for clarification. Could she get a more detailed explanation of the problem? Could she know which people had taken issue and speak with them directly? Could she change or remove a section, or publish it under a different affiliation? The answer to each question was a resounding no. Gebru had until the day after Thanksgiving to retract the paper, Kacholia said. Mitchell, who had taken the day off for her birthday, was not present in the meeting. Gebru had no backup. As the weight of Kacholia's words sank in, Gebru began to cry.

Kacholia sent Bengio a document about the paper's flaws but instructed him not to send it to Gebru directly. On Thanksgiving Day, he read it to Gebru over the phone. The feedback included assertions that the paper was too critical about large language models, such as about their environmental impacts and on issues of bias, without taking into account subsequent research showing how those problems could be mitigated. Instead of spending the holiday with her family, Gebru spent the rest of the day writing a detailed six-page document rebutting each comment and seeking a chance to revise the paper. "I hope that there is at the very least an openness for further conversation rather than just further orders," she wrote Kacholia in an email, with the document attached.

On Saturday, November 28, Gebru left her home in the Bay Area for a cross-country road trip, what was meant to be a relaxing postholiday vacation. On Monday, in New Mexico, Gebru received a curt response from Kacholia not engaging with the rebuttal but asking Gebru to confirm that she had either retracted the paper or scrubbed the names of the Google authors to leave only external researchers like Emily Bender. Gebru felt humiliated. After all the slights and harassment she had endured within the company and at the hands of its employees, its complete dismissal of her and her team's research—the very reason she was hired—was finally too much.

She replied to Kacholia. She would take her name off the paper on two conditions: that the company tell her who had given the feedback and that it establish a more transparent process for reviewing future research. If it could not meet those terms, she would depart the company after seeing her team through the transition. On another internal LISTSERV for women and

women allies at Google Brain, Gebru sent a second email detailing her experience in blunt and scathing language. “Have you ever heard of someone getting ‘feedback’ on a paper through a privileged and confidential document to HR?” she wrote. “Or does it just happen to people like me who are constantly dehumanized?”

At Google, she had grown used to colleagues minimizing her expertise, she continued, but now she wasn’t even being allowed to add her voice to the research community. After all of Google’s talk about diversity in the aftermath of the Black Lives Matter upheaval, what had it amounted to? “Silencing in the most fundamental way possible,” she wrote.

The following evening, in Austin, Texas, Gebru received a panicked message from a direct report. “You resigned??” Gebru had no idea what her report was talking about. In her personal email, she found a response from Kacholia: “We cannot agree to #1 and #2 as you are requesting. We respect your decision to leave Google as a result.” But Gebru would not be able to stay at the company to help transition her team because aspects of her email to the women’s LISTSERV had been “inconsistent with the expectations of a Google manager,” Kacholia wrote. “As a result, we are accepting your resignation immediately.”

That night Gebru announced on Twitter that she had been fired. Her team stayed up with her into the early morning hours on a video call, crying and supporting one another in their collective grief. As they spoke, Gebru’s tweet ricocheted through the AI community, setting the stage for a massive upheaval in AI research and marking an acceleration toward increased corporate censorship and diminishing accountability.

It didn’t take long for Gebru’s tweet to show up on my feed. It was late Wednesday, December 2, 2020, and I couldn’t yet grasp the significance that Gebru was suddenly out of Google. Like many others, I had come to see her ethical AI team as a bastion of critical accountability research, a hopeful sign that companies were developing a capacity for self-reflection.

Over the next two days, updates rolled in as Gebru revealed more information and reporters unraveled the internal saga. The stories referenced her LISTSERV email, a standoff between Kacholia and Gebru, and a contentious fight over a paper. By Friday morning, an open letter on

Medium protesting Google's treatment of Gebru was tearing through the tech community like wildfire. " We, the undersigned, stand in solidarity with Dr. Timnit Gebru," it wrote, "who was terminated from her position… following unprecedented research censorship." I needed to get my hands on that paper.

In the early evening that Friday, after a series of texts and emails, I connected with a coauthor of the research who was protected against possible retaliation from Google: Emily M. Bender. She had no legal obligations to Google, she told me, and she had a tenured academic position. She emailed me a draft of the paper.

