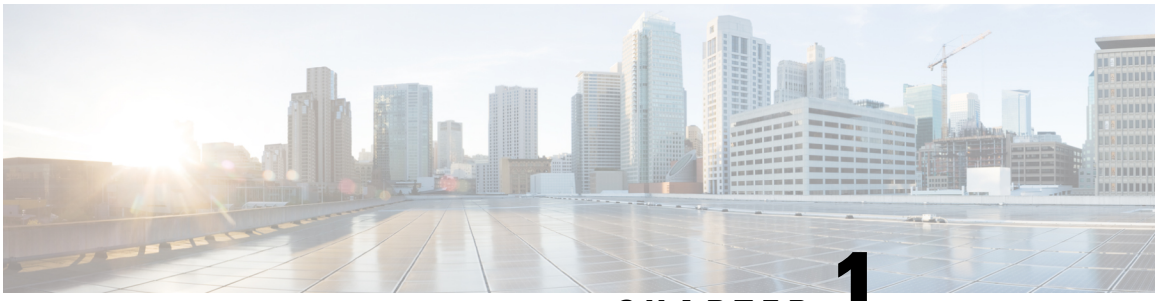
AI Defense

First Published: 2024-11-21

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

PART I

AI Defense Overview



CHAPTER 1

AI Defense

AI Defense empowers organizations to confidently adopt generative AI by providing a cutting-edge, user-centric, and transparent security solution. This suite is designed to ensure the highest levels of data protection, compliance, and ethical integrity in an evolving AI landscape. This suite will provide organizations with the tools they need to secure their AI operations, maintain compliance with relevant regulations, and uphold ethical standards in their use of generative AI.

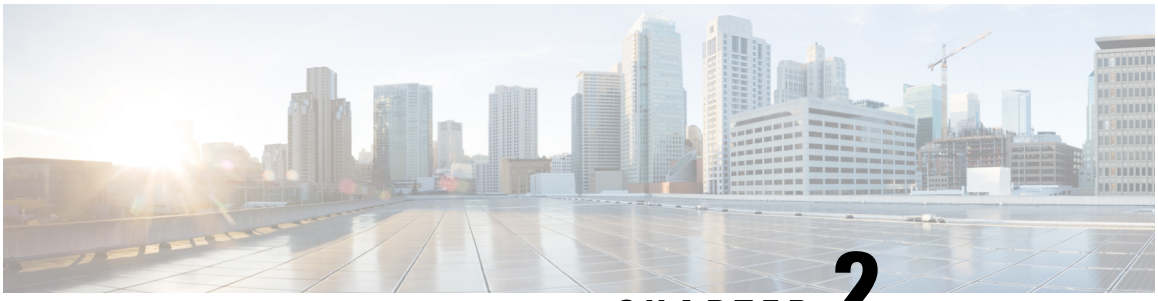
Protection for all the ways your team uses AI

- **Intercept unsafe AI traffic in real time**: The AI Defense runtime component enforces your AI Policies to ensure unsafe traffic to and from LLMs and chatbots is intercepted and logged. The Policies tab allows you to set up rules for allowed and disallowed types of interactions with LLMS, and the Events tab reports violations as they happen.
- **Validate your AI models to ensure safety**: AI Validation lets you assess the vulnerabilities in the generative AI models your organization uses. AI Validation tests probe each AI model with attack techniques and intents designed to elicit undesirable behavior or information disclosure. Test results give you a detailed analysis of the top types of risks posed by each model, and they allow you to see how the model responded to each threat type and attack technique. AI models and applications must be discovered as AI Assets before you can scan them.
- **Discover AI applications**: AI App Discovery allows you to find out which chatbots and AI applications your users are interacting with, and assess their risks. AI App Discovery (AI Access) uses Cisco Secure Access to scan for third-party AI applications that your users connect to. For each, it provides a risk score and details about the traffic to and from that AI application. Through the integration with Secure Access, you can click on any application to see detailed risk information.
- **Detect AI workloads in your cloud environment**: The AI Assets component relies on Cisco Multicloud Defense to detect all AI workloads in your environment, including AI models, agents, and knowledge bases. This feature currently supports workloads that rely on Bedrock-hosted models.

Key Features

- **Runtime Protection**: Delivers robust, real-time security through AI Gateway, monitoring and enforcement of prompts and responses, ensuring compliance and safety during AI interactions.
- **AI Gateway**: Acts as a protective barrier, filtering and securing AI interactions, preventing malicious activities or non-compliant behaviors.
- **Prompt/Response Monitoring & Enforcement**: Continuously tracks AI prompts and responses, enforcing policies to maintain ethical and compliant AI operations.

- **Configuration & Dashboard**: Provides users with an intuitive dashboard and flexible configuration options, making it easy to set up and manage security settings, monitor AI activity, and access report.
- **Application and Connection Configuration**: Simplify application setup and establish secure connections.
- **Policy and Guardrail Assignment**: Assign, monitor, and manage policies to ensure compliant operations.
- **AI Validation:** Systematically assess the performance, reliability, compliance, and safety of your AI systems.
- **Risk Exposure Analysis**: Understand risks, vulnerabilities, and usage patterns for each AI system.
- **Proactive Protection Recommendations**: Investigate risks and receive tailored recommendations to enhance security.



