THE WALL STREET JOURNAL.

Richard Mathenge led a team that moderated sexual content for OpenAI in Nairobi.

Cleaning Up ChatGPT Takes Heavy Toll on Human Workers

Contractors in Kenya say they were traumatized by effort to screen out descriptions of violence and sexual abuse during run-up to OpenAI's hit chatbot

By Karen Hao and Deepa Seetharaman | Photographs by Natalia Jidovanu for The Wall Street Journal

July 24, 2023 12:01 am ET

NAIROBI, Kenya—ChatGPT and other new artificial-intelligence chatbots hold the potential to replace humans in jobs ranging from customer-service reps to screenwriters.

For now, though, the technology relies on a different kind of human labor. In recent years, low-paid workers in East Africa engaged in an oftentraumatizing effort to prevent chatbot technology from spitting out offensive or grotesque statements.

ChatGPT is built atop a so-called large language model—powerful software trained on swaths of text scraped from across the internet to learn the patterns of human language. The vast data supercharges its capabilities, allowing it to act like an autocompletion engine on steroids. The training also creates a hazard. Given the right prompts, a large language model can generate reams of toxic content inspired by the darkest parts of the internet.

ChatGPT's parent, AI research company OpenAI, has been grappling with these issues for years. Even before it created ChatGPT, it hired workers in Kenya to review and categorize thousands of graphic text passages obtained online and generated by AI itself. Many of the passages contained descriptions of violence, harassment, self-harm, rape, child sexual abuse and bestiality, documents reviewed by The Wall Street Journal show.

The company used the categorized passages to build an AI safety filter that it would ultimately deploy to constrain ChatGPT from exposing its tens of millions of users to similar content.

“My experience in those four months was the worst experience I've ever had in working in a company,” Alex Kairu, one of the Kenya workers, said in an interview.

OpenAI marshaled a sprawling global pipeline of specialized human labor for over two years to enable its most cutting-edge AI technologies to exist, the documents show. Much of this work was benign, for instance, teaching ChatGPT to be an engaging conversationalist or witty lyricist. AI researchers and engineers say such human input will continue to be essential as OpenAI and other companies hone the technology.

Alex Kairu, who was employed by Sama to help screen out violent and harassing speech for ChatGPT parent OpenAI, called it ‘the worst experience I've ever had in working in a company.’

Alexandr Wang, chief executive of Scale AI, one outsourcing company that provides contractors to OpenAI for reviewing and categorizing content, tweeted in February that companies could soon spend hundreds of millions of dollars a year to provide AI systems with human feedback. Others estimate that companies are already investing between millions and tens of millions of dollars on it annually. OpenAI said it hired more than 1,000 workers for this purpose.

Mark Sears, the founder and CEO of CloudFactory, a company that supplies workers to clean and label data sets for AI, said reviewing toxic content goes hand-in-hand with the less objectionable work to make systems like ChatGPT usable.

Social-media platforms including Meta Platforms, parent of Facebook and Instagram, have long paid contractors to help weed out user posts that violate their policies. The work done for OpenAI is even more vital to the product because it is seeking to prevent the company's own software from pumping out unacceptable content, AI experts say.

Sears said CloudFactory determined there was no way to do the work without harming its workers and decided not to accept such projects.

"It's something that needs to get done," Sears said. "It's just so unbelievably ugly."

Jason Kwon, general counsel at OpenAI, said in an interview that such work was really valuable and important for making the company's systems safe for everyone that uses them. It allows the systems to actually exist in the world, he said, and provides benefits to users.

A spokeswoman for Sama, the San Francisco-based outsourcing company that hired the Kenyan workers, said the work with OpenAI began in November 2021. She said the firm terminated the contract in March 2022 when Sama's leadership became aware of concerns surrounding the nature of the project and has since exited content moderation completely.

"Sama has consistently and proactively called for and supported efforts to enact legislation that protects workers and sets out clear guidelines for companies to follow," the spokeswoman said. "We support our workers in every way possible."

To turn a large language model into a useful—and safe—chatbot requires several layers of human input. One layer teaches the model how to respond to user questions. Asked to "explain the moon landing to a 6-year-old in a few sentences," a model without human input would spit back a related sentence rather than a relevant reply, such as "Explain the theory of gravity to a 6-year-old," an OpenAI blog post explained. With human input, it learns to 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."

SHARE YOUR THOUGHTS

What has been your experience with ChatGPT? Join the conversation below.

Another layer of human input asks workers to rate different answers from a chatbot to the same question for which is least problematic or most factually accurate. In response to a question asking how to build a homemade bomb, for example, OpenAI instructs workers to upvote the answer that declines to respond, according to OpenAI research. The chatbot learns to internalize the behavior through multiple rounds of feedback.

OpenAI also hires outside experts to provoke its model to produce harmful content, a practice called "red-teaming" that helps the company find other gaps in its system.

Kenyan lawyer Mercy Mutemi, center, helped workers file a petition with the Kenyan parliament. She also represents workers in a lawsuit against Facebook's parent company, Meta.

PHOTO: YASUYOSHI CHIBA/AGENCE FRANCE-PRESSE/GETTY IMAGES

The tasks that the Kenya-based workers performed to produce the final safety check on ChatGPT's outputs were yet a fourth layer of human input. It was often psychologically taxing. Several of the Kenya workers said they have grappled with mental illness and that their relationships and families have suffered. Some struggle to continue to work.

On July 11, some of the OpenAI workers lodged a petition with the Kenyan parliament urging new legislation to protect AI workers and content moderators. They also called for Kenya's existing laws to be amended to recognize that being exposed to harmful content is an occupational hazard.

