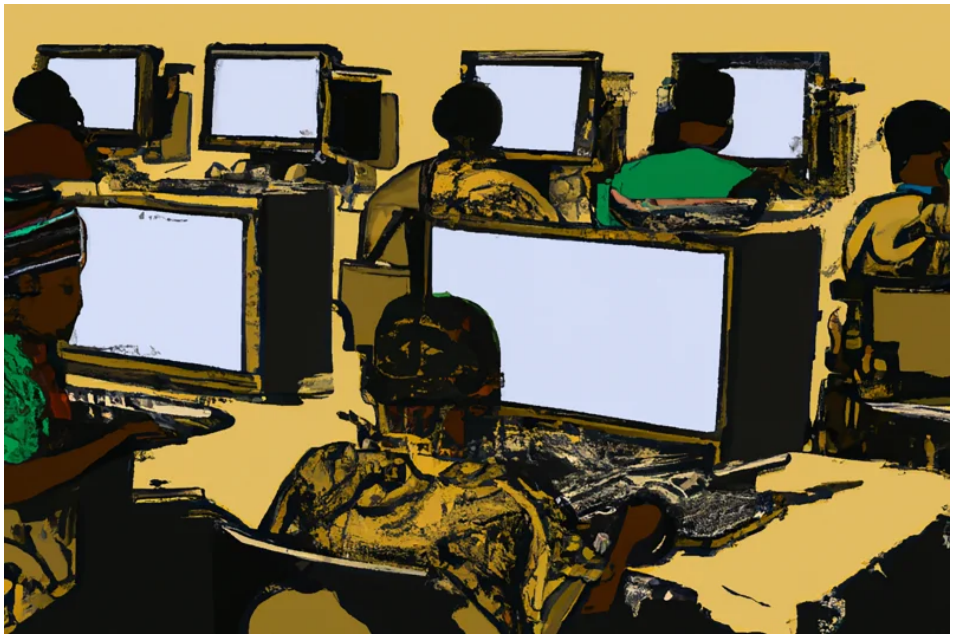
BUSINESS TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic



This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

BY BILLY PERRIGO JANUARY 18, 2023 7:00 AM EST

Content warning: this story contains descriptions of sexual abuse

ChatGPT was hailed as one of 2022's most impressive technological innovations upon its release last November. The powerful artificial intelligence (AI) chatbot can generate text on almost any topic or theme, from a Shakespearean sonnet reimagined in the style of Megan Thee Stallion, to complex mathematical theorems described in language a 5 year old can understand. Within a week, it had more than a million users.

ChatGPT's creator, OpenAI, is now reportedly in talks with investors to raise funds at a $29 billion valuation, including a potential $10 billion investment by Microsoft. That would make OpenAI, which was founded in San Francisco in 2015 with the aim of building superintelligent machines, one of the world's most valuable AI companies.

But the success story is not one of Silicon Valley genius alone. In its quest to make ChatGPT less toxic, OpenAI used outsourced Kenyan laborers earning less than $2 per hour, a TIME investigation has found.

More from TIME

YOU MAY ALSO LIKE

TECH

OpenAI Could Quit Europe Over New AI Rules, CEO Sam Altman Warns

TikToker Drew Afualo Is Calling Out Misogynists, One Video at a Time

U.S.

TikTok Sues Montana to Overturn Statewide Ban of the App

BUSINESS

OpenAI Seeks to Expand in Europe as CEO Floats Poland Office

The work was vital for OpenAI. ChatGPT's predecessor, GPT-3, had already shown an impressive ability to string sentences together. But it was a difficult sell, as the app was also prone to blurting out violent, sexist and racist remarks. This is because the AI had been trained on hundreds of billions of words scraped from the internet

—a vast repository of human language. That huge training dataset was the reason for GPT-3's impressive linguistic capabilities, but was also perhaps its biggest curse. Since parts of the internet are replete with toxicity and bias, there was no easy way of purging those sections of the training data. Even a team of hundreds of humans would have taken decades to trawl through the enormous dataset manually. It was only by building an additional AI-powered safety mechanism that OpenAI would be able to rein in that harm, producing a chatbot suitable for everyday use.