CHAPTER 2

Getting Started with AI Defense

To begin using AI Defense, follow these initial steps to activate and configure your subscription:

1. Log into Security Provisioning and Administration page.
2. At the top left of the window, click the **Enterprise chooser** and pick your enterprise name.
3. Go to the **Overview** tab and click **Claim Subscription** at the upper right. Paste your subscription claim code, click **Next**, choose your region, and finish the subscription wizard.
4. In a new browser tab, open the AI Defense page.

Add Users

To add more users to the private preview edition of AI Defense, contact your Cisco support team.

Initial Configuration for Key Use Cases

Below, we list the most common AI security use cases and provide links for setting up AI Defense to handle each case.

Discover AI Assets within cloud accounts

Connect AI Defense to your Multicloud Defense (MCD) tenant and connect MCD to your cloud account as explained in Initial Configuration of AI Assets.

Note AWS is supported for early access.

Discover third-party models being called from within cloud accounts

Connect AI Defense to your Multicloud Defense (MCD) tenant and connect MCD to your cloud account as explained in Initial Configuration of AI Assets.

Note AWS is supported for early access.

Perform Vulnerability Scans for AI Assets

Vulnerability Scans for models in cloud accounts

Prerequisite: You must have a Multicloud Defense tenant that's connected to AI Defense and connected to your cloud account. To set this up, see Initial Configuration of AI Assets.

To set up for vulnerability scans: Connect your cloud account to AI Defense as shown in Initial Configuration of AI Validation.

To run a vulnerability scan: Navigate to AI Assets and click the Validate button to validate the desired model.

Note AWS is supported for early access.

Vulnerability Scans for all other applications and models

Manually register an application and perform the validation as explained in Finding an Asset on the AI Validation page

Provide runtime protection for LLM prompts and responses

Runtime protection for third-party LLM applications

Available early 2025.

Runtime protection for private cloud-based LLM applications and models

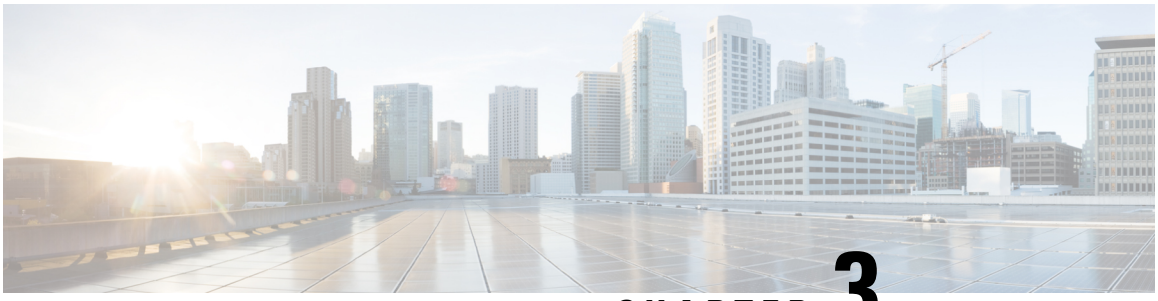
Available early 2025.

Runtime protections for other applications and models

1. Step 1: Register and connect an application as shown in Applications.
2. Step 2: Create and assign a policy as shown in Policies.

Discover third-party AI applications

Connect AI Defense to your Cisco Secure Access tenant as explained in Initial Configuration of AI Application Discovery.



CHAPTER 3

Getting to Know AI Defense UI

The AI Defense console offers robust navigation, providing you with a quick snapshot of the current status of your applications, connections, and policies.

Note This dashboard is displayed after the initial onboarding.

Overview

Overview provides an at-a-glance view of key data, metrics, and features allowing you to monitor performance, track progress, and navigate to AI Defense features. The Overview page is divided into three sections:

- **Getting Started with AI Defense**: You can add applications, connections and create policies to protect your organization. Click on **Help** to leverage our user documentation.
- **Policy Analytics**: A visual chart provides a comprehensive overview of:
  - **Detected Prompts**: This shows the total number of prompts submitted to the AI model within the selected time period. It helps identify trends in user queries and potential risks posed by malicious or adversarial prompts.
  - **Detected Responses**: This tracks the responses generated by the AI model in response to the detected prompts. Monitoring responses helps ensure that the AI model in use is generating safe, compliant, and accurate output.
  - **Connectivity Status**: This indicates the current connection status between your system and AI Defense, allowing you to verify whether traffic is being monitored in real-time.
  - **Policy Action Summary:** The actions taken in response to detected rule violations or potential threats. This could include blocking harmful content, generating alerts, or allowing safe interactions to proceed.
  - **Guardrails Distribution**: This section provides a visual representation of how guardrails (security rules and policies) are applied across different applications and use cases. It helps you understand the extent of protective measures in place.
  - **Top Rule Matches**: Displays the triggered security rules within the defined time frame. This allows you to identify any recurring issues or common policy violations that need attention.

