White paper
Cisco public

Securing AI with Cisco AI Defense

Contents

Over the last few years, new Artificial Intelligence (AI) technologies, such as Generative AI (GenAI), have caught the public's attention. Enterprises and their employees are no exception. As a result, AI is being incorporated into numerous enterprise applications. Simultaneously, enterprise employees' usage of third-party AI tools and applications is proliferating. Enterprises are already at a point where many of their applications are AI-enabled, and most employees use third-party AI tools at least some of the time.

Unfortunately, incorporating AI into enterprise applications has increased their attack surface. AI incorporation has also led to unintended consequences, as AI models are probabilistic in nature. Further, many employees' use of third-party AI tools is unsanctioned and inadequately protected.

As a result, security teams face two big questions: First, how will they secure the development and deployment of AI-infused enterprise applications? Second, how will they secure employees using third-party AI tools?

Clearly, these security teams need a comprehensive AI security solution that covers all uses of AI in the enterprise. Cisco AI Defense is precisely such a solution. As we will see below, it addresses both AI-infused enterprise applications and employees' usage of third-party AI tools.

Applications' AI Attack Surface

Enterprise applications are incorporating AI models, including Large Language Models (LLMs), to enhance the functionality offered to users. This AI infusion adds incremental attack surface(s) to enterprise applications.

Attackers can use the newly available attack surface to harm enterprises by inducing loss-making transactions or reverse-engineering proprietary information from the application's behavior. Further, some attacks may exploit the probabilistic nature of AI models, resulting in the enterprise applications producing incorrect or inappropriate information. If left un-mitigated, the information can cause reputational damage to the enterprise.

Such concerns are not just theoretical but have already happened. For example, as per a news report, an international airline was forced to compensate a customer misled by its AI-based customer service application about airfare policies.[1] Instances of AI applications providing consumers with dangerous food recipes have also been publicly documented.[2] [3]

As new AI technology is absorbed into enterprise applications, the AI attack surface changes, sometimes rapidly and at other times unpredictably, making security teams' jobs even more challenging.

[1] What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case, Forbes, February 2024.

[2] Supermarket AI Gives Horrifying Recipes For Poison Sandwiches And Deadly Chlorine Gas, Forbes, August 2023.

[3] Google's AI Recommended Adding Glue To Pizza And Other Misinformation–What Caused The Viral Blunders?, Forbes, May 2024

Shadow AI

Third-party AI tools and AI-powered applications continue to grow in popularity, including among most enterprises' employees. Their use can lead to productivity gains and innovation. However, much of the use of third-party AI tools falls into the category of shadow AI, which we define as employees' use of third-party AI tools and applications without the oversight of an enterprise's security team.

As a result, using these AI tools can lead to the inadvertent leakage of proprietary information and the unauthorized or inappropriate use of AI-generated artifacts within the enterprise.

The danger is real, as demonstrated in a recent news report on employees of a global electronics manufacturer. The employees inadvertently leaked proprietary source code to a third-party AI tool as they attempted to debug and optimize their code using the tool.[4]

Attempts to block all AI or even just GenAI applications have proven unrealistic, as applications increasingly incorporate AI in some form or another. Such drastic action is also counterproductive, as employees are prevented from using helpful technology.

Requirements to secure AI

A security solution that addresses the AI attack surface and shadow AI challenges described above needs to meet a few fundamental requirements:

1. The solution should enable the discovery of all AI user access, workloads, applications, models, and data across distributed cloud environments.
2. It should detect misconfigurations, security vulnerabilities, and adversarial attacks that put AI- enabled applications at risk.
3. It should protect AI-enabled applications against rapidly evolving threats such as prompt injections, denial of service, and data leakage guided by policies (also known as guardrails) configured by a security team. The solution should also protect employee access to third-party AI tools.

Further, such a system should protect applications throughout their development and production lifecycle. The protection should be automatic and continuous, enabling application developers to innovate rapidly.

Finally, a capable system should be deployable widely enough to cover all employees of an enterprise.

[4] Whoops, Samsung workers accidentally leaked trade secrets via ChatGPT, Mashable, April 2023.

AI Defense to the rescue

AI Defense is a new Cisco product specifically designed to address applications' AI attack surface and shadow AI. AI Defense meets all three requirements – discovery, detection, and protection – discussed above.