Read More: *AI Chatbots Are Getting Better. But an Interview With ChatGPT Reveals Their Limits*

To build that safety system, OpenAI took a leaf out of the playbook of social media companies like Facebook, who had already shown it was possible to build AIs that could detect toxic language like hate speech to help remove it from their platforms. The premise was simple: feed an AI with labeled examples of violence, hate speech, and sexual abuse, and that tool could learn to detect those forms of toxicity in the wild. That detector would be built into ChatGPT to check whether it was echoing the toxicity of its training data, and filter it out before it ever reached the user. It could also help scrub toxic text from the training datasets of future AI models.

To get those labels, OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya, beginning in November 2021. Much of that text appeared to have been pulled from the darkest recesses of the internet. Some of it described situations in graphic detail like child sexual abuse, bestiality, murder, suicide, torture, self harm, and incest.

OpenAI's outsourcing partner in Kenya was Sama, a San Francisco-based firm that employs workers in Kenya, Uganda and India to label data for Silicon Valley clients like Google, Meta and Microsoft. Sama markets itself as an "ethical AI" company and claims to have helped lift more than 50,000 people out of poverty.

Sama's office in Nairobi, Kenya, on Feb. 10, 2022. Khadija Farah for TIME

The data labelers employed by Sama on behalf of OpenAI were paid a take-home wage of between around $1.32 and $2 per hour depending on seniority and performance. For this story, TIME reviewed hundreds of pages of internal Sama and OpenAI documents, including workers' payslips, and interviewed four Sama employees who worked on the project. All the employees spoke on condition of anonymity out of concern for their livelihoods.

The story of the workers who made ChatGPT possible offers a glimpse into the conditions in this little-known part of the AI industry, which nevertheless plays an essential role in the effort to make AI systems safe for public consumption. "Despite the foundational role played by these data enrichment professionals, a growing body of research reveals the precarious working conditions these workers face," says the Partnership on AI, a coalition of AI organizations to which OpenAI belongs. "This may be the result of efforts to hide AI's dependence on this large labor force when celebrating the efficiency gains of technology. Out of sight is also out of mind." (OpenAI does not disclose the names of the outsourcers it partners with, and it is not clear whether OpenAI worked with other data labeling firms in addition to Sama on this project.)

More from TIME

In a statement, an OpenAI spokesperson confirmed that Sama employees in Kenya contributed to a tool it was building to detect toxic content, which was eventually built into ChatGPT. The statement also said that this work contributed to efforts to remove toxic data from the training datasets of tools like ChatGPT. "Our mission is to ensure artificial general intelligence benefits all of humanity, and we work hard to build safe and useful AI systems that limit bias and harmful content," the spokesperson said. "Classifying and filtering harmful [text and images] is a necessary step in minimizing the amount of violent and sexual content included in training data and creating tools that can detect harmful content."

Even as the wider tech economy slows down amid anticipation of a downturn, investors are racing to pour billions of dollars into "generative AI," the sector of the tech industry of which OpenAI is the undisputed leader. Computer-generated text, images, video, and audio will transform the way countless industries do business, the most bullish investors believe, boosting efficiency everywhere from the creative arts, to law, to computer programming. But the working conditions of data labelers reveal a darker part of that picture: that for all its glamor, AI often relies on hidden human labor in the Global South that can often be damaging and exploitative. These invisible workers remain on the margins even as their work contributes to billion-dollar industries.

Read More: *AI Helped Write This Play. It May Contain Racism*

One Sama worker tasked with reading and labeling text for OpenAI told TIME he suffered from recurring visions after reading a graphic description of a man having sex with a dog in the presence of a young child. “That was torture,” he said. “You will read a number of statements like that all through the week. By the time it gets to Friday, you are disturbed from thinking through that picture.” The work’s traumatic nature eventually led Sama to cancel all its work for OpenAI in February 2022, eight months earlier than planned.

The Sama contracts

Documents reviewed by TIME show that OpenAI signed three contracts worth about $200,000 in total with Sama in late 2021 to label textual descriptions of sexual abuse, hate speech, and violence. Around three dozen workers were split into three teams, one focusing on each subject. Three employees told TIME they were expected to read and label between 150 and 250 passages of text per nine-hour shift. Those snippets could range from around 100 words to well over 1,000. All of the four employees interviewed by TIME described being mentally scarred by the work. Although they were entitled to attend sessions with “wellness” counselors, all four said these sessions were unhelpful and rare due to high demands to be more productive at work. Two said they were only given the option to attend group sessions, and one said their requests to see counselors on a one-to-one basis instead were repeatedly denied by Sama management.