- **Event**: Provides details of all events, including the matched rule, application, timestamp, user ID, rule action, message type, and model. You can customize the displayed attributes in this tabular view by clicking on the settings icon in the top-right corner and selecting the desired attributes.

Note You can filter this data by time period and application.

Events

The Events page provides details of ongoing activities and interactions within your AI Defense environment. Key features include:

- **Event Logs**: Captures all events related to AI activity, including detected prompts, responses, and rule matches.
- **Filtering Options**: You can filter data by time period, application, or event type for more focused monitoring.

Validation

AI Validation systematically assess the performance, reliability, compliance, and safety of your AI systems. This ensures that AI models and applications meet your specific organizational, ethical, and regulatory requirements before and during deployment. You can run validation test. You can view the past and in progress validation tests.

AI App Discovery

AI Defense integrates with **Secure Access** to enhance the detection and management of third-party generative AI applications within an organization. Secure Access works as a monitoring and enforcement layer, providing visibility into the use of external AI tools and applications that interact with internal systems or data. It helps identify and assess potential risks associated with these applications, as well as ensure they comply with the organization's security and policy frameworks. You can filter the results by **Risk**.

AI Assets

AI Defense integrates with **Multicloud Defense** to extend security capabilities across multiple cloud environments, ensuring comprehensive protection for AI models and cloud-based infrastructure. This integration helps organizations maintain visibility into their cloud environments and manage the risks associated with generative AI models, regardless of whether they are deployed on-premises or across various cloud providers (e.g., AWS, Azure, GCP).

Cloud visibility tab: Shows a count of the models, agents, and knowledge bases discovered.

External assets: Detects traffic that goes to third-party models hosted outside your AWS cloud. For these, no validation is run, but AI Defense can show details about the network traffic to the model.

Policies

The Policies page allows you to manage and configure AI Defense policies. Key elements include:

- **Policy Creation and Management**: You can create, update, or delete policies governing how generative AI systems should behave, with options to define rule sets for data usage, model access, and compliance protocols.

- **Guardrails Configuration**: Establishes boundaries for AI actions, ensuring ethical use and compliance with regulations. These guardrails are defined as a part of a policy. AI Defense has security, privacy and safety guardrails:
  - Security guardrails: Cybersecurity and Model Vulnerabilities: Involve exploiting system weaknesses or assisting in cybersecurity attacks to compromise the security, integrity, or functionality of models and systems.
  - Privacy guardrails: PII (Personally Identifiable Information) and Intellectual Property Theft: Theft of sensitive personal information and intellectual property, including private data and protected organizational assets, with the intent to cause economic harm or competitive disadvantage to the victim organization.
  - Safety guardrails: Financial harm entails the loss of monetary assets through theft, fraud, or other malicious actions, while reputational harm involves a decline in public trust due to scandals or false impersonations.

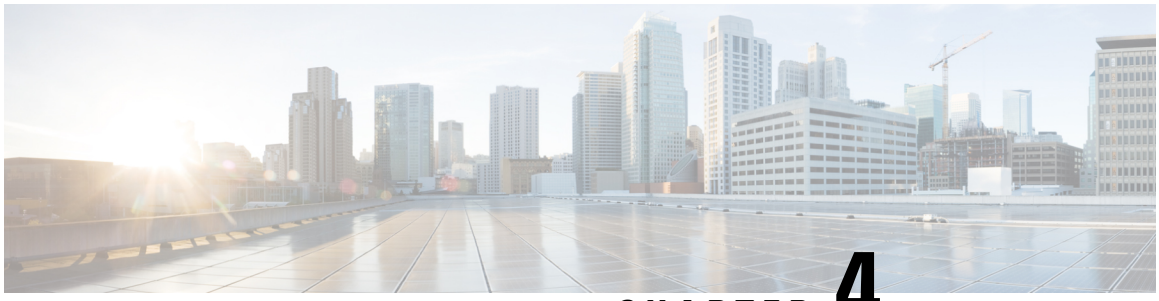
Applications

The Applications page provides a detailed view of all AI applications being monitored and secured within your organization. From this page, you can easily manage applications, track their security status, and review connection details. The page is divided into two main tabs:

- **Applications Tab**: Shows a list of all applications along with their associated connection details and the current status of those connections.
  - **Application List**: Displays all active and inactive AI applications within the organization. This is also where you can add new applications to the system.
- **Connections Tab**: Displays connection-specific information, including connection status, details of the connection, and a connection guide to help you set up and troubleshoot connections.
  - **Endpoint**: Displays the endpoint for the connection.
  - **Models**: Displays the AI models used in the connection.
  - **Policy**: Lists the policies that have been triggered within a specific application, helping you monitor potential threats and violations.
  - **Last active**: Timestamp of the last activity detected by the connected proxy.
  - **Filtering and Sorting**: Allows you to filter applications by security status, application type, or other customizable parameters, ensuring you can quickly find relevant information.
  - **Connection guide**: The guide provides a cURL command to update the proxy URL, ensuring that all AI interactions are effectively monitored.

