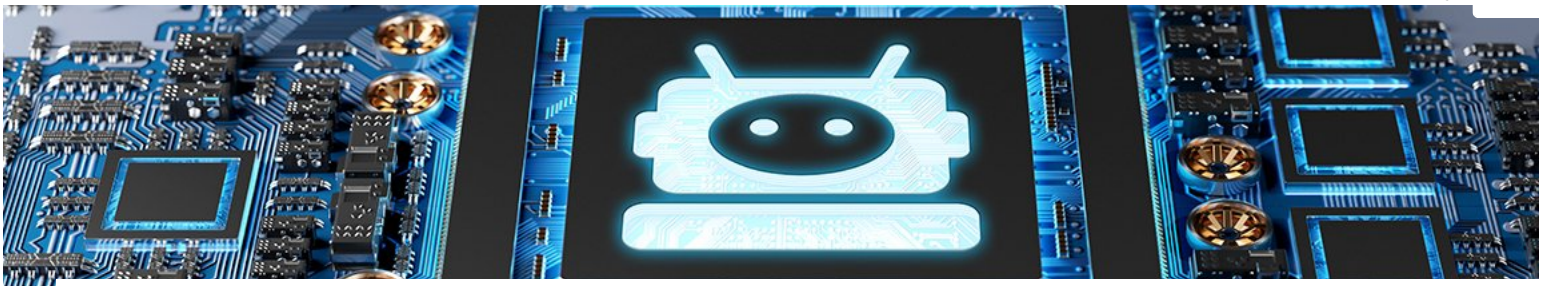


NEWS

How hackers can abuse ChatGPT to create malware

ChatGPT's capabilities for producing software code are limited. But researchers have observed cybercriminals bypassing the chatbot's safeguards to produce malicious content.

By **Alexis Zacharakos,** Student Co-op Published: **22 Feb 2023**

ChatGPT can't replace skilled threat actors -- at least not yet -- but security researchers say there is evidence that it has helped low-skill hackers create malware, raising concerns about cybercriminal abuse of the technology.

Since the launch of ChatGPT in November, the OpenAI chatbot has aided over 100 million users -- about 13 million individuals a day -- in writing text, music, poetry, stories and plays per request. It can also answer test questions and even write software code.

Malintent seems to accompany powerful technology, especially when it's available to the public. As experts predicted, there is evidence on the dark web that people have exploited ChatGPT for malicious content creation despite anti-abuse restrictions designed to prevent illicit requests.

Check Point Software Technologies earlier this month reported that cybercriminals had bypassed ChatGPT's safeguards to generate malware. Check Point researchers monitored dark web forums following the release of ChatGPT and found instances where threat actors discussed used the chatbot to "improve" malware code. The researchers tested the coding capabilities of the chatbot to see what help it could provide to hackers.

"Using ChatGPT and ChatGPT-3, we didn't see ability to create something really sophisticated,

per se," said Sergey Shykevich, threat intelligence group manager at Check Point. "It's mostly about the more people that will enter their circle of potential developer cybercriminals and develop tools, and it can improve the cost efficiency of sophisticated threats."

Though the chatbot lacks complexity in terms of coding, ChatGPT can help facilitate the process. The chatbot is a helpful platform in creating simple malware or improving malicious for those who need assistance, such as low-skill hackers and script kiddies.

"They can submit the chunks of code and verify that there are no bugs or vulnerabilities in the code of ransomware or to improve the code or write phishing emails as part of their infection chains," Shykevich said. "Someone who was never a developer and never created a code or malicious code now all of the sudden was able to experiment to try to create a malicious code."

Illicit use of ChatGPT

As early as December 2022, Reddit users have discussed "jailbreaks" or certain prompts that have been successful in overriding ChatGPT's defenses. Users share that they tricked the chatbot into making discriminatory or ridiculous statements.

Though poking holes into the chatbot's defenses were initially meant to showcase the chatbot's flaws, threat actors have made illicit use of its weaknesses. Check Point researchers observed three ways through which threat actors used OpenAI's API to generate malware.

On December 21, 2022, a threat actor with the username USDoD disclosed on an underground forum that he generated his first script using ChatGPT. The multi-layer encryption Python tool could be used to encrypt one's files, but it can also be used as a ransomware model.

Related Content

The impact of generative AI on the datacentre
– ComputerWeekly.com

Should we be worried about malicious use of AI ..
– ComputerWeekly.com

5 ChatGPT security risks in the enterprise
– Security

USDoD admitted that he had never created a script before, raising concern that the chatbot will allow those without technical skills to create malware and cause significant damage. Chris Anley, chief scientist at NCC Group, told TechTarget Editorial that while ChatGPT may not be able to produce full ransomware, it's easy for users to request an encryption script that could be used maliciously.

Though this code could be potentially harmful, Shykevich said that the decode USDoD derived from ChatGPT "did require some small adjustments to really make it work properly."

