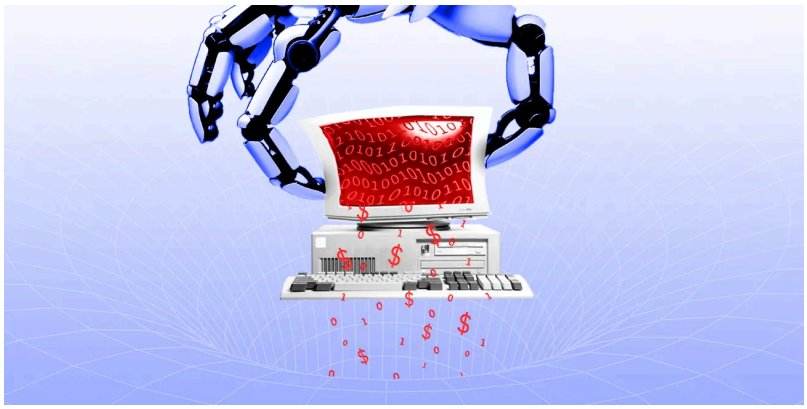
AI

Like digital locusts, OpenAI and Anthropic AI bots cause havoc and raise costs for websites

Darius Rafieyan Sep 19, 2024, 9:00 AM UTC Share Save



Getty Images; Alyssa Powell/BI

Edd Coates' Game UI Database was crippled by traffic from an OpenAI IP address.

AI companies are aggressively crawling the web, causing disruptions.

Website owners see cloud bills spike due to AI botnet traffic.

Edd Coates knew something was wrong. His online database was under attack.

Coates is a game designer and the creator of the Game UI Database. It's a labor of love for which he spent five years cataloging more than 56,000 screenshots of video game user interfaces. If you want to know what the health bar looks like in Fallout 3 and compare that to the inventory screen in Breath of the Wild, Coates has you covered.

A few weeks ago, he says, the website slowed to a crawl. It was taking 3 times as long to load pages, users were getting 502 Bad Gateway Errors, and the homepage was being reloaded 200 times a second.

"I assumed it was some sort of petty DDoS attack," Coates told Business Insider.

But when he checked the system logs, he realized the flood of traffic was coming from a single IP address owned by OpenAI.

In the race to build the world's most advanced AI, tech companies have fanned out across the web, releasing botnets like a plague of digital locusts to scour sites for anything they can use to fuel their voracious models.

It's often high quality training data they're after, but also other information that may help AI models understand the world. The race is on to collect as much information as possible before it runs out, or the rules change on what's acceptable.

One study estimated that the world's supply of usable AI training data could be depleted by 2032. The entire online corpus of recorded human experience may soon be inadequate to keep ChatGPT up to date.

A resource like the Game UI Database, where a human has already done the painstaking labor of cleaning and categorizing images, must have looked like an all-you-can-eat-buffet.

Bigger cloud bills

For small website owners with limited resources, the costs of playing host to swarm of hungry bots can present a significant burden.

"Within a space of 10 minutes we were transferring around 60 to 70 gigabytes of data," said Jay Peet, a fellow game designer who manages the servers that host the Coates' database. "Based on Amazon's on-demand bandwidth pricing that would cost $850 per day."

Coates makes no money from the Game UI Database and in fact operates the site at a loss, but he worries that the actions of giant AI companies may endanger independent creators who rely on their websites to make a living.

"The fact that OpenAI's behavior has crippled my website to the point where it stopped functioning is just the cherry on top," he said.

An OpenAI spokesperson said the company's bot was querying Coates' website roughly twice per second. The representative also stressed that OpenAI was crawling the site as part of an effort to understand the web's structure. It wasn't there to scrape data.

Related stories

Meta has 2 new sneaky bots scooping up free AI-training data from the web

It's getting harder to make big leaps at the frontier of AI. There will be huge winners and losers.

"We make it easy for web publishers to opt out of our ecosystem and express their preferences on how their sites and content work with our products," the spokesperson added. "We've also built systems to detect and moderate site load to be courteous and considerate web participants."

Planetary problems

Joshua Gross, founder of digital product studio Planetary, told BI that he encountered a similar problem after redesigning a website for one of his clients. Shortly after launch, traffic jumped and the client saw their cloud computing costs double from previous months.

"An audit of traffic logs revealed a significant amount of traffic from scraping bots," Gross said. "The problem was primarily Anthropic driving an overwhelming amount of nonsense traffic," he added, referring to repeated requests all resulting in 404 errors.

Jennifer Martinez, a spokesperson for Anthropic said the company strives to make sure its data-collection efforts are transparent and not intrusive or disruptive.

Eventually, Gross said, he was able to stem the deluge of traffic by updating the site's robots.txt code. Robots.txt is protocol, in use since the late 1990s, that lets bot crawlers know where they can and can't go. It is widely accepted as one of the unofficial rules of the web.