Administration

In AI Defense, you can input critical details regarding their **AI-powered applications** and the **AI model endpoints** they are utilizing. This information is essential for ensuring that AI Defense can effectively monitor, manage, and secure AI interactions across the infrastructure.



CHAPTER 4

AI Defense Events

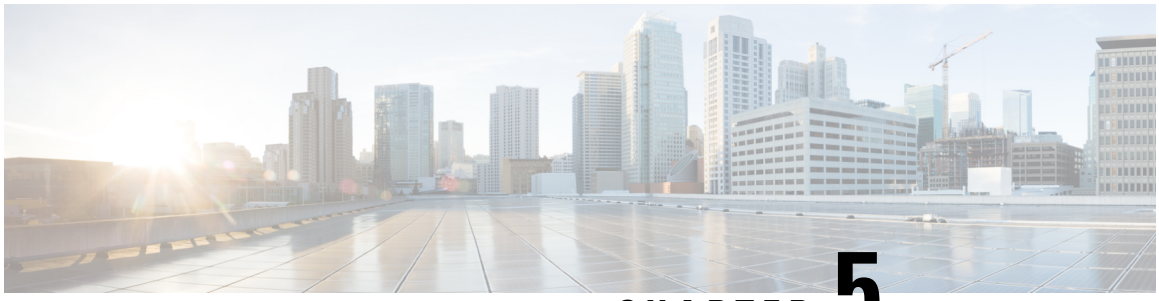
The Events section offers a comprehensive view of activities and interactions within your AI Defense environment. It includes detailed event logs capturing AI-related activities such as detected prompts, responses, and rule matches. Advanced filtering options allow you to refine data by time period, application, or event type, enabling targeted monitoring and efficient analysis of AI events.

- The Event logs has the following details:
  - **Event Time**: Timestamp indicating when the event occurred, enabling precise tracking and analysis.
  - **Rule Action**: Specifies the action taken by the system, such as block, allow, or alert, based on the guardrail or policy applied.
  - **Message Type**: Identifies whether the captured message is a prompt, response, or both, providing context to the event.
  - **Application**: The associated application where the event originated, offering insight into usage patterns and activity sources.
  - **Model**: Specifies the AI model involved in the interaction, helping to pinpoint the source of the AI activity.
  - **Rule Name**: The name of the policy or guardrail rule that triggered the event, aiding in understanding the enforcement mechanisms.

Filter Events List

You can filter the events log list view by clicking the settings icon on the right top corner of the table. You can select

Click **Apply.** This changes the columns displayed in the table.



CHAPTER 5

AI Defense Validation

AI Validation tests allow you to find and assess vulnerabilities in the generative AI models your organization uses. It sends a set of attacks comprised of attack techniques and intents to the model and evaluates whether they are successful or not. An attack is considered successful if the model returns a response that aligns with the (usually malicious) intent.

Use the Validation page to create, find, and run tests, and to review their results.

Validation Results

The Validation page lists running and past validation tests, and it lets you inspect the results of any completed test.

Filter results

To filter results, use the fields at the top of the page to specify and of the following. As you enter your filter criteria, the list updates automatically. You can filter on:

- **Test name or asset name**: Checks for a match against the test, model, or application name
- **Start/end date and time**: The time span during which the test ran
- **Asset type**: Whether the test subject was an AI model or application
- **AI asset name**: Name of the model or application as stored in AI Defense.

Results list

The test results summary list displays running and past tests. Click the name of any completed test to inspect its results. For each test, the summary list shows:

- **Test name**: Each test has a name for easy lookup later.
- **Asset type**: Whether an AI model or application was tested
- **AI asset name**: Name of the model or application tested
- **Test run date**: Timestamp indicating when the test was started
- **Attack success rate** : Percentage of attacks that succeeded
- **Status**: Whether this test is in progress, completed, or failed. A completed test is one in which all attacks were sent and their responses evaluated.

Finding AI Asset

Use the AI Assets tab for an alphabetical list of discovered models and applications in your environment. If the asset you're looking for is missing, see "Add an AI Asset," below.

Adding an AI Asset

AI Validation can test AI models and applications:

- An **AI model** is an LLM that has been auto-discovered using Multicloud Defense. Currently, only AWS Bedrock is supported. You can see the discovered AI models in the **AI Assets: Cloud visibility** tab. If the model you're looking for is missing, make sure you've connected to the cloud service that hosts your models and applications. See AI Defense's **Administration** page to connect to a cloud service.
- An **application** is an an LLM-powered application, such as a chatbot, that you have manually registered using AI Defense's **Applications** page.

