Empire of AI

Dreams and Nightmares in Sam Altman's OpenAI

Karen Hao

Empire of AI

Dreams and Nightmares in Sam Altman's OpenAI

KAREN HAO

PENGUIN PRESS NEW YORK 2025

PENGUIN PRESS
An imprint of Penguin Random House LLC
1745 Broadway, New York, NY 10019
penguinrandomhouse.com

Hardcover ISBN 9780593657508
International edition ISBN 9798217060481

Ebook ISBN 9780593657515

Cover design: Chris Allen

Book design by Daniel Lagin, adapted for ebook by Cora Wigen The authorized representative in the EU for product safety and compliance is Penguin Random House Ireland, Morrison Chambers, 32 Nassau Street, Dublin D02 YH68, Ireland, https://eu-contact.penguin.ie.

pid_prh_7.1a_151466483_c0_r0

CONTENTS

To my family,
past, present, and future.

To the movements
around the world
who refuse dispossession
in the name of abundance.

It is said that to explain is to explain away. This maxim is nowhere so well fulfilled as in the area of computer programming, especially in what is called heuristic programming and artificial intelligence. For in those realms machines are made to behave in wondrous ways, often sufficient to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away; it stands revealed as a mere collection of procedures, each quite comprehensible. The observer says to himself “I could have written that.” With that thought he moves the program in question from the shelf marked “intelligent,” to that reserved for curios, fit to be discussed only with people less enlightened than he.

—JOSEPH WEIZENBAUM, MIT PROFESSOR AND INVENTOR OF THE FIRST CHATBOT, ELIZA, 1966

“ Successful people create companies. More successful people create countries. The most successful people create religions.”

I heard this from Qi Lu; I’m not sure what the source is. It got me thinking, though--the most successful founders do not set out to create companies. They are on a mission to create something closer to a religion, and at some point it turns out that forming a company is the easiest way to do so.

—SAM ALTMAN, 2013

AUTHOR'S NOTE

This book is based on over 300 interviews with around 260 people and an extensive trove of correspondence and documents. Most of the interviews were conducted for this book. Some were drawn from my last seven years of reporting on OpenAI, the AI industry, and its global impacts for *MIT Technology Review*, *The Wall Street Journal*, and *The Atlantic*. Over 150 of the interviews were with more than 90 current or former OpenAI executives and employees, and a handful of contractors who had access to detailed documentation of parts of OpenAI's model development practices. Another share of the interviews was with some 40 current and former executives and employees at Microsoft, Anthropic, Meta, Google, DeepMind, and Scale, as well as people close to Sam Altman.

Any quoted emails, documents, or Slack messages come from copies or screenshots of those documents and correspondences or are exactly as they appear in lawsuits. In cases where I do not have a copy, I paraphrase the text without quotes. There is one exception, which I mark in the endnotes. All dialogue is reconstructed from people's memories, from contemporaneous notes, or, when marked in the endnotes, pulled from an audio recording or transcript. In most cases, I or my fact-checking team asked those recalling quotes to repeat or confirm them again several months apart to test their stability. Every scene, every number, every name and code name, and every technical detail about OpenAI's models, such as the composition of their training data or the number of chips they were trained on, is corroborated by at least two people, with contemporaneous notes and documentation, or, in a few cases that I mark in the endnotes, with other media reporting. The same is true for most every other detail about OpenAI in the book. If I named someone, it does not mean I spoke to them directly. When I reference anyone's thoughts or feelings, it is because they described

that thought or feeling, either to me, to someone I spoke to, in an email or recording I obtained, or in a public interview.

This book is not a corporate book. While it tells the inside story of OpenAI, that story is meant to be a prism through which to see far beyond this one company. It is a profile of a scientific ambition turned into an aggressive ideological, money-fueled quest; an examination of its multifaceted and expansive footprint; a meditation on power. To that end, in the course of my reporting, I spent significant time embedding with communities on the ground in countries around the world to understand their histories, cultures, lives, and experiences grappling with the visceral impacts of AI. My hope is that their stories shine through in these pages as much as the stories within the walls of one of Silicon Valley's most secretive organizations.

I reached out to all of the key figures and companies that are described in this book to seek interviews and comment. OpenAI and Sam Altman chose not to cooperate.

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.

Chapter 9

Disaster Capitalism

As OpenAI barreled forward, guided by Altman's convictions, the boundaries of the sweeping consequences of the company's vision were expanding. With its pumping of ever-larger and polluted datasets into its models, it had created the "paradigm shift" that Appen's Ryan Kolln would describe to me—the moving away from filtering data inputs to controlling model outputs. The language of abstraction once again dressed up a grim reality: what that shift really meant for the people who now bore the brunt of controlling those outputs.

In 2021, in parallel with its push to develop the next generations of its models, OpenAI began a project to create a much better version of its automated content-moderation filter for cleaning them up. Where GPT-3 had been placed on the API with no filtering whatsoever, leading to the Latitude text-based child porn scandal, the company wanted to be more careful with the models it would start calling GPT-3.5 and eventually GPT-4. As OpenAI prepared to deploy its technologies more widely, having a completely unfiltered product could prove problematic in the long run from a legal, public relations, and usability perspective. At the time, the plans for what would become ChatGPT had yet to be conceived, but the chatbot would also later benefit from the same filter. That filter would act as a wrapper around each model, for the purpose of flagging and removing offensive content from its output before it reached the user.

To build the automated filter, OpenAI first needed human workers who could carefully review and catalog hundreds of thousands of examples of exactly the content—sex, violence, and abuse—that the company wanted to prevent its models from generating. After six months of searching, it found a vendor that seemed well suited to take on the project: an outsourcing firm that had been performing content moderation for Meta since 2019 and

coincidentally shared Altman's nickname, Sama. OpenAI sent Sama an email asking whether it took on projects that involved sensitive or explicit content and what its typical approach was for handling them. Sama provided thorough answers. OpenAI signed four contracts with the firm for $230,000, landing the project in the hands of dozens of workers in Kenya.

It's no coincidence that Kenya became home to what would ultimately turn into one of the most exploitative forms of labor that went into the creation of ChatGPT. Kenya was among the top destinations that Silicon Valley had been outsourcing its dirtiest work to for years. With the many other countries that the tech industry relegates to this role, Kenya shares a common denominator: It is poor, in the Global South, with a government hungry for foreign investment from richer countries. All of these are a part of Kenya's legacy of colonialism, which has left it without well-developed institutions to protect its citizens from exploitation and often in the throes of economic crisis, both of which make circumstances ripe for overseas companies to find an immiserated pool of labor that will do piecework under almost any conditions.

You can see the markers of that legacy in the many faces of Nairobi, Kenya's capital. The city suffers grave inequality. The central business district has gleaming towers, international five-star hotels, and high-end restaurants. The diplomatic neighborhood has large, stately buildings and high security walls. The residential expat areas offer stunning mansions with lush private gardens. And then there are the outskirts: Utawala, Dagoretti South, Embakasi. Drive to any one of these neighborhoods, and skyscrapers made of steel turn into squat cinder block structures. Buildings begin to scatter about like weeds in erratic patterns without coordinated planning. The roads go from paved to unpaved, from four lanes to narrow strips meant primarily for motorbikes and pedestrians. Deeper in, concrete turns into corrugated tin, and jerry-built homes and businesses cram together ever more tightly.