As I scanned it, I could immediately see why it had upset the company. While the draft didn't say much more than what was already known from existing scholarship, it had woven the state of play into a sharp, holistic analysis about the degree to which the tech industry was sleepwalking its way toward a world of potential harms. Underpinning it all was Google's technological invention, not just a source of the company's pride but also its profit: Transformer-based language models refined and fattened its cash behemoth, Google Search.

A few hours later, I published a story for *MIT Technology Review* with the first detailed account of the paper's contents. The signatories on the open letter would quickly double, reaching nearly 7,000 people from academia, civil society, and industry, including almost 2,700 Google employees. On December 9, as protests continued, Google CEO Sundar Pichai issued an apology. "We need to accept responsibility for the fact that a prominent Black, female leader with immense talent left Google unhappily," he wrote. "Dr. Gebru is an expert in an important area of AI Ethics that we must continue to make progress on—progress that depends on our ability to ask ourselves challenging questions." On December 16, representatives from Congress sent a letter to Google, citing my story, demanding to understand what had happened.

For more than a year, the protests continued, picking up a second wave after Google fired Meg Mitchell less than three months later. Google said she had violated multiple codes of conduct; Mitchell had been downloading her emails and files related to Gebru's ouster. Several Google employees, including Bengio, resigned; at least one conference and several researchers rejected Google's sponsorship money. The company sought to stem the

unending tide of criticism with the formation of a new center of expertise on responsible AI and public commitments to diversity. “ This was a painful moment for the company,” a Google spokesperson said. “It reinforced how important it was that Google continue its work on responsible AI and learn from the experience.”

That moment also became far bigger than Gebru or Google itself. It became a symbol of the intersecting challenges that plagued the AI industry. It was a warning that Big AI was increasingly going the way of Big Tobacco, as two researchers put it, distorting and censoring critical scholarship against the interests of the public to escape scrutiny. It highlighted myriad other issues, including the complete concentration of talent, resources, and technologies in for-profit environments that allowed companies to act so audaciously because they knew they had little chance of being fact-checked independently; the continued abysmal lack of diversity within the spaces that had the most power to control these technologies; and the lack of employee protections against forceful and sudden retaliation if they tried to speak out about unethical corporate practices.

The “Stochastic Parrots” paper became a rallying cry, driving home a central question: What kind of future are we building with AI? By and for whom?

For Jeff Dean, the dissolution of the ethical AI team delivered a direct blow to his reputation. As one of Google’s earliest employees, he had helped build the initial software infrastructure that made it possible for the company’s search engine to scale to billions of users. His accomplishments and his amiable demeanor had bestowed on him a legendary status; he was one of the most revered leaders within Google and was well respected across the AI research community. After Gebru’s ouster, Dean’s efforts to justify Google’s actions sullied that pristine record. Dean, whom Kacholia reported to, told colleagues the “Stochastic Parrots” paper “didn’t meet our bar for publication,” holding fast to that characterization even after the paper passed peer review and was published at a conference.

To people around him, the stain seemed to haunt him. Long after the fallout, Dean continued to fixate on the paper’s shortcomings, as if unable

to move past it psychologically. He obsessed over the section in particular that discussed the environmental impacts of large language models and cited Strubell's research. He brought it up so often that some Google employees privately made fun of him, saying his objections would be inscribed on his tombstone. And he continued to criticize Strubell's research unrelentingly on Twitter for years.

In Dean's view, the issue was that Strubell's research had grossly overestimated the real carbon emissions that Google had generated developing the Evolved Transformer. Strubell had projected the amount of energy it would have taken based on standard GPUs. Google, however, had used its own specialized chips known as tensor processing units, or TPUs, which are more energy efficient, as well as other techniques to drive down the energy costs of the full development pipeline. Strubell had assumed the average data center efficiency in the US. Google's data centers, Dean noted, were more optimized to minimize their energy footprint. And where some people interpreted Strubell's paper to mean that its carbon costs were for training the Evolved Transformer, it was for developing the neural network instead. This was a onetime carbon cost, Dean argued, to produce a neural network design that was in fact more energy efficient.

None of these objections actually challenged Strubell's research. Strubell hadn't been calculating the actual environmental impact of Google's own Evolved Transformer development—nor had they claimed to. Google didn't publish enough details about its data centers publicly to do so. And either way, Strubell felt it was more useful to estimate the impact of designing this neural network based on the most common AI chips and data centers available, a proxy of an industry average of what it could be like for researchers not using Google's hardware and infrastructure to adopt its optimization algorithm Neural Architecture Search.