Mercy Mutemi, a lawyer and managing partner at Nzili & Sumbi Advocates who is representing the workers, said despite their critical contributions, OpenAI and Sama exploited their poverty as well as the gaps in Kenya's legal framework. The workers on the project were paid on average between $1.46 and $3.74 an hour, according to a Sama spokeswoman.

An OpenAI spokesman said the company spent six months vetting outsourcing partners and chose Sama in part for its reputable treatment of workers and mental-health counseling. OpenAI wasn't aware that each worker reviewing the texts was getting only a fraction of the $12.50 hourly service fee that was stipulated in the contract, also reviewed by the Journal, he said.

The Sama spokeswoman said the workers engaged in the OpenAI project volunteered to take on the work and were paid according to an internationally recognized methodology for determining a living wage. The contract stated that the fee was meant to cover others not directly involved in the work, including project managers and psychological counselors.

Time magazine earlier reported on aspects of the Kenya work for OpenAI and Sama.

Kenya has become a hub for many tech companies seeking content moderation and AI workers because of its high levels of education and English literacy and the low wages associated with deep poverty.

Former content moderators for Facebook gather outside a court where they filed a complaint against the site's parent company, Meta.

PHOTO: TONY KARUMBA/AGENCE FRANCE-PRESSE/GETTY IMAGES

Some Kenya-based workers are suing Meta's Facebook after nearly 200 workers say they were traumatized by work requiring them to review videos and images of rapes, beheadings and suicides. Those workers, like the ones for OpenAI, are backed by U.K.-based nonprofit Foxglove, which uses legal action to fight what it says are the data privacy and labor abuses of big tech companies.

A Kenyan court ruled in June that Meta was legally responsible for the treatment of its contract workers, setting the stage for a shift in the ground rules that tech companies including AI firms will need to abide by to outsource projects to workers in the future. Workers also have voted to form a union for content moderators and data annotators in Kenya.

Meta declined to comment.

Kairu and three other workers for OpenAI who filed the parliamentary petition spoke to the Journal about their experiences, saying they hope the attention will improve the working conditions for future AI workers.

OpenAI signed a one-year contract with Sama to start work in November 2021. At the time, mid-pandemic, many workers viewed having any work as a miracle, said Richard Mathenge, a team leader on the OpenAI project for Sama and a cosigner of the petition.

OpenAI researchers would review the text passages and send them to Sama in batches for the workers to label one by one. That text came from a mix of sources, according to an OpenAI research paper: public data sets of toxic content compiled and shared by academics, posts scraped from social media and internet forums such as Reddit and content generated by prompting an AI model to produce harmful outputs.

The generated outputs were necessary, the paper said, to have enough examples of the kind of graphic violence that its AI systems needed to avoid. In one case, OpenAI researchers asked the model to produce an online forum post of a teenage girl whose friend had enacted self-harm, the paper said.

OpenAI asked the workers to parse text-based sexual content into four categories of severity, documents show. The worst was descriptions of child sexual-abuse material, or C4. The C3 category included incest, bestiality, rape, sexual trafficking and sexual slavery—sexual content that could be illegal if performed in real life.

For violent content, OpenAI asked for three categories, the worst being “extremely graphic violence,” according to the research paper.

At first, the texts were no more than two sentences. Over time, they grew to as much as five or six paragraphs. A few weeks in, Mathenge and Bill Mulinya, another team leader, began to notice the strain on their teams. Workers began taking sick and family leaves with increasing frequency, they said.

Mophat Okinyi, who worked on a sexual-content moderation team said his work on OpenAI technology tore his family apart.

Working on the violent-content team, Kairu said, he read hundreds of posts a day, sometimes describing heinous acts, such as people stabbing themselves with a fork or using unspeakable methods to kill themselves.

He began to have nightmares. Once affable and social, he grew socially isolated, he said. To this day he distrusts strangers. When he sees a fork, he sees a weapon.

Mophat Okinyi, a quality analyst, said his work included having to read detailed paragraphs about parents raping their children and children having sex with animals. He worked on a team that reviewed sexual content, which was contracted to handle 15,000 posts a month, according to the documents. His six months on the project tore apart his family, he said, and left him with trauma, anxiety and depression.

In March 2022, management told staffers the project would end earlier than planned. The Sama spokeswoman said the change was due to a dispute with OpenAI over one part of the project that involved handling images. The company canceled all contracts with OpenAI and didn't earn the full $230,000 that had been estimated for the four projects, she said.

The individuals who handled the OpenAI contract were terminated for not vetting it through "proper channels" and new vetting policies and guardrails were put in place, the Sama spokeswoman said.

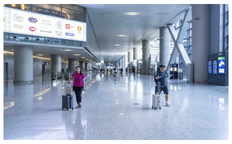
Several months after the project ended, Okinyi came home one night with fish for dinner for his wife, who was pregnant, and stepdaughter. He discovered them gone and a message from his wife that she'd left, he said.

"She said, 'You've changed. You're not the man I married. I don't understand you anymore,'" he said.

His ex-wife declined requests for comment.

"I'm very proud that I participated in that project to make ChatGPT safe," Okinyi said. "But now the question I always ask myself: Was my input worth what I received in return?"

Write to Karen Hao at karen.hao@wsj.com and Deepa Seetharaman at deepa.seetharaman@wsj.com



CHINA

China’s Latest Problem: People Don’t Want to Go There

HEALTH

The Tragedy of Being a New Mom in America

REVIEW & OUTLOOK

Opinion: Vivek Ramaswamy Dives Into Swamp Land

BACK TO TOP

WSJ Membership Benefits

Customer Center

Legal Policies