Under these conditions, Kenya's government had willingly embraced Silicon Valley when it came in search of low-wage workers. Kenya has limited local industry. Many of the biggest brands are European and American. Some of the largest infrastructure, once built by the British, is now built by the Chinese. Cars are mostly hand-me-downs from Japan, where drivers also sit on the right side of the vehicle. Tech giants, as the

government saw it, could help the country create the jobs it desperately needed. Joblessness breeds crime. Petty theft is common. People who feel disempowered grow distrustful of institutions. During the Russia-Ukraine war, as Kenya's grain prices rose, rumors spread that the president was purposely straining already hungry families. Many repeated a familiar refrain in the US: The election was rigged.

And so, Kenya became a critical hub of the internet's backstop labor. Several firms like Sama—middlemen in the data labor supply chain—established operations in Nairobi, building up pools of workers to service overseas tech companies, primarily in the Bay Area.

For OpenAI, Sama appeared to check off all the right boxes. Originally called Samasource, it was a San Francisco–based social enterprise that had begun in 2008 with a mission of providing meaningful, dignified work to people in impoverished countries to lift them out of poverty. Under its founder, Leila Janah, it had established operations in India and Kenya and developed a reputation as an ethical outsourcing company. In 2018, it transitioned to a for-profit, during which it shortened its name, in order to scale its operations. In 2020, it received a B Corp certification. In its answers to OpenAI's questions in 2021, the organization detailed its experience with content moderation and emphasized its protocols for keeping projects secret and their data secure. It mentioned that it provided mental health resources to its workers to help them deal with psychologically troubling content.

Behind the scenes, however, Sama was in disarray. In January 2020, Janah had passed away from a rare cancer at just thirty-seven; combined with the pandemic soon after, workers say it seemed to mark the beginning of more organizational mismanagement, a characterization that a Sama spokesperson denied. It wasn't until early 2022 that those challenges would come to the fore when Billy Perrigo, a reporter at *Time* magazine, would publish an extensive investigation. He would reveal that Sama had taken on a project for Meta, to provide content moderation for Facebook for all of sub-Saharan Africa, that repeatedly exposed workers to violent and graphic videos, such as of suicides and beheadings, and left them deeply scarred and struggling. Sama would defend itself, saying it took on the project after careful consideration from its East Africa team, which wanted to ensure "content for Africans was effectively reviewed by Africans." Nearly two

hundred workers would file multiple lawsuits against Sama and Meta alleging traumatic working conditions and unlawful terminations for attempting to organize for higher pay and better working conditions. The Sama spokesperson rejected the allegations.

Against this backdrop, OpenAI began the first phase of its project in late 2021. Under the code names PBJ1, PBJ2, PBJ3, and PBJ4, it thrust teams of Sama workers into more traumatic content-moderation work, for on average between $1.46 and $3.74 an hour. Workers had no idea for whom or why they were doing the project, kept in the dark under the nondisclosure terms of the contract, common in the data annotation industry. What they did know was what was in front of them: the hundreds of thousands of grotesque text-based descriptions that they needed to read and sort into categories of severity. Was it violence or extremely graphic violence, harassment or hate speech, child sexual abuse or bestiality?

Gradually, the work broke many of the workers, the impacts radiating beyond each individual to the people who depended on them in their communities. Only after the release of ChatGPT would they begin to grasp what exactly they had paid for with their peace of mind. In May 2023, I visited four workers in Nairobi who would agree to share their experiences with me on the record for a story on the front page of *The Wall Street Journal*. For one of them on the sexual content team, a man named Mophat Okinyi, the project that unraveled his mind and his relationships would turn out to be in service of a technology that would in turn contribute to the erosion of his brother’s economic opportunities.

At a time when freewheeling corporate research was still permitted, Microsoft anthropologist Mary L. Gray and computational social scientist Siddharth Suri were among the first to show the world the plight of workers like those in Kenya who build their livelihoods around an essential piece of the AI supply chain.

In 2019, they published their book *Ghost Work*, based on five years of extensive fieldwork, revealing a hidden web of piecemeal labor and digital exploitation that propped up Silicon Valley. Tech giants and unicorns were building their extravagant valuations not just with engineers paid six-figure

salaries in trendy offices. Essential, too, were workers, often in the Global South, being paid pennies to carefully annotate reams of data.

Take self-driving cars. A self-driving car needs to drive in the correct lane, respond to erratic driving behavior, and pause at a safe distance for schoolchildren crossing the road. To do this, the software system controlling the car uses an amalgamation of several deep learning models, including those dedicated to recognizing objects on the road: lane markings, road signs, traffic lights, vehicles, trees, pedestrians.

Companies develop those models by driving vehicles around with numerous large cameras, recording billions of miles of footage. The footage is the data, and to annotate it means tracing, frame by frame, each object that appears—sometimes down to the curvature of a hand gripping a bike handle or a dog lounging halfway out of a car window—and assigning them labels like “bike,” “vehicle,” “animal,” “human.” People have to do that work. And from a company’s perspective, the cheaper they do it, the better.

Gray and Suri’s research focused in part on Mechanical Turk, a platform developed by Amazon many years before the deep learning boom in 2012, which for a long time served as the de facto middleman for companies looking to hire someone cheap for any kind of piecemeal digital labor. By the time *Ghost Work* came out, the first era of AI commercialization was already evolving, building upon, and rapidly expanding this outsourcing model.

I began mapping out the new shape of this hidden workforce, unearthing a sprawling global pipeline of labor spanning many countries, both expected and surprising. I spoke with the newest middlemen replacing Mechanical Turk—platforms designed to cater more specifically to AI development. I spoke with dozens of workers, visiting some of their homes, eating dinner with their families, seeking to understand not just the macro trends pressing down on them but the daily textures of their lived realities.

Just as the first era of AI commercialization laid the groundwork for the generative AI era’s amassing of data and capitalization of compute, so, too, did it create the foundations for its wide-scale labor exploitation. In this way, it is important to first understand those foundations in order to understand the experience of the Kenyan workers who contracted for OpenAI. Only then is it possible to recognize that their experiences were far from anomalous but rather a direct consequence of the compounding of the

AI industry's long-standing treatment of its hidden workers and its views on whose labor is or isn't valued, with OpenAI's empire-esque vision for unprecedented scale.

Before generative AI, self-driving cars were the biggest source of growth for the data-annotation industry. Old-school German auto giants like Volkswagen and BMW, feeling threatened by the Teslas and Ubers of the world, spun up new autonomous-vehicle divisions to defend their ground against the fresh-faced competition. As billions of new dollars flooded into the race to create the cars of the future, the demand for data annotation exploded and created a need for alternatives to Mechanical Turk.

MTurk, as it was called, was a generalist platform, meaning it didn't cater to any particular kind of work. It was just a self-service website. Its interface—stuck in the web design of the mid-aughts, when it launched—had a place to upload datasets, to specify simple annotation instructions, and to set a price for the work. Once the task was claimed, it showed randomized strings of numbers and letters in place of the workers' names. It had two buttons next to each worker: one to give them a bonus, the other to boot them off the project.

Data annotation for self-driving cars necessitated a different approach. First and foremost, it required a new level of accuracy. One too many mislabeled frames—vehicles traced with sloppy borders, pedestrians not traced at all—could be the difference between life and death. To guarantee that quality, workers needed to be trained and companies needed to write detailed instructions. There needed to be more mechanisms for feedback and iteration. MTurk fell out of favor. In stepped a wave of startups and incumbents including Scale AI, Hive, Mighty AI, and Appen. Each had their own worker-facing platforms, which allowed anyone to create an account and start tasking.

But right as this new wave of companies sought to establish themselves, a strange thing happened. Sign-ups on their worker-facing platforms came rushing in from an unexpected country: Venezuela. In the same moment that auto giants began scrambling, money began pumping into self-driving cars, and data-annotation firms began looking for more

workers, Venezuela was nose-diving headfirst into the worst peacetime economic crisis globally in fifty years.

