NEW YORK
June 19–July 2, 2023

How many humans does it take to make tech seem human? Millions.

Inside the AI Factory

By Josh Dzieza

Taskers near Nairobi tagging data for self-driving cars

AI Is a Lot of Work

As the technology becomes ubiquitous, a vast **tasker underclass** is emerging—and not going anywhere.

By Josh Dzieza

Photo-illustration by Paul Sahre

This article is a collaboration between New York *Magazine and The Verge.*


EXPIRED
Label
ATTRIBUTE
PROVIDE HOVER-OVER HINT
Fill in the blank
Ask for advice
DO Label
View
UNCLEAR TO ME
33% ACCURATE
EXTRANEOUS
REVIEW IMAGES
IMPORTANT UPDATES
ADD MISSING
DO
TASK

A FEW MONTHS after graduating from college in Nairobi, a 30-year-old I'll call Joe got a job as an annotator—the tedious work of processing the raw information used to train artificial intelligence. AI learns by finding patterns in enormous quantities of data, but first that data has to be sorted and tagged by people, a vast workforce mostly hidden behind the machines. In Joe's case, he was labeling footage for self-driving cars—identifying every vehicle, pedestrian, cyclist, anything a driver needs to be aware of—frame by frame and from every possible camera angle. It's difficult and repetitive work. A several-second blip of footage took eight hours to annotate, for which Joe was paid about $10.

Then, in 2019, an opportunity arose: Joe could make four times as much running an annotation boot camp for a new company that was hungry for labelers. Every two weeks, 50 new recruits would file into an office building in Nairobi to begin their apprenticeships. There seemed to be limitless demand for the work. They would be asked to categorize clothing seen in mirror selfies, look through the eyes of robot vacuum cleaners to determine which rooms they were in, and draw squares around lidar scans of motorcycles. Over half of Joe's students usually dropped out before the boot camp was finished. "Some people don't know how to stay in one place for long," he explained with gracious understatement. Also, he acknowledged, "it is very boring."

But it was a job in a place where jobs were scarce, and Joe turned out hundreds of graduates. After boot camp, they went home to work alone in their bedrooms and kitchens, forbidden from telling anyone what they were working on, which wasn't really a problem because they rarely knew themselves. Labeling objects for self-driving cars was obvious, but what about categorizing whether snippets of distorted dialogue were spoken by a robot or a human? Uploading photos of yourself staring into a webcam with a blank expression, then with a grin, then wearing a motorcycle helmet? Each project was such a small component of some larger process that it was difficult to say what they were actually training AI to do. Nor did the names of the projects offer any clues: Crab Generation, Whale Segment, Woodland Gyro, and Pillbox Bratwurst. They were non sequitur code names for non sequitur work.

As for the company employing them, most knew it only as Remotasks, a website offering work to anyone fluent in English. Like most of the annotators I spoke with, Joe was unaware until I told him that Remotasks is the worker-facing subsidiary of a company called Scale AI, a multibillion-dollar Silicon Valley data vendor that counts OpenAI and the U.S. military among its customers. Neither Remotasks' or Scale's website mentions the other.

Much of the public response to language models like OpenAI's ChatGPT has focused on all the jobs they appear poised to automate. But behind even the most impressive AI system are people—huge numbers of people labeling data to train it and clarifying data when it gets confused. Only the companies that can afford to buy this data can compete, and those that get it are highly motivated to keep it secret. The result is that, with few exceptions, little is known about the information shaping these systems' behavior, and even less is known about the people doing the shaping.

For Joe's students, it was work stripped of all its normal trappings: a schedule, colleagues, knowledge of what they were working on or whom they were working for. In fact, they rarely called it work at all—just "tasking." They were taskers.

The anthropologist David Graeber defines "bullshit jobs" as employment without meaning or purpose, work that should be automated but for reasons of bureaucracy or status or inertia is not. These AI jobs are their bizarro twin: work that people want to automate, and often think is already automated, yet still requires a human stand-in. The jobs have a purpose; it's just that workers often have no idea what it is.

THE CURRENT AI BOOM—the convincingly human-sounding chatbots, the artwork that can be generated from simple prompts, and the multibillion-dollar valuations of the companies behind these technologies—began with an unprecedented feat of tedious and repetitive labor.