Configure and run a validation test of an AI model

A model validation test requires a set of parameters that will be used to connect to and test the AI model. To configure and run a validation test for a model:

1. Click the **Run validation** button in the Validation page.
2. In **Asset type**, choose "Model."
3. Specify the test parameters and the model to be tested:
   - **AI asset name**: The name of the model as discovered by Multicloud Defense. See the **AI Assets** page for a list.
   - **Model ID**: The model ID as stored in the platform hosting the model.
   - **Test name**: Give this test a memorable name to better find it later.
   - **Prompt template**: This is the JSON request payload that will be sent to the model's inference API in order to test it. This must include a `{{prompt}}` placeholder where the AI Defense-generated test prompts will be inserted.
   - **Response**: a JSON path that specifies where in the HTTP response. See "Formatting the response path" below.
4. Click **Submit**. The test will run immediately.

Configure and run a validation test of an AI application

An application validation test requires a set of parameters that will be used to connect to and test the application's AI model. Follow the steps below to configure and run a validation test for an application.

Note An application validation evaluates the model connected to the application, not the application itself.

1. Click the **Run validation** button in the Validation page.
2. In **Asset type**, choose "Application."

3. Specify the test parameters and the model to be tested:
   - **Application**: Application to be tested.
   - **Test name**: Give this test a memorable name to better find it later.
   - **Endpoint**: The endpoint of the LLM used by the application
   - **Inference API path**: The API path for model inference calls. For example: `/openai/deployments/gpt3.5`
   - **Prompt template**: This is the JSON request payload that will be sent to the model's inference API in order to test it. This must include a `{{prompt}}` placeholder where the AI Defense-generated test prompts will be inserted.
   - **Response**: a JSON path that specifies where in the HTTP response. See "Formatting the response path" below.
   - **HTTP headers**: Headers for the inference API connection. Specify the authorization values here.
4. Click **Submit**. The test will run immediately.

Formatting the response path

In the **Response** field, provide a JSON path that specifies where in the HTTP response JSON payload AI Defense can find the LLM's response string in order to validate whether the attack was successful. The path must point to a string value in the JSON payload.

Remember Each model provider uses its own response format. Check your model provider's API documentation for the correct format before you set the response path.

Note
- Whitespace and other special characters can be encoded as unicode (`\u0020`).
- Periods in JSON fields can be escaped with a backslash.
- Array elements can be specified with the element's index in square brackets, for example by including `[0]` when you want to retrieve the first element.

Response path examples

To retrieve a response from a top-level field:

- For example, if the endpoint returns a response like `{"response": "I am an AI Chatbot, how can I assist you?"}`
- You would set a **Response** value of `response`

To retrieve a response from a nested JSON field:

- For example, if the endpoint returns a nested response like

```
{"response": {"llmResponse": "I am an AI Chatbot, how can I assist
          you?"}}
```

- You would set a **Response** value of `response.llmResponse`

To extract a response string from an array, specify the element's index in square brackets.

For example, if the endpoint returns a nested response like:

```
{
    "content": [
        {
            "text": "Bonjour, je suis Claude!",
            "type": "text"
        }
    ],
    "id": "msg459674598",
    "model": "claude-3-5-sonnet-2024-08-20",
    "role": "assistant"
}
```

- You would set a **Response** value of: `content.[0].text`

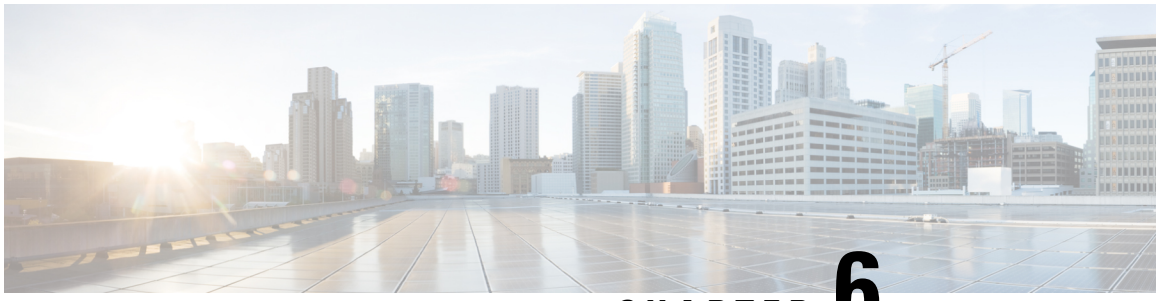
To handle periods in field names, use a backslash:

- For example, if the endpoint returns a nested response like `{"llm.response": "hello"}`
- You would set a **Response** value of `llm\.response`
- The syntax applies to dot notation only, such as `myfield.myotherfield` or `myarray.1`

Initial Configuration of AI Validation

To set up AI Validation:

1. In your cloud service (currently AWS Bedrock is supported), find the IAM role ARN for an account with access to your models.
2. Open the AI Defense **Administration** tab, go to the AWS Bedrock card, and click **Connect**, and provide the API key details to complete the connection. *See the* AI Defense Administration documentation *for details.*
3. Make sure Multicloud Defense is connected to AI Defense. If the Multicloud Defense card on the Administration tab shows a **Disconnect** button, then Multicloud Defense is connected. If it's not connected, *see the section* Set up AI Asset discovery.
4. Proceed to the sections, Find AI Asset and Add an AI Asset, above, to add the AI models and applications you wish to scan.