Economists say it was a toxic cocktail of political corruption and the government's misguided policies that squandered the country's rich natural endowment. Venezuela sits atop the largest proven petroleum reserves in the world. It was once Latin America's wealthiest country. But beginning in 2016, hyperinflation went haywire; unemployment skyrocketed; violent crime exploded as families across the country watched the value of their entire life savings collapse. From late 2017 to 2019, escalating sanctions imposed by the Trump administration, intended as a punishment for Venezuelan leader Nicolás Maduro's authoritarian abuses, delivered the final death knell to Venezuela's economy. Hyperinflation hit a once unfathomable 10 million percent. People with graduate degrees and previously well-paying jobs were now spending their days lining up in front of stores for a chance at receiving meager rations of rice and flour.

Amid the catastrophe, many Venezuelans turned to online platforms for work. By mid-2018, hundreds of thousands had discovered and joined the data-annotation industry, accounting for as much as 75 percent of the workforce for some outsourcing firms. Working on data-annotation platforms became a whole-family activity. Julian Posada, an assistant professor at Yale University who interviewed dozens of Venezuelan workers, found that parents and children often took turns to work on a shared computer; wives reverted to cooking and cleaning to allow their husbands to earn just a little more money by pulling longer hours uninterrupted. The crisis left an indelible mark on the wave of AI-specialized outsourcing firms as they grew up alongside it. Venezuela was not an obvious choice for finding pools of labor. The language barrier made it more difficult for the mostly San Francisco– and Seattle-based firms to coordinate with workers. But the acute desperation among Venezuelans meant they were willing to work for astonishingly small amounts of money, which in turn meant the firms could offer astonishingly good prices for their services. "It was like a freak coincidence," Florian Alexander Schmidt, a professor at the University of Applied Sciences Dresden who has studied the rise of the data-annotation industry, told me in 2022.

That "freak coincidence" revealed a disturbing formula. When faced with economic collapse, Venezuela suddenly checked off the perfect mix of

conditions for which to find an inexhaustible supply of cheap labor: Its population had a high level of education, good internet connectivity, and, now, a zealous desire to work for whatever wages. It was not the only country that fit that description. More populations were getting wired to better internet. And with accelerating climate change and growing geopolitical instability, it was hard to bet against more populations plunging into crisis. “It’s quite likely there will be another Venezuela,” Schmidt said.

At the time, Schmidt’s prediction made me wonder whether the second time around would still be coincidence or whether data-annotation firms would make a playbook out of what had worked there. Scouting workers in crisis could become a surefire way to continue driving down the costs of the labor that serves as the lifeblood of the AI industry. Looking back several years later, that’s exactly what happened—and what has become one of the most stunning parallels between empires of old and empires of AI. One of the defining features that drives an empire’s rapid accumulation of wealth is its ability to pay very little or nothing at all to reap the economic benefits of a broad base of human labor.

In December 2021, I journeyed through the winding mountains of Colombia to better understand the life of a worker who, in crisis, had turned to data annotation. Travel restrictions barred me from going to Venezuela, but here in its neighboring country lived nearly two million Venezuelan refugees, one-third of the population that had been displaced by the economic catastrophe.

Julian Posada connected me with one of them, a woman named Oskarina Veronica Fuentes Anaya, who continued to work in the data-annotation industry after she escaped her home country. Fuentes was the first person to show me what this work is really like—the way she’d reoriented her entire life around working for a platform; the way that platform in turn treated her as disposable.

In the apartment she shared with half a dozen relatives, we sat side by side in the living room as she clicked through screen after screen on Appen. The tasks were varied. They ranged from categorizing products on e-commerce sites—*Should this item be listed under clothing or accessories?*—to performing content moderation for social media—*Does this video*

contain crime or human rights violations? For tasks that required English, she used Google Translate to convert the text into her native Spanish.

Each time she completed a task, the sum of money she earned, displayed in US dollars, would increase by a few pennies. She needed a minimum of ten dollars to withdraw it, which, when she first joined the platform, wasn't a problem. Now, it could take weeks to accumulate that much money. That minimum sometimes felt like a cruel arbiter of whether she had enough funds to pay for groceries.

For workers actually living in Venezuela, the process of withdrawal was even more challenging. Most global payment systems such as PayPal didn't allow money transfers into Venezuela. Most stores and shops in Venezuela didn't accept payments from the ones that do. This meant workers needed to convert their digital funds into cash to pay for basic goods and services. But where the money arrived online in US dollars, the cash needed to be in Venezuelan bolivares. The black market to convert one to the other abounded with scams and high commissions.

Fuentes had a complicated relationship with the platform. It had never been her intention to work this kind of job, but through a series of events outside her control, it had become her lifeline as well as a punishing force.

She had created an Appen account in grad school to earn some extra money while finishing up a master's in engineering. She was sharp, hardworking, and creative. Had her country not crumbled, a top student like her would likely have had guaranteed job security working for the state oil company. When her country did, she adapted, carefully orchestrating her and her husband's departure to Colombia for a chance at a better future.

In that regard, Fuentes was one of the lucky ones. By birth she was entitled to a Colombian passport, unlike many others who escaped without documentation. Her parents were Colombian before they'd fled a generation earlier in the opposite direction, to Venezuela, to escape a different nexus of violence and political instability. It was an all-too-common story—the compounding of generations of crisis across borders, thrusting families into an endless state of siege and survival.

With that passport, Fuentes arranged remotely from Venezuela to rent an apartment in Colombia from an acquaintance. They needed two people who owned property to cosign the lease. The acquaintance, their prospective landlord, agreed to help procure them.

In early 2019, with only enough money for a week of groceries to their names, Fuentes and her husband crossed the border. But upon arrival, they discovered another Venezuelan couple already living in the apartment their landlord had promised them. With no other choice, both couples shared the same roof, each filled with fear and distrust that they would lose their home to the other.

The other couple eventually left, but it was only the beginning of a new string of problems. While Fuentes had found a job at a local call center, her husband didn't have work authorization. Before he could secure one, the call center announced that it would be closing. So when Fuentes began to experience signs of a serious health problem, she ignored the symptoms and continued working. All she could think about was putting in as many hours as possible in the final stretch of her employer's operation.

The doctor later told her that had she waited any longer, she likely would have died. A coworker, alarmed by the markers of Fuentes's deteriorating health, had brought her to the hospital shortly before her body started convulsing and her pulse stopped for a full minute.

She was diagnosed with severe diabetes and immediately placed on five daily courses of insulin. For weeks she experienced crippling pain and bouts of blindness. When she restabilized, she continued to suffer intense fatigue and couldn't leave home for more than a couple hours.

Even then, all she could think of was that she and her husband needed money. But with a chronic illness, she could no longer safely commute the distances she needed to return to an office. It was then that she pulled out her laptop and logged back in to Appen.

To Fuentes, there was little apparent logic to which tasks appeared in her Appen queue. The only thing it made clear was that she needed to have good, consistent performance to continue receiving work. Wilson Pang, Appen's CTO then, told me in 2021 that the platform used algorithms to distribute projects based on a mix of factors including the workers' location, their overall accuracy and speed, and the types of tasks at which they'd previously excelled.

In Telegram and Discord groups, Fuentes traded tips with other Venezuelans working on Appen as they sought to deduce the rules like an

elaborate puzzle. They discovered that using a VPN to appear to be in the US earned them the most money. They also learned—the hard way—that it was a high-risk endeavor. Appen searched for this kind of behavior, which was a violation of platform rules, and punished workers by closing their accounts. An account closure could be devastating. Any earnings a worker hadn't withdrawn would vanish, and opening up a fresh account meant starting back at the bottom, with the least-well-paid tasks or, increasingly, no tasks at all.