In 2007, the AI researcher Fei-Fei Li, then a professor at Princeton, suspected the key to improving image-recognition neural networks, a method of machine learning that had been languishing for years, was training on more data—millions of labeled images rather than tens of thousands. The problem was that it would take decades and millions of dollars for her team of undergrads to label that many photos.

Li found thousands of workers on Mechanical Turk, Amazon's crowdsourcing platform where people around the world complete small tasks for cheap. The resulting annotated dataset, called ImageNet, enabled breakthroughs in machine learning that revitalized the field and ushered in a decade of progress.

Annotation remains a foundational part of making AI, but there is often a sense among engineers that it's a passing, inconvenient prerequisite to the more glamorous work of building models. You collect as much labeled data as you can get as cheaply as possible to train your model, and if it works, at least in theory, you no longer need the annotators. But annotation is never really finished. Machine-learning systems are what researchers call "brittle," prone to fail when encountering something that isn't well represented in their training data. These failures, called "edge cases," can have serious consequences. In 2018, an Uber self-driving test car killed a woman because, though it was programmed to avoid cyclists and pedestrians, it didn't know what to make of someone walking a bike across the street. The more AI systems are put out into the world to dispense legal advice and medical help, the more edge cases they will encounter and the more humans will be needed to sort them. Already, this has given rise to a global industry staffed by people like Joe who use their uniquely human faculties to help the machines.

Over the past six months, I spoke with more than two dozen annotators from around the world, and while many of them were training cutting-edge chatbots, just as many were doing the mundane manual labor required to keep AI running. There are people classifying the emotional content of TikTok videos, new variants of email spam, and the precise sexual provocativeness of online ads. Others are looking at credit-card transactions and figuring out

what sort of purchase they relate to or checking e-commerce recommendations and deciding whether that shirt is really something you might like after buying that other shirt. Humans are correcting customer-service chatbots, listening to Alexa requests, and categorizing the emotions of people on video calls. They are labeling food so that smart refrigerators don't get confused by new packaging, checking automated security cameras before sounding alarms, and identifying corn for baffled autonomous tractors.

"There's an entire supply chain," said Sonam Jindal, the program and research lead of the nonprofit Partnership on AI. "The general perception in the industry is that this work isn't a critical part of development and isn't going to be needed for long. All the excitement is around building artificial intelligence, and once we build that, it won't be needed anymore, so why think about it? But it's infrastructure for AI. Human intelligence is the basis of artificial intelligence, and we need to be valuing these as real jobs in the AI economy that are going to be here for a while."

The data vendors behind familiar names like OpenAI, Google, and Microsoft come in different forms. There are private outsourcing companies with call-center-like offices, such as the Kenya- and Nepal-based CloudFactory, where Joe annotated for $1.20 an hour before switching to Remotasks. There are also "crowdworking" sites like Mechanical Turk and Clickworker where anyone can sign up to perform tasks. In the middle are services like Scale AI. Anyone can sign up, but everyone has to pass qualification exams and training courses and undergo performance monitoring. Annotation is big business. Scale, founded in 2016 by then-19-year-old Alexandr Wang, was valued in 2021 at $7.3 billion, making him what *Forbes* called "the youngest self-made billionaire," though the magazine noted in a recent profile that his stake has fallen on secondary markets since then.

This tangled supply chain is deliberately hard to map. According to people in the industry, the companies buying the data demand strict confidentiality. (This is the reason Scale cited to explain why Remotasks has a different name.) Annotation reveals too much about the systems being developed, and the huge number of workers required makes leaks difficult to prevent. Annotators are warned repeatedly not to tell anyone about their jobs, not even their friends and co-workers, but corporate aliases, project code names, and, crucially, the extreme division of labor ensure they don't have enough information about them to talk even if they wanted to. (Most workers requested pseudonyms for fear of being booted from the platforms.) Consequently, there are no granular estimates of the number of people who work in annotation, but it is a lot, and it is growing. A recent Google Research paper gave an order-of-magnitude figure of "millions" with the potential to become "billions."

Automation often unfolds in unexpected ways. Erik Duhaime, CEO of medical-data-annotation company Centaur Labs, recalled how, several years ago, prominent machine-learning engineers were predicting AI would make the job of radiologist obsolete. When that didn't happen, conventional wisdom shifted to radiologists using AI as a tool. Neither of those is quite what he sees occurring. AI is very good at specific tasks, Duhaime said, and that leads work to be broken up and distributed across a system of specialized algorithms and to equally specialized humans. An AI system might be capable of spotting cancer, he said, giving a hypothetical example, but only in a certain type of imagery from a certain type of machine; so now, you need a human to check that the AI is being fed the right type of data and maybe another human who checks its work before passing it to another AI that writes a report, which goes to another human, and so on. "AI doesn't replace work," he said. "But it does change how work is organized."