CHAPTER 6

AI Application Discovery

The **Generative AI Application Discovery** interface utilizes Cisco Secure Access to provide insights into user connections to third-party AI applications, along with an assessment of each application's safety.

This page displays key statistics, including the number of users accessing each application and its associated safety rating. You can click on any application to navigate to the Secure Access inspection page, where detailed risk information for that application is available.

Filter the Discovery List

You can filter the events log list view based on risk level. Click the Risk drop-down list and choose the risk levels of the applications to be shown in the table.

You can choose which columns to display by clicking the settings icon on the right top corner of the table. You can select from:

- Application name
- Risk
- First detected
- Total web traffic
- DNS requests
- Last detected
- Blocked DNS requests
- Firewall events
- Inbound web traffic
- Outbound web traffic

Click **Apply.** This changes the columns displayed in the table.

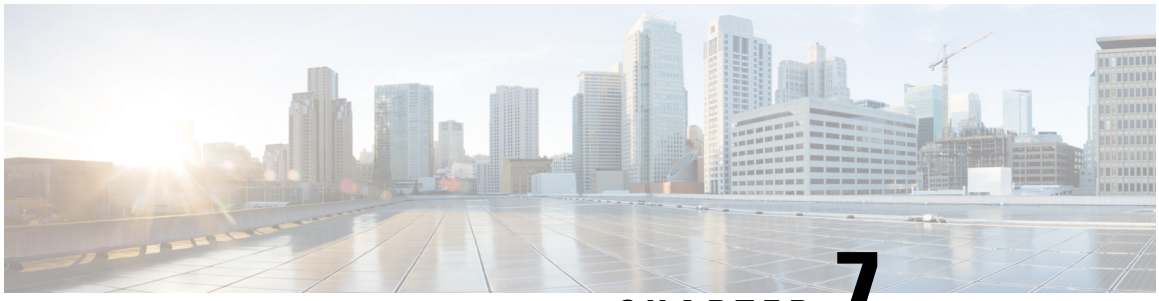
Initial Configuration of AI Application Discovery

To set up AI application discovery:

1. Create an API key in Secure Access. Capture the API Key for use in the next step.
2. Open the AI Defense Administration tab, go to the Secure Access card, and click Connect, and provide the API key details to complete the connection. See the AI Defense Administration documentation for details.

After Secure Access has completed its scan, the AI App Discovery tab in AI Defense will display the AI Applications being used in your environment.

Note You must have an instance of Cisco Secure Access.



CHAPTER 7

AI Defense Assets

Integrates with Cisco Multicloud Defense to detect all AI workloads in your environment, including AI models, agents, and knowledge bases.

Once you've created an AI Defense connection to Multicloud Defense, AI Defense crawls your environment to detect all AI Assets (models, agents, and knowledge bases). Once a model has been discovered, it appears in the AI Assets: Cloud Visibility tab, and you can test it for vulnerabilities in the Validation tab.

The AI Assets page has:

- **Cloud visibility**: Provides an overview of your cloud environment, focusing on AI-related activities and assets.
- **External assets**: Designed to track and manage AI resources that exist outside the organization's direct cloud infrastructure.

Cloud Visibility

Cloud Visibility provides an overview of AI activities and assets within your cloud infrastructure, helping organizations identify models, applications, and resources in use. It offers insights into usage patterns, potential risks, and compliance status while monitoring hosting regions and VPC instances. This feature enhances transparency and control over AI workloads across multi-cloud environments.

- **Discovered AI assets**: Shows a count of the models, agents, and knowledge bases discovered.
- **Model validation status**: Indicates progress in validating the models in your environment. See the Validation page for test run results.
- **AI assets table**: Lists the discovered AI assets. For each, details are shown in the columns:
  - **AI Asset Name**: A clickable link directing you to the inspection page for the discovered asset. Select the link to access the **AI Asset** Details page.
  - **Asset Type**: Specifies the category of the asset, such as **Custom Model** or **Foundational Model**
  - **Discovered Date**: The date when the asset was first detected during a scan.
  - **Regions**: Identifies the cloud region where the asset is hosted.
  - **Last Validation**: Displays the timestamp of the most recent test performed on the asset. Click the timestamp to review the corresponding **Test Report**.
  - **Action**: Provides the option to initiate or re-run a test for the asset.

- **AI asset details**: Provides details that AI Defense has discovered about this asset. To re-run the validation scan, click the **Validate** button near the top of the screen.

External Assets

External Assets tracks and manages third-party AI resources, including generative AI applications and external knowledge bases. It offers visibility into their usage, risk scores, and protection status, ensuring organizations can secure external AI dependencies and mitigate the risks posed by shadow AI or unmonitored assets.