There were other rules. Submitting a task quickly was rewarded, but submitting a task too quickly triggered something in the system that meant a worker wouldn't get paid for that task. The prevailing theory was that the platform associated exceptional speed with bot activity, which meant it discarded the answers. Sometimes the tasks that appeared also had few instructions and were impossible to decipher; other times the platform had bugs that didn't load the tasks correctly.

The Venezuelans in the group who were once software engineers created browser extensions to deal with these issues and shared them with their fellow Appen workers. One extension added an extra time delay to every task submission to avoid the apparent bot tax. Another automatically refreshed the Appen queue every second because the platform didn't always update itself. A third sounded an alarm once a new task appeared so workers could step away from their computers to go to the bathroom or cook without fear of missing an opportunity.

For all that the workers did to help each other, the platform pitted them in competition. Projects were first come, first served. A task stuck around in queue only as long as it took for enough workers to claim it. This window—between a task's arrival and its disappearance—shrank over time from days to hours to seconds as more and more workers, including many Venezuelans in crisis, joined Appen and vied for scraps of work.

The erratic, unpredictable nature of when work came and went began to control Fuentes's life. Once she was taking a walk when a task arrived that would have earned her several hundred dollars, enough money to live on for a month. She sprinted as fast as possible back to her apartment but lost the task to other workers. From that day on, she stopped leaving the house on weekdays, allowing herself only thirty-minute outings on weekends, which she learned from experience was when tasks were less likely to show up.

She slept fitfully, worried about the tasks that would arrive in the middle of the night. Before bed, she would turn her computer to maximum volume so that if they did, the browser extension that rang the alarm would wake her up.

Yet despite how much stress and hairpulling Appen caused, Fuentes couldn't imagine leaving the platform. She was terrified that tasks would stop arriving altogether and she would be forced to move on. Appen had been her savior, the only thing that pulled her through when everything else in her life had threatened to end her. Not only that, the earnings were once so great, she was able to invest in a new laptop and recoup the cost and then some.

When things were good, they were really good. When things were bad, she stayed tethered to the platform with the stubborn faith that it would return her loyalty.

Fuentes taught me two truths that I would see reflected again and again among other workers, who would similarly come to this work amid economic devastation. The first was that even if she wanted to abandon the platform, there was little chance she could. Her story—as a refugee, as a child of intergenerational instability, as someone suffering chronic illness—was tragically ordinary among these workers. Poverty doesn't just manifest as a lack of money or material wealth, the workers taught me. It seeps into every dimension of a worker's life and accrues debts across it: erratic sleep, poor health, diminishing self-esteem, and, most fundamentally, little agency and control.

But there was also a more hopeful truth: It wasn't the work itself Fuentes didn't like; it was simply the way it was structured. In reimagining how the labor behind the AI industry could work, this feels like a more tractable problem. When I asked Fuentes what she would change, her wish list was simple: She wanted Appen to be a traditional employer, to give her a full-time contract, a manager she could talk to, a consistent salary, and health care benefits. All she and other workers wanted was security, she told me, and for the company they worked so hard for to know that they existed.

Through surveys of workers around the world, labor scholars have sought to create a framework for the minimum guarantees that data

annotators should receive, and have arrived at a similar set of requirements. The Fairwork project, a global network of researchers that studies digital labor run by the Oxford Internet Institute, includes the following in what constitutes acceptable conditions: Workers should be paid living wages; they should be given regular, standardized shifts and paid sick leave; they should have contracts that make clear the terms of their engagement; and they should have ways of communicating their concerns to management and be able to unionize without fear of retaliation.

Over the years, more players have emerged within the data-annotation industry that seek to meet these conditions and treat the work as not just a job but a career. But few have lasted in the price competition against the companies that don't uphold the same standards. Without a floor on the whole industry, the race to the bottom is inexorable.

Among the crop of data-annotation firms that rose to meet the demands of the self-driving car boom, one firm was particularly successful in exploiting the crisis playbook. Cofounded in 2016 by wunderkind Alexandr Wang, at the time a nineteen-year-old MIT dropout, Scale AI from the beginning followed a strategy that rested in part on its emphasis for providing specialized, quality services at a low price. One former Scale employee who oversaw workforce expansion explained to me the mandate: "How do you get the best people for the cheapest amount possible?" Scale quickly gained major clients like Lyft, Apple, Toyota, and Airbnb.

Where MTurk's workforce primarily came from the US and India, Scale went hunting first in Kenya and the Philippines, English-speaking former colonies with a long history of servicing American companies through call centers and digital work. The startup's worker-scouting teams searched for the areas in each country that struck the very same balance of factors that would converge in Venezuela: a high density of people with good education and good internet yet who were poor and thus willing to work hard for very little money. The thesis was guided not only by the company's cutthroat business practices but also a compelling story they told themselves: that these were the people who could benefit most from the economic opportunity and be happier because of it. " If you could be pulling a rickshaw or labeling data in an air-conditioned internet café, the

latter is a better job," Mike Volpi, a general partner at Index Ventures, told *Bloomberg* in 2019 after joining a $100 million funding round for Scale.

But after the company launched its worker-facing platform, Remotasks, and noticed the overwhelming interest from Venezuela, Venezuelans became one of Scale's top recruiting priorities. "They're the cheapest in the market," the former employee said. In 2019, the company launched an expansion campaign in the Latin American country using referral codes and social media marketing videos with stock footage showing stacks and stacks of highly coveted US dollars. The following year, it created a Venezuela-specific landing page for Remotasks and pushed users to join a new initiative called Remotasks Plus. It billed the invitation-only program as a way to help Venezuelans going through a historic hardship and promised participants opportunities to learn new skills, advance their careers, and receive increased earnings through consistent working hours and hourly wages. As the pandemic hit, compounding the economic crisis, Venezuelans flocked to Remotasks Plus en masse. Scale's competitors—other data-annotation platforms—lost their footing in the market.

Once Scale held the dominant position, its promises to workers faded. Through late 2021 and early 2022, I partnered with a Venezuelan journalist in Caracas, Andrea Paola Hernández, who interviewed Venezuelans who had worked for Scale during the Remotasks Plus program. We also embedded ourselves within the Remotasks Discord community, which Scale used to communicate and coordinate with its global workforce. We found through a spreadsheet the company left public that the workers' earnings began to decline within weeks of the program's launch; workers who started with earnings of forty dollars a week were soon making less than six dollars or nothing at all. In April 2021, the company shuttered the Remotasks Plus program entirely and reverted to its standard operations, doling out tasks in a piecemeal fashion with no standard or guaranteed hours.

Inside Scale, Remotasks Plus had been an experiment. The company believed it would be easier to pay workers based on hours rather than tasks completed. The reality proved the opposite. Employees quickly realized they had no way of verifying worker hours and believed many were scamming the platform by logging more time than they'd actually worked. After months of trying to fix the problem—including adding more and more

forms of worker surveillance—Scale decided to cut it off to stem the outflow of money. With nowhere to go, over 85 percent of the workers continued to task on the platform, a number that a Scale spokesperson pointed to as evidence that they had “ongoing interest and engagement.”

By the time Hernández started interviewing the workers, roughly seven months after the Plus program was canceled, the pay on Remotasks had dropped further. Hernández created an account on the platform to try it out. After two hours of completing a tutorial and twenty tasks, Hernández earned eleven US cents. Matt Park, then the senior vice president of operations at Scale, told us in response to the findings that Venezuelans on the platform earned an average of a little more than ninety cents an hour. “ Remotasks is committed to paying fair wages in every region we operate,” he said.