You might miss this if you believe AI is a brilliant, thinking machine. But if you pull back the curtain even a little, it looks more familiar, the latest iteration of a particularly Silicon Valley division of labor, in which the futuristic gleam of new technologies hides a sprawling manufacturing apparatus and the people who make it run. Duhaime reached back farther for a comparison, a digital version of the transition from craftsmen to industrial manufacturing: coherent processes broken into tasks and arrayed along assembly lines with some steps done by machines and some by humans but none resembling what came before.

Worries about AI-driven disruption are often countered with the argument that AI automates tasks, not jobs, and that these tasks will be the dull ones, leaving people to pursue more fulfilling and human work. But just as likely, the rise of AI will look like past labor-saving technologies, maybe like the telephone or typewriter, which vanquished the drudgery of message delivering and handwriting but generated so much new correspondence, commerce, and paperwork that new offices staffed by new types of workers—clerks, accountants, typists—were required to manage it. When AI comes for your job, you may not lose it, but it might become more alien, more isolating, more tedious.

EARLIER THIS YEAR, I signed up for Scale AI's Remotasks. The process was straightforward. After entering my computer specs, internet speed, and some basic contact information, I found myself in the "training center." To access a paying task, I first had to complete an associated (unpaid) intro course.

The training center displayed a range of courses with inscrutable names like Glue Swimsuit and Poster Macadamia. I clicked on something called GFD Chunking, which revealed itself to be labeling clothing in social-media photos.

The instructions, however, were odd. For one, they basically consisted of the same direction reiterated in the idiosyncratically colored and capitalized typography of a collaged bomb threat.

"DO LABEL items that are real and can be worn by humans or are intended to be worn by real people," it read.

"All items below SHOULD be labeled because they are real and can be worn by real-life humans," it reiterated above photos of an Air Jordans ad, someone in a Kylo Ren helmet, and mannequins in dresses, over which was a lime-green box explaining, once again, "DO Label real items that can be worn by real people."

I skimmed to the bottom of the manual, where the instructor had written in the large bright-red font equivalent of grabbing someone by the shoulders and shaking them, "THE FOLLOWING ITEMS SHOULD NOT BE LABELED because a human could not actually put wear any of these items!" above a photo of C-3PO, Princess Jasmine from *Aladdin*, and a cartoon shoe with eyeballs.

Feeling confident in my ability to distinguish between real clothes that can be worn by real people and not-real clothes that cannot, I proceeded to the test. Right away, it threw an onto-

logical curveball: a picture of a magazine depicting photos of women in dresses. Is a photograph of clothing real clothing? *No,* I thought, *because a human cannot wear a photograph of clothing.* Wrong! As far as AI is concerned, photos of real clothes are real clothes. Next came a photo of a woman in a dimly lit bedroom taking a selfie before a full-length mirror. The blouse and shorts she's wearing are real. What about their reflection? Also real! Reflections of real clothes are also real clothes.

After an embarrassing amount of trial and error, I made it to the actual work, only to make the horrifying discovery that the instructions I'd been struggling to follow had been updated and clarified so many times that they were now a full 43 printed pages of directives: DO NOT label open suitcases full of clothes; DO label shoes but do NOT label flippers; DO label leggings but do NOT label tights; do NOT label towels even if someone is wearing it; label costumes but do NOT label armor. And so on.

There has been general instruction disarray across the industry, according to Milagros Miceli, a researcher at the Weizenbaum Institute in Germany who studies data work. It is in part a product of the way machine-learning systems learn. Where a human would get the concept of "shirt" with a few examples, machine-learning programs need thousands, and they need to be categorized with perfect consistency yet varied enough (polo shirts, shirts being worn outdoors, shirts hanging on a rack) that the very literal system can handle the diversity of the real world. "Imagine simplifying complex realities into something that is readable for a machine that is totally dumb," she said.

