INSIGHTS
Cyber made simple.

one brightly cyber

AI HALLUCINATIONS MAY BE USED BY CYBERCRIMINALS

BY JAY BORDEN

AI hallucinations describe a situation where an AI system makes something up with no facts or input at all but presents it as accurate. We wrote about it in an earlier Insight.

People rarely check on the results or answers given by an AI system just accepting them as true. This is risky.

In a recent example attorneys used an AI system to find legal precedents relevant to their case. The AI system responded with some, and the attorneys used them in their legal case. All good except for one thing, the precedents never existed and were entirely made up by the AI system. That is a hallucination.

In another example, CNET, the technical site that is well-respected, had to print a retraction after it published financial advice that was extremely inaccurate. That advice was the result of an AI system hallucination.

Both of these, and other situations, are accidental in that the AI system hallucinated and the results were used without being checked. Fortunately, no serious damage was done.

With the use of AI systems growing daily it is inevitable that cybercriminals will find ways to exploit the tendency of AI systems to hallucinate.

And it has been found. Programmers are using AI to develop code, improve code, understand the logic, find errors, and automate tasks. A survey showed that developers believe AI adds to their work.

In a recent case researchers looking for hallucinated computer code libraries found chat bots referring to a specific Python package called huggingface-cli. But it doesn't exist.

AI HALLUCINATIONS MAY BE USED BY CYBERCRIMINALS

CONTINUED

A researcher decided to see what would happen with the chat bot reference and uploaded an empty package with that name to the library. Coders assumed the chat bot reference was trustworthy and within 90 days the package created by the researcher was downloaded over 35,000 times. This number has been verified and is not an AI hallucination.

A number of large companies recommended or used huggingface-cli in their code repositories. So, the hallucinations were widely accepted and spread.

The same researcher tested a few AI systems asking them to find hallucinated packages. All of the tested AI systems found hallucinated packages. Some of the AI systems hallucinated the packages 20% of the time while one hallucinated a package over 60% of the time. Not very comforting.

With these results you can see where cybercriminals can find or create a hallucinated package and upload malicious code to it. The uploaded code may be disguised by providing something useful, or not. But it will contain malware that will come free to any coder downloading the package. Then that package will be used by developers in creating code for the company.

Not all hallucinated packages pose equal risks. Some languages use protected prefixes or extensions. For instance, .net is a Microsoft controlled extension and has some controlled prefixes. Some languages don't have centralized libraries making it harder for developers to find and download a hallucinated package. .

It is unclear if any attacker has exploited this type of hallucination yet. If they did, it hasn't been discovered. It they haven't, they may well do so soon.

What to do? Educate your development team to check on the validity of libraries and packages before downloading or referencing. Be especially wary if they were found or recommended by an AI system.

Yes, we realize that this reduces the value of using AI systems. But the time spent checking will be far less than what will be required for removing all uses and references to it and then finding the malware it has installed.

To learn all the ways we can help make your company and family safer, visit onebrightlycyber.com, contact OneBrightlyCyber at info@onebrightlycyber.com, or call (888) 773-1920.