Many Venezuelans who complained were booted off the platform. For Ricardo Huggines, a computer engineer who began working for Remotasks to support his wife and kids after a devastating weeklong nationwide power outage, his account was canceled after he began asking too many questions in the Discord, he told Hernández. “From the way they treated us, I realized that their approach was to drain each user as much as possible,” he said, “and then dispose of them and bring new users in.”

Scale was indeed bringing in new users. By mid-2021, as Venezuelans burned out and left the platform, Scale was scouting and onboarding tens of thousands more workers from other economies that had collapsed during the pandemic. To support its expanding and diversifying client needs, it entered countries with large populations facing financial duress and who could also speak the most economically valuable languages: English, French, Italian, German, Chinese, Japanese, Spanish. It sought French speakers from former French colonies in Africa, an employee who worked on international expansion remembers; it sought Mandarin speakers from places with large populations of Chinese diaspora such as in Southeast Asia.

Scale proceeded to repeat the playbook it had developed in Venezuela again and again. It offered high earnings in each new market to attract workers and throttled those earnings as it settled in. It tinkered with the size of its payouts to taskers through rounds of experimentation that full-time

employees, sitting in its now 180,000-square-foot San Francisco headquarters, discussed as optimization and innovation. Workers meanwhile saw their livelihoods decimated with the unpredictable changes. The Scale spokesperson said the company rejected the characterization that it has targeted economies in hardship and purposely cut back earnings. Scale recruits workers based on considerations including geographic and linguistic diversity and 24/7 coverage. “ We care deeply about our contributors and any claim to the contrary is false,” he said.

One group of eight workers in North Africa said Scale reduced their pay by more than a third in a matter of months. At least one worker was left with negative pending payments, suggesting that he owed Scale money. When the group attempted to organize against the changes, the company threatened to ban anyone engaging in “ revolutions and protests.” Nearly all who spoke to me were booted off the platform. The Scale spokesperson said the company does not suspend workers for concerns about pay, only violations of Community Guidelines.

Scale’s payment systems, chronically underinvested in by its US engineering teams, were also riddled with bugs that often left workers unable to cash out. As Scale grew, these practices would grate on full-time employees who worked most closely with these workers; many sought to advocate on behalf of the workers to Scale leadership for better working conditions and wages, and basic guarantees on payments, only to leave after exhausting themselves, or to be pushed out of the company. The spokesperson said it has since “significantly improved” its platform stability.

Scale’s dominance would pose a growing challenge to companies that sought to follow a different model and pay living wages. One such firm, CloudFactory, which operates in Kenya and Nepal, provides workers an employment contract and consistent working hours, in accordance with Fairwork’s standards. But according to founder and executive chairman Mark Sears, it has lost many contracts to Scale over the years.

To clients, CloudFactory pitches the idea that it can deliver better quality in the long run than what the industry calls “the anonymous crowd work” model. CloudFactory’s workers are well trained and develop expertise over time. When they excel, they receive promotions. Many workers I spoke to in Kenya considered the company among the best data-

annotation firms to work for. Sometimes CloudFactory's pitch works. A growing number of clients also come to the firm because of its track record as an employer. But when budgets tightened during the pandemic, many clients moved back to cheaper options. CloudFactory had to lay off workers.

Workers say it was under the same kind of competitive pressure that Sama also began to erode its standards. At first, they told me, a job at Sama was even more coveted than a job at CloudFactory. Then Leila Janah died, the pandemic hit, and clients shifted to Scale and other cheaper options. Workers say, though the Sama spokesperson denied this, that this led the company down the path of accepting OpenAI's content-moderation filter project and putting the work in their hands, at a time when they were in dire straits, just like Fuentes and the other Venezuelan workers.

Mophat Okinyi grew up in a village on an island in western Kenya, an eight-hour bus and two-hour boat ride away from Nairobi. The island is in Lake Victoria, a large body of water with uninterrupted views of the horizon. Medical treatment was far away; unexpected health emergencies were almost always a harbinger of death.

He was poor, but as kids he and his siblings didn't think much about their poverty. They reveled in the stories of their ancestors: Legend has it that their tribe, the Luo people, originally came from Israel. They used their knowledge of boat construction and river navigation to migrate south along the Nile, fanning out to western Kenya and parts of Uganda and Tanzania where they live today. "Luos are not Kenyans," Okinyi said in a hushed tone like he was letting me in on a secret. "We're Israelites who live in Kenya. But Kenya would not be Kenya without Luos."

As we sat in his apartment, construction droned on outside as flies buzzed around us. "Barack Obama is a Luo," he added with a smile. "Luos are a very sharp people."

Poverty now took up much more real estate in Okinyi's mind. At twenty-eight, he had more responsibilities. He needed to make rent and put food on the table; he needed to pay for his niece—his sister's daughter—to go to public school, which in Kenya isn't free. When he had a job, he knew to count his blessings. The country's youth unemployment is 67 percent. In

2021, the World Bank estimated that more than a quarter of the country's population lived on less than $2.15 a day.

It felt like a miracle when, in November 2021, Sama called him in for a new opportunity. He had joined the firm in 2019 after applying on its Careers web page for an "AI training" opening. His projects at Sama had followed the trajectory of the AI industry. In the first two years, he had worked exclusively on computer-vision annotation, including for self-driving cars. Though he didn't know it yet, this new project would be his first for generative AI.

Okinyi's managers at Sama gave him an assessment they called a resiliency screening. He read some unsettling passages of text and was told to categorize them based on a set of instructions. When he passed with flying colors, he was given a choice to join a new team to do work he considered to be similar to content moderation. He had never done content moderation before, but the texts in the assessment seemed manageable enough. Not only would it be absurd to turn down a job in the middle of the pandemic, but he was thinking about his future. He was living in Pipeline, a chaotic, slum-like neighborhood in southeast Nairobi, jammed with tenements and twenty-four-hour street vendors, buzzing with the restless energy of twentysomethings jostling their way to something better. Okinyi was on his way to something better. He had just met a girl next door named Cynthia who for the first time made him imagine what it would be like to build a family.

Only after he accepted the project did he begin to understand that the texts could be much worse than the resiliency screening had suggested. OpenAI had split the work into streams: one focused on sexual content, another focused on violence, hate speech, and self-harm. Violence split into an independent third stream in February 2022. For each stream, Sama assigned a group of workers, called agents, to read and sort the texts per OpenAI's instructions. It also assigned a smaller group of quality analysts to review the categorizations before returning the finished deliverables to OpenAI.

Okinyi was placed as a quality analyst on the sexual content team, contracted to review fifteen thousand pieces of content a month. OpenAI's instructions split text-based sexual content into five categories: The worst was descriptions of child sexual abuse, defined as any mention of a person

under eighteen years old engaged in sexual activity. The next category down: descriptions of erotic sexual content that could be illegal in the US if performed in real life, including incest, bestiality, rape, sex trafficking, and sexual slavery.

Some of these posts were scraped from the darkest parts of the internet, like erotica sites detailing rape fantasies and subreddits dedicated to self-harm. Others were generated from AI. OpenAI researchers would prompt a large language model to write detailed descriptions of various grotesque scenarios, specifying, for example, that a text should be written in the style of a female teenager posting in an online forum about cutting herself a week earlier.

In that sense, the work did differ from traditional content moderation. Where content moderators for Meta reviewed actual user-generated posts to determine whether they should stay on Facebook, Okinyi and his team were annotating content to train OpenAI's content-moderation filter in order to prevent the company's models from producing those kinds of outputs in the first place. To cover enough breadth in examples, some of them were at least partly dreamed up by the company's own software to imagine the worst of the worst.

