FORBES INNOVATION

PREMIUM EDITORS' PICK

Hackers Have Uploaded Thousands Of Malicious Files To AI's Biggest Online Repository

Hugging Face has become the launching pad for large language models but its popularity has also proven a draw for cyber criminals.

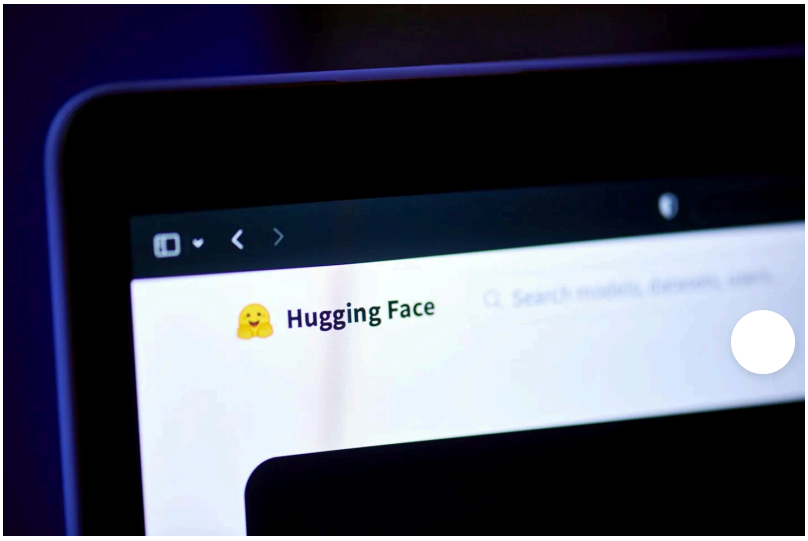
Iain Martin Forbes Staff

I'm a senior editor at Forbes and cover tech and venture capital.

Follow

Oct 22, 2024, 06:00am EDT

Updated Oct 22, 2024, 03:04pm EDT



Security researchers Protect AI have found tens of thousands "malicious" models on Hugging Face, the Github for artificial intelligence.

Hugging Face, the primary online repository for generative AI, has hosted thousands of files containing hidden code that can poison data and steal information, including the tokens used to pay AI and cloud operators, according to security researchers.

Researchers from security startups ProtectAI, Hiddenlayer and Wiz have warned for months that hackers have uploaded "malicious models" to Hugging Face's site, which now hosts more than a million models available for download.

"The old Trojan horse computer viruses that tried to sneak malicious code onto your system have evolved for the AI era," said Ian Swanson, Protect AI's CEO and founder. The Seattle, Washington-based startup found over 3,000 malicious files when it began scanning Hugging Face earlier this year.

Some of these bad actors are even setting up fake Hugging Face profiles to pose as Meta or other technology companies to lure downloads from the unwary, according to Swanson. A scan of Hugging Face uncovered a number of fake accounts posing as companies like Facebook, Visa, SpaceX and Swedish telecoms giant Ericsson.

One model, which falsely claimed to be from the genomics testing startup 23AndMe, had been downloaded thousands of times before it was spotted, Swanson said. He warned that when installed, the malicious code hidden in the fake 23AndMe model would silently hunt for AWS passwords, which could

be used to steal cloud computer resources. Hugging Face deleted the model after being alerted to the risk.

Hugging Face has now integrated ProtectAI's tool that scans for malicious code into its platform, showing users the results before they download anything.

The company told *Forbes* it has verified the profiles of big companies like OpenAI and Nvidia starting in 2022. In November 2021, it began scanning the files often used to train machine learning models on the platform for unsafe code. "We hope that our work and partnership with Protect AI, and hopefully many more, will help better trust machine learning artifacts to make sharing and adoption easier," said Julien Chaumond, CTO of Hugging Face in an email to *Forbes*.

The risk from malicious models has been substantial enough to warrant a joint warning from the United State's Cybersecurity and Infrastructure Security Agency and Canada and Britain's security agencies in April. The NSA and its British and Canadian counterparts cautioned businesses to scan any pre-trained models for dangerous code, and then only run them away from critical systems.

The hackers that have targeted Hugging Face typically inject rogue instructions into the code that developers download from the site, using it to hijack the model when it is run by an unsuspecting target. "These are classic attacks but they're just hidden within models," Swanson said. "Nobody would know that the model is doing these nefarious things and it would be incredibly hard for them to be able to trace that back."

Hugging Face was last valued at $4.5 billion when it raised $235 million in August 2023. The eight-year-old startup founded by Clément Delangue, Julien Chaumond and Thomas Wolf pivoted from running a teenage-focused chatbot app to a platform for machine learning in 2018. It's now raised $400 million to date and has been dubbed the Github for AI researchers.

"For a long time, AI was a researcher's field and the security practices were quite basic," said Chaumond. "As our popularity grows, so does the number of potentially bad actors who may want to target the AI community."

Update: Protect AI clarified that the number of malicious models it found was in the thousands, not tens of thousands.

MORE FROM FORBES

The $2 Billion Emoji: Hugging Face Wants To Be Launchpad For A Machine Learning Revolution

By

AI Startup Hugging Face Is Raising Fresh VC Funds At $4 Billion Valuation

By

AI Unicorn Hugging Face Acquires A Startup To Eventually Host Hundreds Of Millions Of Models

By

'Like Wikipedia And ChatGPT Had A Kid': Inside The Buzzy AI Startup Coming For Google's Lunch

By

Send me a secure tip.

Iain Martin

Follow

Iain Martin is a senior editor who covers tech, startups and venture capital out of London, England. Iain edits the Midas List Europe, Midas Seed and Forbes 30 Under 30 Europe... **Read More**

Editorial Standards Forbes Accolades