The act of simplifying reality for a machine results in a great deal of complexity for the human. Instruction writers must come up with rules that will get humans to categorize the world with perfect consistency. To do so, they often create categories no human would use. A human asked to tag all the shirts in a photo probably wouldn't tag the reflection of a shirt in a mirror because they would know it is a reflection and not real. But to the AI, which has no understanding of the world, it's all just pixels and the two are perfectly identical. Fed a dataset with some shirts labeled and other (reflected) shirts unlabeled, the model won't work. So the engineer goes back to the vendor with an update: DO label reflections of shirts. Soon, you have a 43-page guide descending into red all-caps.

"When you start off, the rules are relatively simple," said a former Scale employee who requested anonymity because of an NDA. "Then they get back a thousand images and then they're like, *Wait a second,* and then you have multiple engineers and they start to argue with each other. It's very much a human thing."

The job of the annotator often involves putting human understanding aside and following instructions very, *very* literally—to think, as one annotator said, like a robot. It's a strange mental space to inhabit, doing your best to follow nonsensical but rigorous rules, like taking a standardized test while on hallucinogens. Annotators invariably end up confronted with confounding questions like, Is that a red shirt with white stripes or a white shirt with red stripes? Is a wicker bowl a "decorative bowl" if it's full of apples? What color is leopard print? When instructors said to label traffic-control directors, did they also mean to label traffic-control directors eating lunch on the sidewalk? Every question must be answered, and a wrong guess could get you banned and booted to a new, totally different task with its own baffling rules.

Most of the work on Remotasks is paid at a piece rate with a single task earning anywhere from a few cents to several dollars. Because tasks can take seconds or hours, wages are hard to predict. When Remotasks first arrived in Kenya, annotators said it paid relatively well—averaging about $5 to $10 per hour depending on the task—but the amount fell as time went on.

Scale AI spokesperson Anna Franko said that the company's economists analyze the specifics of a project, the skills required, the regional cost of living, and other factors "to ensure fair and competitive compensation." Former Scale employees also said pay is determined through a surge-pricing-like mechanism that adjusts for how many annotators are available and how quickly the data is needed.

According to workers I spoke with and job listings, U.S.-based Remotasks annotators generally earn between $10 and $25 per hour, though some subject-matter experts can make more. By the beginning of this year, pay for the Kenyan annotators I spoke with had dropped to between $1 and $3 per hour.

That is, when they were making any money at all. The most common complaint about Remotasks work is its variability; it's steady enough to be a full-time job for long stretches but too unpredictable to rely on. Annotators spend hours reading instructions and completing unpaid trainings only to do a dozen tasks and then have the project end. There might be nothing new for days, then, without warning, a totally different task appears and could last anywhere from a few hours to weeks. Any task could be their last, and they never know when the next one will come.

This boom-and-bust cycle results from the cadence of AI development, according to engineers and data vendors. Training a large model requires an enormous amount of annotation followed by more iterative updates, and engineers want it all as fast as possible so they can hit their target launch date. There may be monthslong demand for thousands of annotators, then for only a few hundred, then for a dozen specialists of a certain type, and then thousands again. "The question is, Who bears the cost for these fluctuations?" said Jindal of Partnership on AI. "Because right now, it's the workers."

To succeed, annotators work together. When I told Victor, who started working for Remotasks while at university in Nairobi, about my struggles with the traffic-control-directors task, he told me everyone knew to stay away from that one: too tricky, bad pay, not worth it. Like a lot of annotators, Victor uses unofficial WhatsApp groups to spread the word when a good task drops.

◀ **Remotasks instructions for labeling clothing.**

COURTESY OF THE WRITER

When he figures out a new one, he starts impromptu Google Meets to show others how it's done. Anyone can join and work together for a time, sharing tips. "It's a culture we have developed of helping each other because we know when on your own, you can't know all the tricks," he said.

Because work appears and vanishes without warning, taskers always need to be on alert. Victor has found that projects pop up very late at night, so he is in the habit of waking every three hours or so to check his queue. When a task is there, he'll stay awake as long as he can to work. Once, he stayed up 36 hours straight labeling elbows and knees and heads in photographs of crowds—he has no idea why. Another time, he stayed up so long his mother asked him what was wrong with his eyes. He looked in the mirror to discover they were swollen.

Annotators generally know only that they are training AI for companies located vaguely elsewhere, but sometimes the veil of anonymity drops—instructions mention a brand or a chatbot says too much. "I read and I Googled and found I am working for a 25-year-old billionaire," said one worker, who, when we spoke, was labeling the emotions of people calling to order Domino's pizza. "I really am wasting my life here if I made somebody a billionaire and I'm earning a couple of bucks a week."