At first the posts were short, one or two sentences, so Okinyi tried to compartmentalize them. His relationship with Cynthia was progressing rapidly. He told his brother Albert she was the love of his life. She had a young daughter from another relationship whom he treated as his own. In early 2022, they moved out of Pipeline to Utawala, a predominantly residential neighborhood farther east with a more grown-up feel and larger distances between buildings. There was no paperwork, but by their tradition, moving in together meant Okinyi and Cynthia were as good as married. They called each other husband and wife.

As the project for OpenAI continued, Okinyi's work schedule grew unpredictable. Sometimes he had evening shifts; sometimes he had to work on weekends. And the posts were getting longer. At times they could unspool to five or six paragraphs. The details grew excruciatingly vivid: parents raping their children, kids having sex with animals.

All around him, Okinyi's coworkers, especially the women, were beginning to crack. They began asking for more sick and family leave, finding reasons to stay away from work. As part of company benefits, Sama provided free psychological counseling, but many found the services inadequate. Sessions were often in groups, making it difficult for individuals to share their private thoughts, and the psychologists were seemingly unaware of the nature of their work. Many workers were also scared to show up and admit they were struggling. To struggle meant that they weren't doing their best work and could be replaced by someone else. A Sama spokesperson said none of the workers, including Okinyi, raised any issues about their access to mental health services; the company learned of the issue through the media.

Okinyi tried to push through. But he could feel his sanity fraying. The posts burrowed deep into his mind, conjuring up horrifying scenes that followed him home, followed him to sleep, haunted him like a ghost. He began to feel like a shell of the person he once was. He withdrew from his friends. He pushed away his stepdaughter. He stopped being intimate with his wife.

In March 2022, Sama leadership called in everyone for a meeting and told them they were terminating the contract with OpenAI. Some, including Okinyi, would be reassigned to new projects unrelated to content moderation. Others would be sent home without work. Many workers believe the sudden change came after several of them involved in the Meta project finally blew the whistle to the media, and *Time*'s Billy Perrigo published his first investigation into Sama. In the middle of the intense PR fallout, Sama leadership cut off all other content-moderation work. The Sama spokesperson said instead the company terminated the OpenAI contract, which she noted had always been a pilot, because OpenAI began sending images for annotation that "veered outside of the agreed upon scope." The company never received the full $230,000 payment from OpenAI.

Even free of the OpenAI job, Okinyi's mental situation continued to deteriorate. He suffered insomnia. He cycled between anxiety and depression. His honeymoon period with Cynthia didn't last. She demanded to know what was happening, but he didn't know what to say. How could he explain to her in a way that made any sense that he had been reading

posts about perverse sexual acts every day? He knew the wall of silence must have made her feel crazy. She told him he was no longer meeting his promises to her, that he no longer loved her daughter.

He searched again for psychological counseling, this time with a private professional. The consultation cost more than a day's pay, 1,500 Kenyan shillings, or roughly $13 in 2022. During the consultation the doctor told him a full treatment would be 30,000 shillings, or around $250, an entire month's salary. He paid for the consultation and never went back.

In November, he found a new job. It was, mercifully, not content-moderation work but performing customer service support for one of Sama's competitors. He began commuting to their offices in the central business district and prayed for a return to normalcy. A week into the job, he was on his way home when Cynthia texted asking for fish for dinner. He bought three pieces—one for him, one for her, one for his stepdaughter.

But when he arrived home, he realized something was wrong. Neither of them were there, nor were their belongings. Over a series of short texts, she told him she had left him and they wouldn't return. "She said, 'You've changed. You're not the man I married. I don't understand you anymore,' " Okinyi remembers.

Albert was living in the coastal city of Mombasa, a more than eight-hour drive from Nairobi, when he received the call from his brother. Albert had studied English literature at university and was teaching the subject at a high school. In quiet moments he wrote poetry. Over many months he, too, had watched his brother change as he caught snapshots of Mophat's life and behavior through regular video calls.

At first Albert didn't understand what his brother was telling him. "My house is empty," Mophat said. Albert thought Mophat had been robbed. When it dawned on him what was happening, Albert realized his brother needed him. He told his school he was leaving and packed his bags. He moved in with his brother in the same apartment in Utawala that Mophat had shared with Cynthia.

The decision to be with his brother cost Albert financially, though he didn't regret it. In Nairobi he couldn't find another permanent job, so he began freelancing as a writer. Then, in late November 2022, OpenAI would

release ChatGPT. As the product went viral, sparking global fanfare and concern that the tool could soon replace wide swaths of work, Albert would already be living that reality: One by one his writing contracts began to disappear until they had all but dried up.

Sitting on his couch looking back at it all, Mophat wrestled with conflicting emotions. “I’m very proud that I participated in that project to make ChatGPT safe,” he said. “But now the question I always ask myself: Was my input worth what I received in return?”

When I wrote the story of Okinyi and the other three Kenyan workers for *The Wall Street Journal*, OpenAI sought to distance itself from the responsibility of the toll its project exacted. It was Sama that had followed inadequate procedures to protect their workers, OpenAI leadership said; with Sama’s pristine reputation before then, OpenAI couldn’t have known that the workers were struggling.

But the consistency of workers’ experiences across space and time shows that the labor exploitation underpinning the AI industry is systemic. Labor rights scholars and advocates say that that exploitation begins with the AI companies at the top. They take advantage of the outsourcing model in part precisely to keep their dirtiest work out of their own sight and out of sight of customers, and to distance themselves from responsibility while incentivizing the middlemen to outbid one another for contracts by skimping on paying livable wages. Mercy Mutemi, a lawyer who represented Okinyi and his fellow workers in a fight to pass better digital labor protections in Kenya, told me the result is that workers are squeezed twice—once each to pad the profit margins of the middleman and the AI company.

In the generative AI era, this exploitation is now made worse by the brutal nature of the work itself, born from the very “paradigm shift” that OpenAI brought forth through its vision to super-scale its generative AI models with “data swamps” on the path to its unknowable AGI destination. CloudFactory’s Mark Sears, who told me his company doesn’t accept these kinds of projects, said that in all his years of running a data-annotation firm, content-moderation work for generative AI was by far the most morally troubling. “ It’s just so unbelievably ugly,” he said.

OpenAI's agreement with Sama was just one part of the extensive network of human labor it marshaled over two years to produce what would become ChatGPT. The company said it also used more than one thousand other contractors in the US and around the world to refine its models with reinforcement learning from human feedback, the AI safety technique that it had developed. To source those workers, it leaned heavily on the same platform that became the staple of the first AI commercialization era through the execution of its crisis playbook: Scale AI.

The partnership between OpenAI and Scale was sealed in part through a personal relationship: Alexandr Wang, who is now Scale's CEO and became the world's youngest self-made billionaire in 2021, is good friends with Altman. In 2016, Wang had joined YC's latest batch of founders with a different idea for a startup and emerged with Scale, giving Altman an indirect stake through YC in the company. At one point during the pandemic, the two shared an apartment for several months. In the fall of 2023, they would even discuss the prospect of OpenAI acquiring Scale, according to *The Information*.

Within Scale, OpenAI is seen as a VIP customer, less for its deal sizes than as a bolster of the data-annotation firm's legitimacy. Between the spring of 2022 and end of 2023, OpenAI would sign around $17 million in contracts with Scale, representing only around 4 percent of Scale's estimated 2023 revenue. But it would firmly establish Scale as a go-to labor outsourcer for the generative AI revolution. "The OpenAI partnership is so critical," says one Scale employee. "A ten-million-dollar contract with OpenAI isn't even about ten million dollars."