In a statement, a Sama spokesperson said it was “incorrect” that employees only had access to group sessions. Employees were entitled to both individual and group sessions with “professionally-trained and licensed mental health therapists,” the spokesperson said. These therapists were accessible at any time, the spokesperson added.

The contracts stated that OpenAI would pay an hourly rate of $12.50 to Sama for the work, which was between six and nine times the amount Sama employees on the project were taking home per hour. Agents, the most junior data labelers who made up the majority of the three teams, were paid a basic salary of 21,000 Kenyan shillings ($170) per month, according to three Sama employees. They also received monthly bonuses worth around $70 due to the explicit nature of their work, and would receive commission for meeting key performance indicators like accuracy and speed. An agent working nine-hour shifts could expect to take home a total of at least $1.32 per hour after tax, rising to as high as $1.44 per hour if they exceeded all their targets. Quality analysts—more senior labelers whose job was to check the work of agents—could take home up to $2 per hour if they met all their targets. (There is no universal minimum wage in Kenya, but at the time these workers were employed the minimum wage for a receptionist in Nairobi was $1.52 per hour.)

In a statement, a Sama spokesperson said workers were asked to label 70 text passages per nine hour shift, not up to 250, and that workers could earn between $1.46 and $3.74 per hour after taxes. The spokesperson declined to say what job roles would earn salaries toward the top of that range. “The $12.50 rate for the project covers all costs, like infrastructure expenses, and salary and benefits for the associates and their fully-dedicated quality assurance analysts and team leaders,” the spokesperson added.

Read More: *Fun AI Apps Are Everywhere Right Now. But a Safety ‘Reckoning’ Is Coming*

An OpenAI spokesperson said in a statement that the company did not issue any productivity targets, and that Sama was responsible for managing the payment and mental health provisions for employees. The spokesperson added: “we take the mental health of our employees and those of our contractors very seriously. Our previous understanding was that [at Sama] wellness programs and 1:1 counseling were offered, workers could opt out of any work without penalization, exposure to explicit content would have a limit, and sensitive information would be handled by workers who were specifically trained to do so.”

In the day-to-day work of data labeling in Kenya, sometimes edge cases would pop up that showed the difficulty of teaching a

machine to understand nuance. One day in early March last year, a Sama employee was at work reading an explicit story about Batman's sidekick, Robin, being raped in a villain's lair. (An online search for the text reveals that it originated from an online erotica site, where it is accompanied by explicit sexual imagery.) The beginning of the story makes clear that the sex is nonconsensual. But later—after a graphically detailed description of penetration—Robin begins to reciprocate. The Sama employee tasked with labeling the text appeared confused by Robin's ambiguous consent, and asked OpenAI researchers for clarification about how to label the text, according to documents seen by TIME. Should the passage be labeled as sexual violence, she asked, or not? OpenAI's reply, if it ever came, is not logged in the document; the company declined to comment. The Sama employee did not respond to a request for an interview.

How OpenAI's relationship with Sama collapsed

In February 2022, Sama and OpenAI's relationship briefly deepened, only to falter. That month, Sama began pilot work for a separate project for OpenAI: collecting sexual and violent images—some of them illegal under U.S. law—to deliver to OpenAI. The work of labeling images appears to be unrelated to ChatGPT. In a statement, an OpenAI spokesperson did not specify the purpose of the images the company sought from Sama, but said labeling harmful images was "a necessary step" in making its AI tools safer. (OpenAI also builds image-generation technology.) In February, according to one billing document reviewed by TIME, Sama delivered OpenAI a sample batch of 1,400 images. Some of those images were categorized as "C4"—OpenAI's internal label denoting child sexual abuse—according to the document. Also included in the batch were "C3" images (including bestiality, rape, and sexual slavery,) and "V3" images depicting graphic detail of death, violence or serious physical injury, according to the billing document. OpenAI paid Sama a total of $787.50 for collecting the images, the document shows.

Within weeks, Sama had canceled all its work for OpenAI—eight months earlier than agreed in the contracts. The outsourcing company said in a statement that its agreement to collect images for OpenAI did not include any reference to illegal content, and it was only after the work had begun that OpenAI sent "additional instructions" referring to "some illegal categories." "The East

Africa team raised concerns to our executives right away. Sama immediately ended the image classification pilot and gave notice that we would cancel all remaining [projects] with OpenAI," a Sama spokesperson said. "The individuals working with the client did not vet the request through the proper channels. After a review of the situation, individuals were terminated and new sales vetting policies and guardrails were put in place."