Victor is a self-proclaimed "fanatic" about AI and started annotating because he wants to help bring about a fully automated post-work future. But earlier this year, someone dropped a *Time* story into one of his WhatsApp groups about workers training ChatGPT to recognize toxic content who were getting paid less than $2 an hour by the vendor Sama AI. "People were angry that these companies are so profitable but paying so poorly," Victor said. He was unaware until I told him about Remotasks' connection to Scale. Instructions for one of the tasks he worked on were nearly identical to those used by OpenAI, which meant he had likely been training ChatGPT as well, for approximately $3 per hour.

"I remember that someone posted that we will be remembered in the future," he said. "And somebody else replied, 'We are being treated worse than foot soldiers. We will be remembered nowhere in the future.' I remember that very well. Nobody will recognize the work we did or the effort we put in."

IDENTIFYING CLOTHING and labeling customer-service conversations are just some of the annotation gigs available. Lately, the hottest on the market has been chatbot trainer. Because it demands specific areas of expertise or language fluency and wages are often adjusted regionally, this job tends to pay better. Certain types of specialist annotation can go for $50 or more per hour.

There are people **classifying** the emotional content of TikTok videos, new variants of email spam, and the **precise sexual provocativeness** of online ads.

A woman I'll call Anna was searching for a job in Texas when she stumbled across a generic listing for online work and applied. It was Remotasks, and after passing an introductory exam, she was brought into a Slack room of 1,500 people who were training a project code-named Dolphin, which she later discovered to be Google DeepMind's chatbot, Sparrow, one of the many bots competing with ChatGPT. Her job is to talk with it all day. At about $14 an hour, plus bonuses for high productivity, "it definitely beats getting paid $10 an hour at the local Dollar General store," she said.

Also, she enjoys it. She has discussed science-fiction novels, mathematical paradoxes, children's riddles, and TV shows. Sometimes the bot's responses make her laugh; other times, she runs out of things to talk about. "Some days, my brain is just like, *I literally have no idea what on earth to ask it now,*" she said. "So I have a little notebook, and I've written about two pages of things—I just Google interesting topics—so I think I'll be good for seven hours today, but that's not always the case."

Each time Anna prompts Sparrow, it delivers two responses and she picks the best one, thereby creating something called "human-feedback data." When ChatGPT debuted late last year, its impressively natural-seeming conversational style was credited to its having been trained on troves of internet data. But the language that fuels ChatGPT and its competitors is filtered through several rounds of human annotation. One group of contractors writes examples of how the engineers want the bot to behave, creating questions followed by correct answers, descriptions of computer programs followed by functional code, and requests for tips on committing crimes followed by polite refusals. After the model is trained on these examples, yet more contractors are brought in to prompt it and rank its responses. This is what Anna is doing with Sparrow. Exactly which criteria the raters are told to use varies—honesty, or helpfulness, or just personal preference. The point is that they are creating data on human taste, and once there's enough of it, engineers can train a second model to mimic their preferences at scale, automating the ranking process and training their AI to act in ways humans approve of. The result is a remarkably human-seeming bot that mostly declines harmful requests and explains its AI nature with seeming self-awareness.

Put another way, ChatGPT seems so human because it was trained by an AI that was mimicking humans who were rating an AI that was mimicking humans who were pretending to be a better version of an AI that was trained on human writing.

This circuitous technique is called "reinforcement learning from human feedback," or RLHF, and it's so effective that it's worth pausing to fully register what it doesn't do. When annotators teach a model to be accurate, for example, the model isn't learning to check answers against logic or external sources or about what accuracy as a concept even is. The model is still a text-prediction machine mimicking patterns in human writing, but now its training corpus has been supplemented with bespoke examples, and the model has been weighted to favor them. Maybe this results in the model extracting patterns from the part of its linguistic map labeled as accurate and producing text that happens to align with the truth, but it can also result in it mimicking the confident style and expert jargon of the accurate text while writing things that are totally wrong. There is no guarantee that the text the labelers marked as accurate is in fact accurate, and when it is, there is no guarantee that the model learns the right patterns from it.