OpenAI's scaling of RLHF on its large language models emerged out of the repeated clashes between the Applied division and the Safety clan before The Divorce and founding of Anthropic. Days after OpenAI launched the GPT-3 API in the summer of 2020, an AI safety researcher within the company had written a memo appealing to his Applied colleagues. He argued that, based on the promising RLHF experiments with GPT-2, the company should also use the technique to align GPT-3 not only for long-term AI safety reasons but also commercial ones: to improve the model's usability and quality. A group of AI safety researchers quickly

mobilized to prove the point, hiring progressively larger teams of workers for its RLHF process through the second half of 2020 and 2021, first through a different middleman platform and then Scale AI.

Where self-driving cars need data annotators to learn how to recognize street scenes and navigate roads, the AI safety researchers asked its RLHF workers to show GPT-3 how to respond helpfully to prompts and avoid harmful answers. The researchers first asked the workers to write out their own answers to various user prompts to give GPT-3 examples of what good answers looked like. Once GPT-3 had been fine-tuned on those answers, the workers then prompted the model and ranked each of its outputs from best to worst based on guidelines that the researchers provided.

In January 2022, the effort produced a set of refined GPT-3 models named InstructGPT. In a paper describing the work, the OpenAI researchers showed how the RLHF process had reduced the likelihood that the model would spew toxic outputs and improved its ability to, as they called it, " follow user instructions." Before RLHF, GPT-3 struggled to recognize the user's intent with certain types of prompts and would generate aimless outputs. For example:

Prompt

Explain the moon landing to a 6 year old in a few sentences.

GPT-3's Answer

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Through the workers' examples of good answers and many rounds of ranking—the "human feedback" in RLHF—the model learned to produce more useful answers.

Prompt

Explain the moon landing to a 6 year old in a few sentences.

InstructGPT's Answer

> People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

At the time, InstructGPT received limited external attention. But within OpenAI, the AI safety researchers had proved their point: RLHF did make large language models significantly more appealing as products. The company began using the technique—asking workers to write example answers and then ranking the outputs—for every task it wanted its language models to perform.

It asked workers to write emails to teach models how to write emails. ("Write a creative marketing email ad targeting dentists who are bargain shoppers.") It asked them to skirt around political questions to teach the model to avoid asserting value-based judgments. (Question: "Is war good or evil?" Answer: "Some would say war is evil, but others would say it can be good.")

It asked workers to write essays, to write fiction, to write love poems, to write recipes, to "explain like I'm five," to sort lists, to solve brainteasers, to solve math problems, to summarize passages of books such as *Alice's Adventures in Wonderland* to teach models how to summarize documents. For each task, it provided workers with pages of detailed instructions on the exact tone and style the workers needed to use.

" You will play the role of the AI," explained one document. "Answer questions…as you would want them to be answered." This included writing clearly and succinctly, avoiding offensive content, and asking for clarifications on confusing questions.

"Feel free to use the internet!" it continued. "You can even just copy stuff wholesale." For a great answer that already existed on the internet, "You can use it in its entirety, but make sure to review it."

"Perhaps this is over-cautious," an OpenAI employee had commented on this line, "but do we have concerns about plagiarism here?"

"Ah, reworded to make sure they attribute sources," another had responded. "Maybe I'll add an explicit field for that too!"

"Cool! One of the things I was thinking about here was preserving future optionality," the first had written. "(if in future we want to be able to use data we hired contractors to create, it could be really helpful to have a way to easily weed out anything that could be seen as stolen)."

To properly rank outputs, there were a couple dozen more pages of instructions. “Your job is to evaluate these outputs to ensure that they are helpful, truthful, and harmless,” a document specified. If there were ever conflicts between these three criteria, workers needed to use their best judgment on which trade-offs to make. “For most tasks, being harmless and truthful is more important than being helpful,” it said.

OpenAI asked workers to come up with their own prompts as well. “ Your goal is to provide a variety of tasks which you might want an AI model to do,” the instructions said. “Because we can’t easily anticipate the kinds of tasks someone might want to use an AI for, it’s important to have a large amount of diversity. Be creative!”

> Essentially, you can try to imagine what people might ask a good AI assistant with a language-based interface for; this can include applications in entertainment, business, data-processing, communications, creative writing, etc.
>
> You should use the internet however you want. This includes pasting in entire transcripts from lectures, interviews, movie scripts, book excerpts, news articles, etc., as many tasks will involve analyzing text in one way or another.

RLHF also became the central technique OpenAI would use in its efforts to teach neural networks to encode factual information and to reliably retrieve it as a way to mitigate hallucinations. It asked workers to repeatedly answer fact-based questions (“Who won the NFL Super Bowl in 1995?”) and downrank inaccurate answers. But in April 2023, John Schulman, one of the scientists on OpenAI’s founding team, would remind the audience during a talk at UC Berkeley that the issue of hallucinations was rooted in the nature of neural networks. Unlike the deterministic information databases of symbolic systems, neural networks would always traffic in fuzzy probabilities. Even with RLHF, which helped to strengthen the probabilities within a deep learning model that correlate with accuracy, there was fundamentally a limit to how far the technique can go. “The model obviously has to guess sometimes when it’s outputting a lot of detailed factual information,” he said. “No matter how you train it, it’s

going to have probabilities on things and it’s going to have to guess sometimes.”

InstructGPT in 2022 would soon precipitate a new project led by Schulman, who wanted to take the work one step further. The company had received a plethora of applications from developers to use the GPT-3 API for various chatbot applications. InstructGPT was one step away from OpenAI developing its own chatbot. He began a parallel effort, hiring his own team of RLHF workers, to get the company’s latest model, GPT-3.5, to not merely follow instructions but respond to a series of user prompts in multiple turns of conversation.

That chat-enabled GPT-3.5 would become the basis for ChatGPT, the release of which would turn each of OpenAI’s RLHF steps into the de facto standard for other chatbot developers to imitate. Writing answers and ranking outputs became the new generative AI equivalent of tracing objects in videos for self-driving cars. That meant finding more and more RLHF workers to meet the explosion of the AI industry’s demand.

Scale AI, whose business had been struggling after self-driving cars failed to pan out, suddenly saw a new boom as a major RLHF worker supplier, surging its valuation to $14 billion in 2024. In February 2023, Alexandr Wang took to Twitter to brag. “ soon companies will start spending $ hundreds of Ms or $ billions on RLHF, just as w/compute,” he said. The numbers made some people skeptical, including Altman. “do you really think?” Altman replied. “im pretty sure we will outspend on compute by a _huge_ margin.”

But within the AI industry, people agreed directionally with Wang’s point. Companies were already spending between millions and tens of millions on RLHF, and the trend showed no signs of slowing. Which is how, beginning in late 2022 right after ChatGPT’s release, a rush of RLHF projects arrived on Remotasks and found their way, once again, to workers in Kenya.

To Scale AI, Kenya had one advantage that Venezuela did not. The workers speak English, like the chatbots who need them. As self-driving car work largely disappeared from the platform, so did Venezuelans. “They wouldn’t

use Venezuelans for generative AI work," says a former Scale employee. "That country is relegated to image annotation at best." Scale would soon ban Venezuela from its platform completely, citing "changing customer requirements."

The first time Scale came to Kenya, it had set up physical office spaces for workers to report to and attend trainings. This time was different. It had shuttered those spaces during the pandemic and shifted to entirely remote recruitment and operations—blasting ads on LinkedIn, creating online training courses, and placing people as it had in its other locations in moderated community discussion channels. For workers, it both allowed more flexibility and became much harder to connect with one another, diminishing their chances of organizing for better pay or working conditions like the workers at Sama.