AI Defense has five conceptual components (see Figure 1):

1. A visibility engine that parses Domain Name System (DNS) and cloud logs to discover the usage of AI – both external and internal – in the enterprise network.
2. A suite of security models. These are proprietary LLMs, Small Language Models (SLMs), and Machine Learning (ML) models that have been carefully trained to detect inappropriate input to and output from other AI models and tools. Each model in the suite has a specific purpose and sports its individual performance – latency, throughput, efficacy – characteristics. The models are frequently retrained and updated to keep up with the changing nature of the AI threat landscape. One or more of the models are chosen for traffic inspection based on the policy configured by the security team.
3. A set of traffic interceptors that funnel traffic to and from AI models – whether external to the enterprise or internal – to the suite of security models. One of these interceptors is embedded inside the enterprise's cloud networking infrastructure, enabling it to intercept traffic between enterprise applications and AI models. A second one is embedded in the enterprise's edge networking infrastructure, allowing it to intercept traffic between employees and third-party AI tools. Additional interceptors enable Application Programming Interface (API) based intercept of AI traffic to cater to situations where network interception is not possible.
4. A validation service that algorithmically generates AI tests. These tests are automatically run against AI models, and the output is passed to the security models. Thereafter, a comprehensive report is produced that flags vulnerabilities in the AI models scanned. The report is intended for application developers so that they can adjust their choice of AI models and understand the risk posed by individual models while the application is still in the development (pre-production) phase. The validation service also recommends policies to mitigate the vulnerabilities discovered. The security team can deploy the recommended policies to protect the application in production. Note that the Cisco threat research team continuously updates the validation service to keep pace with changes in the AI threat landscape.
5. A management console using which the security team can configure the rest of the components. In particular, the security team can configure policies to apply to various slices of traffic flowing through the interceptors. Note that depending on the policy configured for a traffic slice, an appropriate subset of the security models is spun up and deployed to inspect the slice. The management console also provides alerts, reports, and a comprehensive dashboard for the other components of AI Defense.

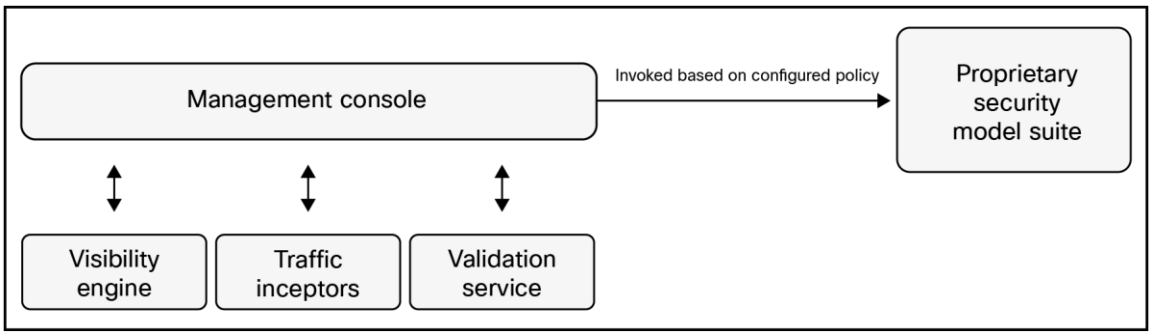

Figure 1.
Conceptual components of AI Defense

Orchestrating these five components into a production-level service with strict latency, throughput, and efficacy constraints is challenging. Fortunately, Cisco's technical team has been able to craft the components to support large, mission-critical deployments.

AI Defense in action

In the following paragraphs, we describe specific situations inspired by customer feedback that demonstrate the use of AI Defense in enterprise networks.

Validating in-development applications- In this scenario, application developers have created a new chat-based customer service application that uses an AI model hosted at the enterprise.

The AI model provides valuable functionality to the customer service application using past customer interactions but is known to sometimes produce output referencing the enterprise's proprietary information. If not managed, these references pose a data leakage risk. The developers want to comprehensively understand the situations in which the AI model's output is unsatisfactory before the application goes live.

The security team automatically sees that a new AI model and related assets have been added to the enterprise's environment. The team configures the validation service, interceptor, and the appropriate security policy via the management console to initiate a scan. Subsequently, automated scans are run against the application to look for new vulnerabilities.

