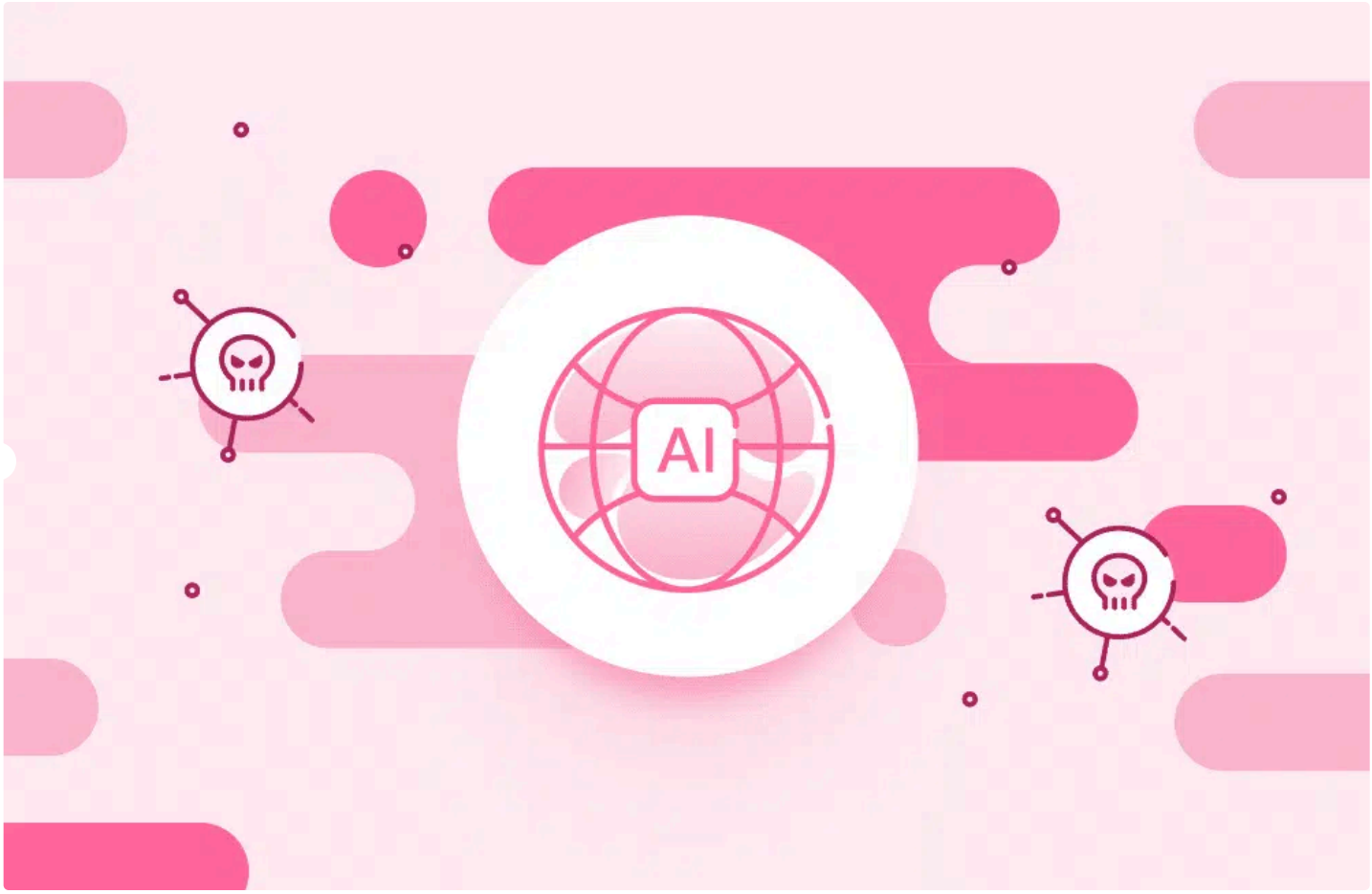
Hallucinated Packages, Malicious AI Models, and Insecure AI-Generated Code

AJ Starita • June 20, 2024 • 3 min read

Application Security

Share article



AI promises many advantages when it comes to application development. But it's also giving threat actors plenty of advantages, too. It's always important to remember that AI models can produce a lot of garbage that is ***really*** convincing—and so can attackers.

"Dark" AI models can be used to purposely write malicious code, but in this blog, we'll discuss three other distinct ways using AI models can lead to attacks. Whether you're building AI models or using AI models to build software, it's important to be cautious and stay alert for signs that things aren't quite what they seem.

Trippin' AIs lead to "hallucination squatting"

Large language machines (LLMs) don't think like humans—in fact, they don't think ***at all*** and they don't really "know" anything, either. They work by finding statistically ***plausible*** responses, not necessarily truthful ones. They are known to

occasionally hallucinate, which is a much cooler way of saying "spit out things that sound right but turn out to be complete BS". They've even been known to hallucinate ***sources*** of information to back themselves up.

One thing they might hallucinate is the name of an open source package. LLM hallucinations tend to have some persistence, so an enterprising threat actor might prompt an LLM to hallucinate a package name and then create a malicious package with the hallucinated name. From there, they only need to wait for the LLM-using developers to come. They might even remember to provide the functionality that the LLM suggests the package ought to have, making it that much more likely that their malicious code will make its way into the developer's environment.

Bad actors and bad AI models

Hugging Face does provide extensive security measures to users, including malware scanning, but many malicious ML models still make their way onto the platform. Bad actors sneaking malicious packages onto public software sharing platforms isn't anything new, but in the excitement of building with AI, developers may forget to be skeptical.

Loading malicious pickle files, which serialize and deserialize Python object structure and are common in AI models on Hugging Face, can easily lead to remote code execution attacks. Developers should be extremely cautious about downloading AI models and only use those from trustworthy sources.

Letting your Co-Pilot take the wheel

Code-producing LLMs have been known to include security vulnerabilities in their outputs, including ones common enough to make the OWASP Top 10. While engineers are working hard every day to make sure these models get better and better at their jobs, developers should still be wary of taking on any code that they don't fully understand. Unfortunately, the developers most likely to use the help from an AI are junior ones who are also less likely to scrutinize the code they've been given.

Takeaways

In all three of these examples, inexperienced developers are more likely to be duped by threat actors or LLMs. Of course, scanning your code and watching your network for signs of attack are a must, but nothing beats developer training here. Remind your teams frequently to be extremely skeptical of third-party code, whether it's generated by an LLM or not.

That said, we don't mean to be all doom and gloom; the world of AI is improving all the time, and AI developers are aware of these problems. Hallucinations are a bigger problem with LLMs that don't get updated frequently enough, so even Chat-GPT 3.5 has been getting updates, Hugging Face is working hard to remove malicious AI models, and Co-Pilot has been steadily improving the security of code output.

Even so, precaution is never a bad thing.