Six Steps Toward AI Security

Defending enterprise AI is a journey that begins with extending security practices already in place.

Author: David Reber Jr.

In the wake of ChatGPT, every company is trying to figure out its AI strategy, work that quickly raises the question: What about security?

Some may feel overwhelmed at the prospect of securing new technology. The good news is policies and practices in place today provide excellent starting points.

Indeed, the way forward lies in extending the existing foundations of enterprise and cloud security. It's a journey that can be summarized in six steps:

Expand analysis of the threats

Broaden response mechanisms

Secure the data supply chain

Use AI to scale efforts

Be transparent

Create continuous improvements

The first step is to get familiar with the new landscape.

Security now needs to cover the AI development lifecycle. This includes new attack surfaces like training data, models and the people and processes using them.

Extrapolate from the known types of threats to identify and anticipate emerging ones. For instance, an attacker might try to alter the behavior of an AI model by accessing data while it's training the model on a cloud service.

The security researchers and red teams who probed for vulnerabilities in the past will be great resources again. They'll need access to AI systems and data to identify and act on new threats as well as help building solid working relationships with data science staff.

Once a picture of the threats is clear, define ways to defend against them.

Monitor AI model performance closely. Assume it will drift, opening new attack surfaces, just as it can be assumed that traditional security defenses will be breached.

Also build on the PSIRT (product security incident response team) practices that should already be in place.

For example, NVIDIA released product security policies that encompass its AI portfolio. Several organizations — including the Open Worldwide Application Security Project — have released AI-tailored implementations of key security elements such as the common vulnerability enumeration method used to identify traditional IT threats.

Adapt and apply to AI models and workflows traditional defenses like:

Keeping network control and data planes separate

Removing any unsafe or personal identifying data

Using zero-trust security and authentication

Defining appropriate event logs, alerts and tests

Setting flow controls where appropriate

Protect the datasets used to train AI models. They're valuable and vulnerable.

Once again, enterprises can leverage existing practices. Create secure data supply chains, similar to those created to secure channels for software. It's important to establish access control for training data, just like other internal data is secured.

Some gaps may need to be filled. Today, security specialists know how to use hash files of applications to ensure no one has altered their code. That process may be challenging to scale for petabyte-sized datasets used for AI training.

The good news is researchers see the need, and they're working on tools to address it.

AI is not only a new attack area to defend, it's also a new and powerful security tool.

Machine learning models can detect subtle changes no human can see in mountains of network traffic. That makes AI an ideal technology to prevent many of the most widely used attacks, like identity theft, phishing, malware and ransomware.

NVIDIA Morpheus , a cybersecurity framework, can build AI applications that create, read and update digital fingerprints that scan for many kinds of threats. In addition, generative AI and Morpheus can enable new ways to detect spear phishing attempts .

Transparency is a key component of any security strategy. Let customers know about any new AI security policies and practices that have been put in place.

For example, NVIDIA publishes details about the AI models in NGC , its hub for accelerated software. Called model cards , they act like truth-in-lending statements, describing AIs, the data they were trained on and any constraints for their use.

NVIDIA uses an expanded set of fields in its model cards, so users are clear about the history and limits of a neural network before putting it into production. That helps advance security, establish trust and ensure models are robust.

These six steps are just the start of a journey. Processes and policies like these need to evolve.

The emerging practice of confidential computing , for instance, is extending security across cloud services where AI models are often trained and run in production.

The industry is already beginning to see basic versions of code scanners for AI models. They're a sign of what's to come. Teams need to keep an eye on the horizon for best practices and tools as they arrive.

Along the way, the community needs to share what it learns. An excellent example of that occurred at the recent Generative Red Team Challenge .

In the end, it's about creating a collective defense. We're all making this journey to AI security together, one step at a time.

Original URL: https://blogs.nvidia.com/blog/2023/09/25/ai-security-steps/