This dynamic makes chatbot annotation a delicate process. It has to be rigorous and consistent because sloppy feedback, like marking material that merely sounds correct as accurate, risks training models to be even more convincing bullshitters. An early OpenAI and DeepMind joint project using RLHF, in this case to train a virtual robot hand to grab an item, resulted in also training the robot to position its hand between the object and its raters and wiggle around such that it only appeared to its human overseers to grab the item. Ranking a language model's responses is always going to be somewhat subjective because it's language. A text of any length will have multiple elements that could be right or wrong or, taken together, misleading. OpenAI researchers ran into this obstacle in another early RLHF paper. Trying to get their model to summarize text, the researchers found they agreed only 60 percent of the time that a summary was good. "Unlike many tasks in [machine learning] our queries do not have unambiguous ground truth," they lamented.

When Anna rates Sparrow's responses, she's supposed to be looking at their accuracy, helpfulness, and harmlessness while also checking that the model isn't giving medical or financial advice or anthropomorphizing itself or running afoul of other criteria. To be useful training data, the model's responses have to be quantifiably ranked against one another: Is a bot that helpfully tells you how to make a bomb "better" than a bot that's so harmless it refuses to answer any questions? In one DeepMind paper, when Sparrow's makers took a turn annotating, four researchers wound up debating whether their bot had assumed the gender of a user who asked it for relationship advice. According to Geoffrey Irving, one of DeepMind's research scientists, the company's researchers hold weekly annotation meetings in which they rerate data themselves and discuss ambiguous cases, consulting with ethical or subject-matter experts when a case is particularly tricky.

Anna often finds herself having to choose between two bad options. "Even if they're both absolutely, ridiculously wrong, you still have to figure out which one is better and then write words explaining why," she said. Sometimes, when both responses are bad, she's encouraged to write a better response herself, which she does about half the time.

Because feedback data is difficult to collect, it fetches a higher price. Basic preferences of the sort Anna is producing sell for about $1 each, according to people with knowledge of the indus-

try. But if you want to train a model to do legal research, you need someone with training in law, and this gets expensive. Everyone involved is reluctant to say how much they're spending, but in general, specialized written examples can go for hundreds of dollars, while expert ratings can cost $50 or more. One engineer told me about buying examples of Socratic dialogues for up to $300 a pop. Another told me about paying $15 for a "darkly funny limerick about a goldfish."

OpenAI, Microsoft, Meta, and Anthropic did not comment about how many people contribute annotations to their models, how much they are paid, or where in the world they are located. Irving of DeepMind, which is a subsidiary of Google, said the annotators working on Sparrow are paid "at least the hourly living wage" based on their location. Anna knows "absolutely nothing" about Remotasks, but Sparrow has been more open. She wasn't the only annotator I spoke with who got more information from the AI they were training than from their employer; several others learned whom they were working for by asking their AI for its company's terms of service. "I literally asked it, 'What is your purpose, Sparrow?'" Anna said. It pulled up a link to DeepMind's website and explained that it's an AI assistant and that its creators trained it using RLHF to be helpful and safe.

UNTIL RECENTLY, it was relatively easy to spot bad output from a language model. It looked like gibberish. But this gets harder as the models get better—a problem called "scalable oversight." Google inadvertently demonstrated how hard it is to catch the errors of a modern-language model when one made it into the splashy debut of its AI assistant, Bard. (It stated confidently that the James Webb Space Telescope "took the very first pictures of a planet outside of our own solar system," which is wrong.) This trajectory means annotation increasingly requires specific skills and expertise.

Last year, someone I'll call Lewis was working on Mechanical Turk when, after completing a task, he received a message inviting him to apply for a platform he hadn't heard of. It was called Taskup.ai, and its website was remarkably basic: just a navy background with text reading GET PAID FOR TASKS ON DEMAND. He applied.

The work paid far better than anything he had tried before, often around $30 an hour. It was more challenging, too: devising complex scenarios to trick chatbots into giving dangerous advice, testing a model's ability to stay in character, and having detailed conversations about scientific topics so technical they required extensive research. He found the work "satisfying and stimulating." While checking one model's attempts to code in Python, Lewis was learning too. He couldn't work for more than four hours at a stretch, lest he risk becoming mentally drained and making mistakes, and he wanted to keep the job.

"If there was one thing I could change, I would just like to have more information about what happens on the other end," he said. "We only know as much as we need to know to get work done, but if I could know more, then maybe I could get more established and perhaps pursue this as a career."