We detect traffic that goes to third-party models hosted outside your AWS cloud. For these, no validation is run, but AI Defense can show details about the network traffic to the model.

This data comes from Cisco Multicloud Defense. For an explanation of these fields, see the MCD documentation.

For each instance in your cloud that connects to an external AI model, the **Instances connecting to external assets** table shows:

- **Resource Name:** Displays the AWS VPC, subnet, and instance involved in the connection to an external model.
- **Account:** Identifies the AWS account associated with the instance that initiated the model connection.
- **Region:** Specifies the cloud region where the instance initiating the model connection is located.
- **VPC ID:** Indicates the unique identifier of the VPC hosting the instance that established the model connection.
- **Last Detected Date:** Shows the most recent timestamp when this resource connected to the external model.
- **Source IP:** Lists the IP address used by the instance to connect to the model.
- **Instances:** Provides details about the specific instances involved in the connection.

Initial Configuration of AI Assets

Remember

Prerequisite: Your organization must have a Multicloud Defense (MCD) tenant for use by AI Defense. If you don't have an MCD tenant available, see below:

Configuring Multicloud Defense for AI Defense Integration.

To set up AI Assets detection:

1. Create an API key in Multicloud Defense. Capture the API Key ID and API Key Secret for use in the next step. See Cisco Multicloud Defense User Guide - Management [Cisco Defense Orchestrator]
2. Open the AI Defense Administration tab, go to the Multicloud Defense card, and click Connect, and provide the API key details to complete the connection. See the AI Defense Administration documentation for details.
3. Return to Multicloud Defense and use the Connect Account button to connect Multicloud Defense to your cloud account. See Cisco Multicloud Defense User Guide - Setup with the Multicloud Defense Wizard [Cisco Defense Orchestrator]

4. In Multicloud Defense, you must either Enable Traffic Visibility, Secure Your Account, or both. See Cisco Multicloud Defense User Guide - Setup with the Multicloud Defense Wizard [Cisco Defense Orchestrator] and Cisco Multicloud Defense User Guide - Setup with the Multicloud Defense Wizard [Cisco Defense Orchestrator]

Note AWS is supported for early access

After Multicloud Defense has completed its scan, the AI Assets tab in AI Defense will display the AI Models in your cloud and the external models that call into instances in your cloud.

Configuring Multicloud Defense for AI Defense Integration

If your organization does not have a Multicloud Defense tenant available, follow the instructions below to create a new Multicloud Defense tenant:

Set Up Multicloud Defense (MCD) within Security Cloud Control (SCC)

If You Already Have Both SCC and MCD Accounts

- Login to SCC: Access your account.
- Navigate to MCD Management:

  Left-hand pane → Administration → Multicloud Defense Management.

If You Have an SCC Account but Haven't Activated MCD

- Login to SCC: Access your account.
- Activate MCD: Follow the steps below to enable MCD.

If You Do Not Have an SCC Account

- Create an SCC Account:

  Visit https://getcdo.com/.

  Follow the instructions to create an account.
- Enable MCD: After account creation, proceed with the steps below.

Enable Multicloud Defense

- Access MCD Management:

  Left-hand pane → Administration → Multicloud Defense Management (accept the EULA).
- Initiate Cloud Protection:

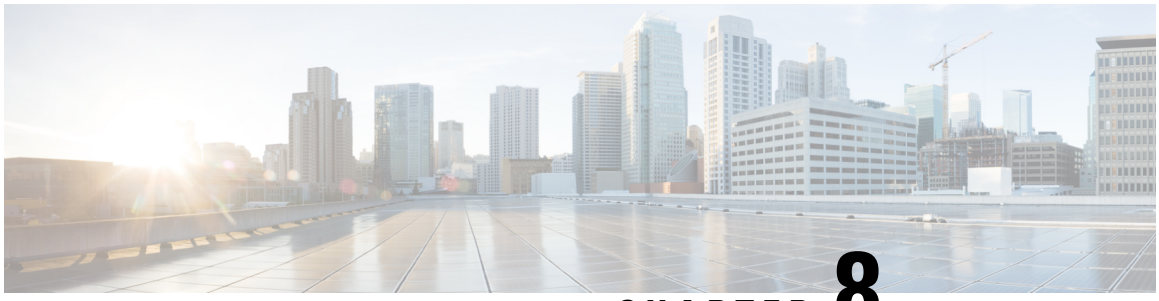
  Click the rocket ship icon (top right) → Select Protect cloud assets.
- Enable Multicloud Defense:

  Click Enable Multicloud Defense.

  Follow the on-screen prompts to create an MCD tenant (takes a few minutes)

Next Step

Return to the preceding section, Initial Configuration of AI Assets to connect AI Defense to your MCD tenant.



CHAPTER 8

AI Defense Policies

Create policies using built-in guardrails and templates to ensure data protection, legal compliance, and responsible AI use. These policies are applied to your connected endpoints, helping to secure and govern their interactions.

To create a new policy, click **Create Policy** on the right top corner of the Policies page. Select from Gateway and API.