In an earlier report, Check Point discovered another hacker explaining that he used ChatGPT to recreate a python-based info stealer code a week later, which could be used to copy and exfiltrate files like PDFs, images and Office documents from a system. The same user then shared a simple Java code he wrote, which uses PowerShell to download a PuTTY software client, though it can be altered to download and operate any program.

Check Point researchers also uncovered a hacker demonstrating ChatGPT's role in the creation of a dark web marketplace. The hacker published code with a third-party API that incorporated current cryptocurrency prices that could be used to make illicit purchases.

Bypassing ChatGPT restrictions

At the outset, some security researchers viewed the restrictions in ChatGPT's user interface as weak and found threat actors could easily circumvent barriers. Since then, Shykevich said OpenAI has hunkered down to improve the chatbot's restraints.

"We see the restriction on the user interface of ChatGPT are every week much higher. So now it's more difficult to do malicious or abusive activity using ChatGPT," he said.

But cybercriminals can still abuse the program by using and avoiding certain words or phrases that allow users to bypass restrictions. Matt Lewis, commercial research director at NCC Group, Lewis referred to interacting with the online models as an "art form" that involves calculation.

"If you avoid use of the word malware and just ask it to show you the example of code that

encrypts files, by design, that's what it's going to do," said Lewis. "It has a way that it likes being instructed and there are interesting ways you can just get it to do what you want it to do in a number of different ways."

In a demonstration for TechTarget Editorial, Lewis showed how the program will "write an encryption script" that, despite falling short of full ransomware, could still be dangerous. "That's going to be a hard problem to solve," Lewis said of the bypasses, adding that policing language for context and intent will be extremely difficult for OpenAI.

Further complicating matters, Check Point researchers observed threat actors employing Telegram bots with APIs for a GPT-3 model, known as text-davinci-003, rather than ChatGPT, to override chatbot restrictions.

ChatGPT is simply a user interface for models of OpenAI. These models are available to developers to integrate back-end models with their own applications. Users consume these models via APIs, which are not guarded by restrictions.

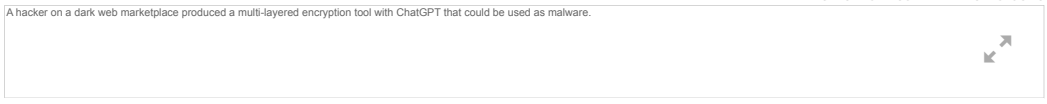
"From what we see, the barriers and restrictions that OpenAI put on the ChatGPT interface are not in place mostly for those who use the models via API," Shykevich said.

Threat actors can also dodge restrictions by being precise in their prompts to the chatbot. CyberArk has tested ChatGPT since its launch and detected blind spots in the restrictions. By repeatedly insisting and demanding, it will deliver the desired coding product. CyberArk researchers also reported that by continuously asking ChatGPT and rendering a new piece of code every time, users can create highly evasive polymorphic malware.

Polymorphic viruses can be extremely dangerous. But Anley noted there are already tools and frameworks available online to produce them. ChatGPT's capability to create one is most beneficial to unskilled coders and script kiddies.

"This is not a new capability in terms of attackers … nor is it a particularly effective way to generate variants of your malware," Anley said. "There are already tools that are better at it. It's potentially new in that it might allow less skilled attackers to generate code that might be dangerous."

CHECK POINT SOFTWARE TECHNOLOGIES



Check Point researchers discovered a threat actor published a multi-layered encryption tool created with ChatGPT.

How high are the stakes?

Because ChatGPT's outputs are often imperfect, they will most likely be harmful only if accompanied by the role of a skilled hacker that can make the necessary adjustments and additions to the code.

"It probably takes someone knowledgeable to turn that into a weaponized exploit, a weaponized

piece of malware," Anley said.

What makes the chatbot even less practical for malware creation is that there are already many tools available to create functional malware. Until we see advancement in the AI program, it will not be completely advantageous to utilize ChatGPT in this way.

"While it's true that you might be able to use the models in this way to generate some bad stuff, we can already generate malware of this type in a way that is probably more effective and safer from the attackers' point of view," Anley said.

However, ChatGPT has been effective for generating realistic phishing emails, which are often unconvincing when marked by spelling mistakes and incorrect grammar. But built with families of OpenAI's GPT-3 large language models, ChatGPT can produce cogent emails.

"The language is pretty good and pretty tight, and so you might not be able to distinguish it as much as you used to," said Lewis.

To address this issue, OpenAI will continue to crack down on ChatGPT's restrictions. But threat actors will most likely continue to find a way to abuse the technology. Shykevich warned that while discussions for chatbot moderation, restrictions and censorship are important, the safeguards will never be enough to completely end malicious activity.

"The abuse can become a monster if we don't discuss or the restrictions are not implemented from the beginning," Shykevich said. "But there is also no way that the abuse will be reduced to zero."

Next Steps

ChatGPT uses for cybersecurity continue to ramp up

Model collapse explained: How synthetic training data breaks AI

Dig Deeper on Security analytics and automation

ChatGPT

How to manage generative AI security risks in the enterprise

AI has a place in cyber, but needs effective evaluation

By: Amanda Hetler

By: Nihad Hassan