I spoke with eight other workers, most based in the U.S., who had similar experiences of answering surveys or completing tasks on other platforms and finding themselves recruited for Taskup.ai or several similarly generic sites, such as DataAnnotation.tech or Gethybrid.io. Often their work involved training chatbots, though with higher-quality expectations and more specialized purposes than other sites they had worked for. One was demonstrating spreadsheet macros. Another was just supposed to have conversations and rate responses according to whatever criteria she wanted. She often asked the chatbot things that had come up in conversations with her 7-year-old daughter, like "What is the largest dinosaur?" and "Write a story about a tiger." "I haven't fully gotten my head around what they're trying to do with it," she told me.

Taskup.ai, DataAnnotation.tech, and Gethybrid.io all appear to be owned by the same company: Surge AI. Its CEO, Edwin Chen, would neither confirm nor deny the connection, but he was willing to talk about his company and how he sees annotation evolving.

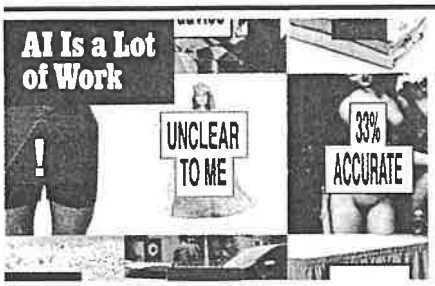
"I've always felt the annotation landscape is overly simplistic," Chen said over a video call from Surge's office. He founded Surge in 2020 after working on AI at Google, Facebook, and Twitter convinced him that crowdsourced labeling was inadequate. "We want AI to tell jokes or write really good marketing copy or help me out when I need therapy or whatnot," Chen said. "You can't ask five people to independently come up with a joke and combine it into a majority answer. Not everybody can tell a joke or solve a Python program. The annotation landscape needs to shift from this low-quality, low-skill mind-set to something that's much richer and captures the range of human skills and creativity and values that we want AI systems to possess."

Last year, Surge relabeled Google's dataset classifying Reddit posts by emotion. Google had stripped each post of context and sent them to workers in India for labeling. Surge employees familiar with American internet culture found that 30 percent of the labels were wrong. Posts like "hell yeah my brother" had been classified as annoyance and "Yay, cold McDonald's. My favorite" as love.

Surge claims to vet its workers for qualifications—that people doing creative-writing tasks have experience with creative writing, for example—but exactly how Surge finds workers is "proprietary," Chen said. As with Remotasks, workers often have to complete training courses, though unlike Remotasks, they are paid for it, according to the annotators I spoke with. Having fewer, better-trained workers producing higher-quality data allows Surge to compensate better than its peers, Chen said, though he declined to elaborate, saying only that people are paid "fair and ethical wages." The workers I spoke with earned between $15 and $30 per hour, but they are a small sample of all the annotators, a group Chen said now consists of 100,000 people. The secrecy, he explained, stems from clients' demands for confidentiality.

Surge's customers include OpenAI, Google, Microsoft, Meta, and Anthropic. Surge specializes in feedback and language annotation, and after ChatGPT launched, it got an influx of requests, Chen said: "I thought everybody knew the power of RLHF, but I guess people just didn't viscerally understand."

The new models are so impressive *(Continued on page 89)*



CONTINUED FROM PAGE 27

they've inspired another round of predictions that annotation is about to be automated. Given the costs involved, there is significant financial pressure to do so. Anthropic, Meta, and other companies have recently made strides in using AI to drastically reduce the amount of human annotation needed to guide models, and other developers have started using GPT-4 to generate training data. However, a recent paper found that GPT-4-trained models may be learning to mimic GPT's authoritative style with even less accuracy, and so far, when improvements in AI have made one form of annotation obsolete, demand for other, more sophisticated types of labeling has gone up. This debate spilled into the open earlier this year, when Scale's CEO, Wang, tweeted that he predicted AI labs will soon be spending as many billions of dollars on human data as they do on computing power; OpenAI's CEO, Sam Altman, responded that data needs will decrease as AI improves.

Chen is skeptical AI will reach a point where human feedback is no longer needed, but he does see annotation becoming more difficult as models improve. Like many researchers, he believes the path forward will involve AI systems helping humans oversee other AI. Surge recently collaborated with Anthropic on a proof of concept, having human labelers answer questions about a lengthy text with the help of an unreliable AI assistant, on the theory that the humans would have to feel out the weaknesses of their AI assistant and collaborate to reason their way to the correct answer. Another possibility has two AIs debating each other and a human rendering the final verdict on which is correct. "We still have yet to see really good practical implementations of this stuff, but it's starting to become necessary because it's getting really hard for labelers to keep up with the models," said OpenAI research scientist John Schulman in a recent talk at Berkeley.