The management console produces a report showing a small number of instances of proprietary data in response to the scans. The security team forwards the report to the application developers for analysis.

The developers consider modifying their application to prevent the leakage of proprietary data. However, given the probabilistic nature of the AI model and the cost of refining and retesting the application,

they conclude that installing additional policies that detect and strip out proprietary information is more practical.

The security team installs the appropriate policies before the application goes live.

Protecting live AI-infused applications- In this scenario, the previously mentioned customer service application has gone live. Customers have been using it, and initial reports of their experience are positive.

Meanwhile, the developers are working on the next version of the application. As they do so, they realize that some of the training data used for the AI model includes references to other vendors that provide adjacent products. Many of the references are unflattering towards the vendors on account of integration challenges that customers have encountered.

The developers are concerned that specific customer interactions with the enterprise application may lead to the application naming the vendors as the cause of integration problems. This poses a reputational and legal risk to the enterprise.

Since the application is already live and active, taking it down is unappetizing. Similarly, retraining the AI model is expensive and unlikely to be approved for several months.

The developers confer with the security team, and they install additional policies that enable security model inspection of responses from the AI model (see Figure 2).

The security models deployed screen and remove discussion of other vendors, temporarily eliminating the perceived reputational and legal risk.

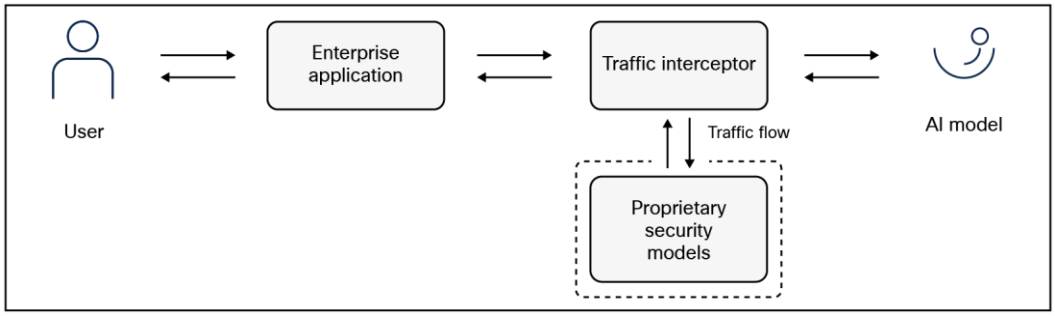

Figure 2.
Traffic flow intercept between the enterprise application and AI model

Safeguarding employees' usage of third-party AI tools- In this scenario, the enterprise's Chief Information Security Officer (CISO) has started a project to reign in shadow AI to reduce proprietary data leakage and legal risk.

As a first step, the security team uses AI Defense's visibility engine to discover the use of third-party AI tools. They find marketing and sales employees using a specific AI tool for competitive analysis.

Some employees download text snippets and documents to their work computers as part of their workflow. Unbeknownst to the employees, some of this information is confidential to competitors, and downloading it to a computer belonging to the enterprise poses a legal risk. The tainted competitor- proprietary information has inadvertently made it into the training data of the third-party AI tool.

Regardless, its presence on the enterprise's infrastructure is unwelcome.

The security team decides to address the shadow AI usage. The team uses the management console to configure interceptors such that user traffic to and from the AI tool is passed through the security models.

Further, the security team configures security policies to disallow proprietary information belonging to other corporations from being downloaded via AI tools. The security policies, in turn, orchestrate the spin-up of the appropriate security models.

The security models are nuanced in interpreting information and block only a tiny fraction of the information retrieved by the AI tool. The information blocked is deemed competitor-proprietary and, therefore, inappropriate for enterprise use.

Consequently, the marketing and sales employees complete their competitive analysis without putting the enterprise at risk.

Conclusion

Cisco has been developing security technologies for over thirty years and networking technologies for over forty. In recent years, it has invested heavily in using AI in traditional security products. With AI Defense, Cisco is expanding its protections to include security for AI: hardening and protecting AI usage, tools, and infrastructure.

As discussed in the sections above, AI Defense leverages Cisco's networking footprint to obtain visibility into AI assets and traffic flows and enable the enforcement of AI-specific security policies to protect the enterprise. The result is an end-to-end AI security solution that provides maximum visibility and enforcement.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-5054486-00 03/25