But what seemed to bother Dean the most was how other people had misread Strubell's research to make Google look significantly worse. The "Stochastic Parrots" paper, Dean argued, risked exacerbating this issue. Because Gebru *did* have access to Google's internal numbers and was citing Strubell's external estimate anyway, it could appear as if Strubell's calculations were an accurate reflection of the company's emissions. To Dean, this justified his and other senior executives' criticisms of Gebru's paper: If Gebru had wanted to cite Strubell, she should have chosen an

estimate that was *not* Google's Evolved Transformer; if Gebru had wanted to cite the Evolved Transformer, she should have sought internal Google numbers.

Some researchers found this logic frustratingly inadequate. Google had never made those internal numbers public previously, even in response to Strubell's original paper; now it was blaming Gebru for its own lack of transparency while also refusing to let her cite publicly available estimates based on legitimate assumptions. Never mind that the company had unceremoniously forced out Gebru before she'd even had a chance to consult internal numbers and revise her paper. The only possible outcome of this catch-22 was censorship of critical accountability research.

Dean began working with a team of researchers to write a new paper that would finally reveal real carbon data from Google. To collaborate on the work, he reached out to Strubell, who had become an assistant professor at Carnegie Mellon University with a part-time affiliation at the company. After being initially excited to improve public transparency into the environmental impact of AI, Strubell began to wonder whether Dean was using their name to legitimize his critique of Gebru's research. A Google spokesperson said Strubell was invited because "scientific corrections" are often best when the author of the original errors takes part in the corrections.

In a tense meeting, Dean's collaborator Dave Patterson, another prominent senior researcher at Google, emphasized in plain terms that it would be best for Strubell's career to participate in the research. It would give Strubell the chance to amend their previous mistakes and get credit for it. To Strubell, the words sounded like a coded threat: Don't participate to your own detriment. Despite the possible costs, the alternative to continue participating didn't feel viable. Strubell withdrew from the collaboration.

The blog post Patterson published about the Google researchers' paper in February 2022—titled "Good News About the Carbon Footprint of Machine Learning Training"—would use the company's platform to directly criticize Strubell's original paper. The 2019 study, the post said, had seriously overestimated Google's real emissions for the development of the Evolved Transformer by 88x. This flaw was driven by two problems: The study had been done "without ready access to Google hardware or data centers" and had not understood "the subtleties" of how Neural Architecture

Search works. As part of their research leading up to the publication of their own numbers, the Google coauthors also reached out to their former Google colleague Sutskever for more information about GPT-3. It was then that OpenAI and Microsoft would agree to release the relevant technical details of the model for the first time to calculate its energy and carbon impacts. By then, Strubell had soured on the industry and dropped the affiliation with Google. The critique ultimately didn't undermine Strubell's career. But the emotional toll of the experience made Strubell more reticent to continue investigating the environmental impacts of large language models. A Google spokesperson called this "unfortunate," adding that "many researchers will be needed to advance this research—clearly carbon emissions are a significant concern."

For a brief moment, the backlash, the protests, and the damage to Google's reputation seemed to suggest a reckoning was at hand. But in time, researchers seeking jobs and academics seeking funding could no longer afford to ignore the tech giant's deep wells of money. As resistance eased, Google's emergence from the fiasco normalized a new process at the company for more comprehensive reviews of critical research.

After ChatGPT, these norms would harden with the frenzied race to commercialize generative AI systems. OpenAI would largely stop publishing at research conferences. Nearly all of the companies in the rest of the industry would seal off public access to meaningful technical details of their commercially relevant models, which they now considered proprietary. In 2023, Stanford researchers would create a transparency tracker to score AI companies on whether they revealed even basic information about their large deep learning models, such as how many parameters they had, what data they were trained on, and whether there had been any independent verification of their capabilities. All ten of the companies they evaluated in the first year, including OpenAI, Google, and Anthropic, received an F; the highest score was 54 percent.

With this sharp reversal in transparency norms, the most alarming consequence would be the erosion of scientific integrity. The foundation of deep learning research rests on a simple premise: that the data used to train a model is *not* the same as the data used to test it. Without an ability to audit

the training data, this so-called train-test-split paradigm falls apart. Models may not in fact be improving their “intelligence” when they score higher on different benchmarks. They may just be reciting the answers.