"I think you always need a human to monitor what AIs are doing just because they are this kind of alien entity," Chen said. Machine-learning systems are just too strange ever to fully trust. The most impressive models today have what, to a human, seems like bizarre weaknesses, he added, pointing out that though GPT-4 can generate complex and convincing prose, it can't pick out which words are adjectives: "Either that or models get so good that they're better than humans at all things, in which case, you reach your utopia and who cares?"

As 2022 ended, Joe started hearing from his students that their task queues were often empty. Then he got an email informing him the boot camps in Kenya were closing. He continued training taskers online, but he began to worry about the future.

"There were signs that it was not going to last long," he said. Annotation was leaving Kenya. From colleagues he had met online, he heard tasks were going to Nepal, India, and the Philippines. "The companies shift from one region to another," Joe said. "They don't have infrastructure locally, so it makes them flexible to shift to regions that favor them in terms of operation cost."

One way the AI industry differs from manufacturers of phones and cars is in its fluidity. The work is constantly changing, constantly getting automated away and replaced with new needs for new types of data. It's an assembly line but one that can be endlessly and instantly reconfigured, moving to wherever there is the right combination of skills, bandwidth, and wages.

Lately, the best-paying work is in the U.S. In May, Scale started listing annotation jobs on its own website, soliciting people with experience in practically every field AI is predicted to conquer. There were listings for AI trainers with expertise in health coaching, human resources, finance, economics, data science, programming, computer science, chemistry, biology, accounting, taxes, nutrition, physics, travel, K-12 education, sports journalism, and self-help. You can make $45 an hour teaching robots law or make $25 an hour teaching them poetry. There were also listings for people with security clearance, presumably to help train military AI. Scale recently launched a defense-oriented language model called Donovan, which Wang called "ammunition in the AI war," and won a contract to work on the Army's robotic-combat-vehicle program.

Anna is still training chatbots in Texas. Colleagues have been turned into reviewers and Slack admins—she isn't sure why, but it has given her hope that the gig could be a longer-term career. One thing she isn't worried about is being automated out of a job. "I mean, what it can do is amazing," she said of the chatbot. "But it still does some really weird shit."

When Remotasks first arrived in Kenya, Joe thought annotation could be a good career. Even after the work moved elsewhere, he was determined to make it one. There were thousands of people in Nairobi who knew how to do the work, he reasoned—he had trained many of them, after all. Joe rented office space in the city and began sourcing contracts: a job annotating blueprints for a construction company, another labeling fruits despoiled by insects for some sort of agricultural project, plus the usual work of annotating for self-driving cars and e-commerce.

But he has found his vision difficult to achieve. He has just one full-time employee, down from two. "We haven't been having a consistent flow of work," he said. There are weeks with nothing to do because customers are still collecting data, and when they're done, he has to bring in short-term contractors to meet their deadlines: "Clients don't care whether we have consistent work or not. So long as the datasets have been completed, then that's the end of that."

Rather than let their skills go to waste, other taskers decided to chase the work wherever it went. They rented proxy servers to disguise their locations and bought fake IDs to pass security checks so they could pretend to work from Singapore, the Netherlands, Mississippi, or wherever the tasks were flowing. It's a risky business. Scale has become increasingly aggressive about suspending accounts caught disguising their location, according to multiple taskers. It was during one of these crackdowns that my account got banned, presumably because I had been using a VPN to see what workers in other countries were seeing, and all $1.50 or so of my earnings were seized.

"These days, we have become a bit cunning because we noticed that in other countries they are paying well," said Victor, who was earning double the Kenyan rate by tasking in Malaysia. "You do it cautiously."

Another Kenyan annotator said that after his account got suspended for mysterious reasons, he decided to stop playing by the rules. Now, he runs multiple accounts in multiple countries, tasking wherever the pay is best. He works fast and gets high marks for quality, he said, thanks to ChatGPT. The bot is wonderful, he said, letting him speed through $10 tasks in a matter of minutes. When we spoke, he was having it rate another chatbot's responses according to seven different criteria, one AI training the other. ■