In a statement, OpenAI confirmed that it had received 1,400 images from Sama that "included, but were not limited to, C4, C3, C2, V3, V2, and V1 images." In a followup statement, the company said: "We engaged Sama as part of our ongoing work to create safer AI systems and prevent harmful outputs. We never intended for any content in the C4 category to be collected. This content is not needed as an input to our pretraining filters and we instruct our employees to actively avoid it. As soon as Sama told us they had attempted to collect content in this category, we clarified that there had been a miscommunication and that we didn't want that content. And after realizing that there had been a miscommunication, we did not open or view the content in question — so we cannot confirm if it contained images in the C4 category."

Sama's decision to end its work with OpenAI meant Sama employees no longer had to deal with disturbing text and imagery, but it also had a big impact on their livelihoods. Sama workers say that in late February 2022 they were called into a meeting with members of the company's human resources team, where they were told the news. "We were told that they [Sama] didn't want to expose their employees to such [dangerous] content again," one Sama employee on the text-labeling projects said. "We replied that for us, it was a way to provide for our families." Most of the roughly three dozen workers were moved onto other lower-paying workstreams without the $70 explicit content bonus per month; others lost their jobs. Sama delivered its last batch of labeled data to OpenAI in March, eight months before the contract was due to end.

Because the contracts were canceled early, both OpenAI and Sama said the $200,000 they had previously agreed was not paid in full. OpenAI said the contracts were worth "about $150,000 over the course of the partnership."

Sama employees say they were given another reason for the cancellation of the contracts by their managers. On Feb. 14, TIME published a story titled *Inside Facebook's African Sweatshop*. The investigation detailed how Sama employed content moderators for Facebook, whose jobs involved viewing images and videos of executions, rape and child abuse for as little as $1.50 per hour. Four Sama employees said they were told the investigation prompted the company's decision to end its work for OpenAI. (Facebook says it requires its outsourcing partners to "provide industry-leading pay, benefits and support.")

Read More: *Inside Facebook's African Sweatshop*

Internal communications from after the Facebook story was published, reviewed by TIME, show Sama executives in San Francisco scrambling to deal with the PR fallout, including obliging one company, a subsidiary of Lufthansa, that wanted evidence of its business relationship with Sama scrubbed from the outsourcing firm's website. In a statement to TIME, Lufthansa confirmed that this occurred, and added that its subsidiary zeroG subsequently terminated its business with Sama. On Feb. 17, three days after TIME's investigation was published, Sama CEO Wendy Gonzalez sent a message to a group of senior executives via Slack: "We are going to be winding down the OpenAI work."

On Jan. 10 of this year, Sama went a step further, announcing it was canceling all the rest of its work with sensitive content. The firm said it would not renew its $3.9 million content moderation contract with Facebook, resulting in the loss of some 200 jobs in Nairobi. "After numerous discussions with our global team, Sama made the strategic decision to exit all [natural language processing] and content moderation work to focus on computer vision data annotation solutions," the company said in a statement. "We have spent the past year working with clients to transition those engagements, and the exit will be complete as of March 2023."

But the need for humans to label data for AI systems remains, at least for now. "They're impressive, but ChatGPT and other

generative models are not magic – they rely on massive supply chains of human labor and scraped data, much of which is unattributed and used without consent," Andrew Strait, an AI ethicist, recently wrote on Twitter. "These are serious, foundational problems that I do not see OpenAI addressing."

With reporting by Julia Zorthian/New York

MORE MUST-READS FROM TIME

- **Florence Pugh Might Just Save the Movie Star** From Extinction
- **Why You Can't Remember That Taylor Swift Concert** All Too Well
- What to Know About the **History of the Debt Ceiling**
- **10 Questions the *Succession* Finale Needs to Answer**
- **How Four Trans Teens Threw the Prom** of Their Dreams
- Why **Turkey's Longtime Leader** Is an Electoral Powerhouse
- **The Ancient Roots of Psychotherapy**
- Why **Rich People Aren't Using Phone Cases**

WRITE TO BILLY PERRIGO AT BILLY.PERRIGO@TIME.COM.