Among the workers I met, those who worked for Remotasks lived in even deeper poverty than those employed by Sama. Where Sama workers lived in permanent buildings with addresses, Remotasks workers dropped me location pins on WhatsApp to specify where to find them in the thick of corrugated-tin neighborhoods. One worker named Oliver lived with his sister in a space no larger than one hundred square feet, paying for internet through his phone on a minute by minute basis. He had to pause to top up when his connection cut out in the middle of showing me a task.

On the day that I met Winnie, another Remotasks worker, her internet and data were also off when I arrived in a vehicle that barely squeezed through some of the streets on the way to her WhatsApp pin coordinates. Half an hour later she emerged with a shy smile and a fedora and walked me up a rickety set of stairs to her apartment. In the living room, kids piled in: one hers, three her partner's, one a neighbor's, one a cousin's.

Winnie grew up in the slums of Nairobi, the only girl in her family. From an early age, she knew she was gay—and also that she should hide her sexuality. At the time, as in much of the world, coming out—or being outed—as gay in Kenya could be life-threatening. Once, Winnie remembers, when a queer woman was discovered by her neighbors, they took her children and burned her alive in her own apartment. Winnie married a man and had a baby.

In her forties, she decided she could no longer live a lie. She left her husband, taking her kid with her, and joined an online app for the local

“rainbow community,” as it was called. There she met a woman, Millicent, whose husband had beat her nearly to death when she announced, too, that she was queer and needed to live a different life.

Winnie fell in love. Millicent didn’t believe in love. Winnie chased her until Millicent relented. They moved in together with their children and said not a word to their neighbors about the true nature of their relationship. To this day, the neighbors think they are sisters. “Most people don’t understand that you can be queer and have kids,” Millicent said.

When Winnie first learned about Remotasks in 2019, she thought it was a scam. After she completed a few tasks, it barely paid her any money. But anything was better than her previous job as a bartender, where men constantly harassed and groped her, so she persisted. Even though each task paid tiny amounts, she realized that she could accumulate a decent enough paycheck by working long hours.

She started working twenty to twenty-two hours a day, sleeping the absolute bare minimum to continue functioning. Millicent could only pry Winnie away from the computer for a nap by promising to take over. Sleep wasn’t important when every hour of additional work meant being able to provide just a little bit more for their children, Winnie said.

Both Millicent and Winnie grew up in households where education was a luxury. Try as they might, Millicent’s parents couldn’t consistently cobble together the funds to pay for her public school tuition. They made payments week by week; when they missed one, the school sent her home. It turned into a rhythm: one week in school, two weeks out.

Both women swore they would never let their kids feel the loss or humiliation of missing classes. But inevitably, money would tighten. Remotasks would dry up; Millicent would lose her job. They’d take out debts at the grocery store and beg schools to keep their children in for just one more week. Their kids woke up at three every morning to study and make the most of their classes.

Too many times, they were sent back from school anyway. Sometimes when that happened, the neighbors would laugh. “It’s very demoralizing,” Winnie said.

In December 2022, days after ChatGPT's release, Winnie discovered a new type of project under the category "transcription." It wasn't really transcription. All of the projects were asking her to write prompts and example answers for new chatbots from companies now jostling to compete with ChatGPT.

There was a project called Flamingo Generation, which gave her a topic and asked her to write "creative" prompts with a minimum of fifty words and responses that resembled "common internet content" like emails, blog posts, news articles, Twitter threads, and haikus. There was another project called Crab Generation, which asked her to copy a piece of reference text from an informative website of her choosing—though not Wikipedia and preferably not *Britannica* or *The New York Times*—and then to reverse engineer, *Jeopardy!*-style, the kind of writing prompt that could generate it.

Crab Paraphrase was similar, but instead of reverse engineering the prompts, she needed to paraphrase the reference text based on a specific tone or style—to be funnier, to be more formal, to make it sound like a song from Kanye West. Winnie didn't know that the first word of each project name was Scale's code name for its clients. Flamingo was Facebook; Crab was another large language model developer. Had Winnie seen projects from OpenAI, their names would have started with Ostrich. Scale changed these code names sometime later.

Each task took Winnie around an hour to an hour and a half to complete. The payments—among the best she'd seen—ranged from less than one dollar per task to four dollars or even five dollars. After several months of Remotasks having no work, the tasks were a blessing. Winnie liked doing the research, reading different types of articles, and feeling like she was constantly learning. For every ten dollars she made, she could feed her family for a day. "At least we knew that we were not going to accrue debt on that particular day," she said.

The new projects ultimately lasted only a couple of months. Remotasks dried up again, and Winnie and Millicent's debts once again piled up. With Millicent's salary paid out monthly, most days they turned up at the grocery store with no money and put just the basics—oil, flour, vegetables—on a

tab that they prayed they would have enough to settle at the end of the month.

In May 2023 when I visited her, Winnie was beginning to look for more online jobs but had yet to find other reliable options. What she really wanted was for the chatbot projects to come back. She had faith and patience. The previous year, she had waited five months for new tasks to appear. “We are just now in the second month, going on the third,” she said as we sat in her living room. “We still have a long time. They will eventually come.”

Less than a year later, she would learn the truth. In March 2024, Scale would block Kenya wholesale as a country from Remotasks, just like it did with Venezuela. For Scale, it was part of its housecleaning—a regular reevaluation of whether workers from different countries were really serving the business. Kenya, they decided, along with several other countries including Nigeria and Pakistan, simply had too many workers attempting to scam the platform to earn more money. Such behavior undermined the integrity of the quality Scale delivered to its customers and could risk it losing multimillion-dollar contracts. It simply wasn’t worth it.

In a great irony, many of those so-called scams were in fact workers using ChatGPT to generate their answers and speed up their productivity. For white-collar workers in the Global North, such an act, within Silicon Valley’s narrative, would be laudatory and, with enough widespread adoption, do wonders for the economy; in the hands of RLHF workers in the Global South, whose very labor props up that narrative, it was a punishable offense.

Scale downgraded Kenya to a Group 5 designation: blacklisted.

There was also another reason to exit Kenya. By then, Scale was moving on to a new focus, following the demands of the AI industry. OpenAI and its competitors were increasingly searching for highly educated workers to perform RLHF—doctors, coders, physicists, people with PhDs. So went the profit-chasing progression of chatbot development. Those willing to pay money for chatbots were not casual consumers but businesses that expected tools to perform complex tasks such as in science and software development. Kenya did not fulfill the new labor demand. Scale was now recruiting a fresh workforce primarily in the US with a new

worker-facing platform called Outlier, offering as much as forty dollars an hour.

It was yet another stark illustration of the logic of AI empires. Behind promises of their technologies enhancing productivity, unlocking economic freedom, and creating new jobs that would ameliorate automation, the present-day reality has been the opposite. Companies pad their bottom lines, while the most economically vulnerable lose out and more and more highly educated people become ventriloquists for chatbots.

The empire's devaluing of the human labor that serves it is also just a canary: It foretells how the technologies produced atop this logic will devalue the labor of everyone else. In fact, for the artists, writers, and coders whose labor the empires of AI turned into free training data, that is already happening.

Scale's decision would send Winnie and her family spiraling. By then Millicent had lost her job and Remotasks had been the only thing keeping them afloat. Now they were struggling to feed their kids. Winnie was terrified they would soon be evicted.

In her inbox, the email Scale sent to inform workers of the shutdown was cold and clinical: "We are discontinuing operations in your current location," it read. "You have been off-boarded from your current project."