Blocking AI bots

Robots.txt restrictions aimed at AI companies have skyrocketed. One study found that between April 2023 and April 2024, nearly 5% of all online data and about 25% of the highest quality data added robots.txt restrictions for AI botnets.

The same study found that 25.9% of such restrictions were for OpenAI, compared to 13.3% for Anthropic, and 9.8% for Google. The authors also found that many data owners banned crawling in their Terms of Service, but did not have robots.txt restrictions in place. That has left them vulnerable to unwanted crawling from bots that rely solely on robots.txt.

OpenAI and Anthropic have said their bots respect robots.txt but BI has reported instances in the recent past in which both companies have bypassed the restrictions.

Key metrics polluted

David Senecal, a principal product architect for fraud and abuse at networking giant Akamai, says his firm tracks AI training botnets managed by Google, Microsoft, OpenAI, Anthropic, and others. He says among Akamai's users the bots are controversial.

"Website owners are generally fine with having their data indexed by web search engines like Googlebot or Bingbot," Senecal said, "however, some do not like the idea of their data being used to train a model."

He says some users complain about increased cloud costs or stability issues from the increased traffic. Others worry the botnets present intellectual property issues or will "pollute key metrics" like conversion rates.

When an AI bot is swarming your website over and over, your traffic metrics will likely be out of whack with reality. That's causes problems for sites that advertise online and need to track how effective this marketing is.

Related stories

Meta has 2 new sneaky bots scooping up free AI-training data from the web

It's getting harder to make big leaps at the frontier of AI. There will be huge winners and losers.

Senecal says robots.txt is still the best way to manage unwanted crawling and scraping, though it's an imperfect solution. It requires domain creators to know the specific names of every single bot they want to block, and it requires the bot operators to comply voluntarily. On top of that, Senecal says Akamai tracks various "impersonator" bots that parade as Anthropic or OpenAI web crawlers, making the task of parsing through them even harder.

In some cases, Senecal says, botnets will crawl an entire website every day just to see what's changed, a blunt approach that results in massive amounts of duplicated data.

"This way of collecting data is very wasteful," he said, "but until the mindset on data sharing changes and a more evolved and mature way to share data exists, scraping will remain the status quo."

"We are not Google"

Roberto Di Cosmo is the director of Software Heritage, a non-profit database created to "collect, preserve and share all publicly available source code for the benefit of society."

Di Cosmo says this past summer he saw an unprecedented surge in AI botnets scraping the online database, causing the website to become unresponsive for some users. His engineers spent hours identifying and blacklisting thousands of IP addresses that were driving the traffic, diverting resources away from other important tasks.

"We are not Google, we have a limited amount of resources to run this operation," Di Cosmo said.

He's an evangelist for open access, and not in theory opposed to AI companies using the database to train models. Software Heritage already has a partnership with Hugging Face, which used the database to help train its AI model StarCoder2.

"Developing machine-learning models that encompass these digital commons can democratize software creation, enabling a wider audience to benefit from the digital revolution, a goal that aligns with our values," Di Cosmo said, "but it must be done in a responsible way."

Software Heritage has published a set of principles governing how and when it agrees to share its data. All models created using the database must be open-source

and not "monopolized for private gain." And the creators of the underlying code must be able to opt out if they wish.

"Sometimes, these people get the data anyway," Di Cosmo said, referring to botnets that scrape hundreds of billions of web pages one by one.

Getting taken offline

"We have been taken offline a couple of times due to AI bots," said Tania Cohen, chief executive of 360Giving, a non-profit database of grants and charitable giving opportunities.

Cohen says that as a small charity with no in-house technical team, the surges in traffic have been highly disruptive. What's even more frustrating, she says, is that much of the information is easily downloadable in other ways and doesn't need to be crawled.

But hungry AI botnets scrape first, ask questions later.

"Utterly sick"

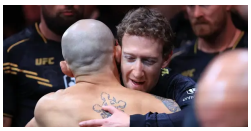
Coates says his Game UI Database is back up and running and he continues to add to it. There are millions of people out there like Coates, obsessive about some tiny corner of the world, compelled to sink thousands of hours into a pursuit that no one else on Earth could find meaning in. It's one of the reasons to love the internet.

And it's yet another area of society buffeted by the ripple effects of the AI revolution. The server costs of a small-fry database operator may seem not worth mentioning. But Coates' story is emblematic of a bigger question: When AI comes to change the world, who bears the cost?

Coates says he maintains the database as a source of reference material for fellow game designers. He worries that generative AI, which depends on the work of human creators, will inevitably replace those very same creators.

"To find that my work is not only being stolen by a large organization, but used to hurt the very people I'm trying to help, makes me feel utterly sick," Coates said.

Read next



AI

Meta has 2 new sneaky bots scooping up free AI-training data from the web

AI

It's getting harder to make big leaps at the frontier of AI. There will be huge winners and losers.

AI

Internal Amazon sales guidelines spread doubt about OpenAI capabilities while bad-mouthing Microsoft and Google

OpenAI | Artificial Intelligence | Cloud Computing | More...