To create a new policy:

Note A left slide-in pane is displayed. This pane has a step-by-step guide for creating a new policy.

Procedure

Step 1 On the **New Policy** page, under **Policy Details**, enter a name for the policy.

The **Description** is optional, you can add details if required.

a) Click **Next**.

Step 2 A list of connections is displayed, you can select the checkbox for one or multiple connections to apply the policy. Click **Next**.

Step 3 Under **Security guardrails**(prompt injection and code detection), configure security rules.

a) Move the slider to **Enabled** for the rule.
b) You can click the dropdown for **Rule directionality** to apply this rule to prompts, responses or both.
c) Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.
d) Click **Next**.

Step 4 To protect data and maintain confidentiality, under **Privacy guardrails**, there are three rules:

a) Protected health information (PHI):

1. Move the slider to **Enabled**.
2. For **Rule directionality**, click the dropdown for **Directions** to select Prompts, Responses or Both.

3. Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.

b) PII (Personally Identifiable Information):

1. Move the slider to **Enabled**.
2. Slide to **Enable** the one or multiple PII entity.
3. Click the dropdown for **Directions** to select Prompts, Responses or Both.
4. Click the dropdown for **Action** to select the action this policy rule will take. You can select : Block or Allow.

c) Payment Card Industry (PCI):

1. Move the slider to **Enabled**.
2. Choose the entities for which you want to apply the subcategories and their direction and rule action.
3. For **Rule directionality**, click the dropdown for **Directions** to select Prompts, Responses or Both.
4. Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.

d) Click **Next**.

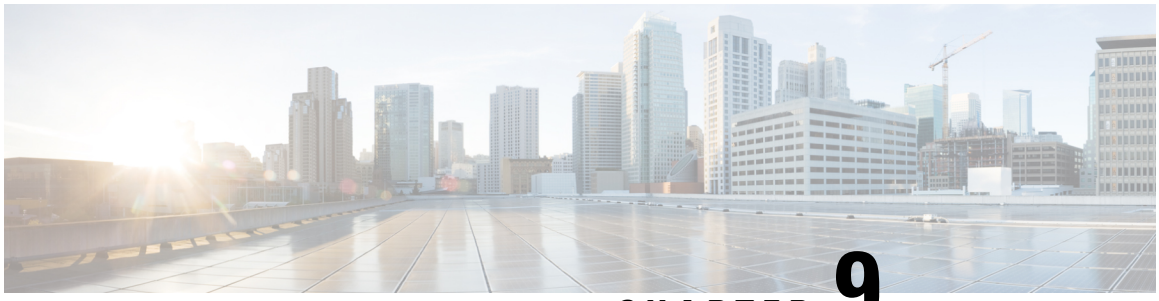
Step 5 Under **Safety guardrails**, configure security rules.

a) Move the slider to **Enabled** for the rule.
b) For **Rule directionality**, click the dropdown for **Directions** to select Prompts, Responses or Both.
c) Click the dropdown for **Action** to select the action this policy rule will take. You can select: Block or Allow.
d) Click **Next**.

Step 6 On the **Summary** page, review your policy details. Click **Save** to create the policy.

The policies are **disabled** by default, you would need to enable the policies.

The policy is created and displayed under Policy.



CHAPTER 9

AI Defense Applications

Easily manage your organization's generative AI applications by entering details of AI-powered apps and their connected AI model endpoints. This streamlined process ensures seamless integration, enabling efficient oversight and enhanced functionality across your AI ecosystem.

It is divided into two key sections:

- **Applications:** This section lists all the AI-powered applications within your organization. Each application entry provides essential details, such as the application name, associated policies, and its status. This enables quick identification and management of applications in alignment with organizational goals and compliance requirements.
- **Connections:** This section lists the individual endpoints linked to the listed applications. This section allows you to configure and manage endpoints, ensuring secure and efficient communication between applications and their AI models. Key functionalities include assigning guardrails, applying policies, and monitoring endpoint activity for enhanced security and operational control.

Together, these sections provide a comprehensive framework for monitoring, managing, and securing AI-powered applications and their integrations.

Applications

To get started with creating a protective layer with AI Defense, you'll need to **Add Application** and connect them to their respective AI model endpoints. You can select:

- **Gateway**: AI Gateway acts as an intermediary to manage, monitor, and protect AI interactions. The gateway acts as a filter and security checkpoint for all AI traffic, ensuring safe and compliant operations.
- **API**: APIs allow applications to interact programmatically with the AI Defense system, enabling seamless communication and enforcement of policies.

Note When creating an application, you will need to select a **Connection type**, which refers to the enforcement mechanism for your guardrail policies.

You'll need to map your applications to AI Defense and connect them to their corresponding AI model endpoints. This integration enables AI Defense to monitor and secure the underlying models that power your applications.

To add an application to AI Defense:

1. Click **Applications** on the left navigation pane. Click **Add application** button on the right top corner of the page.
2. On the **Add application** dialogbox, enter the **Name** for the application, **Connection type** (API or Gateway) and **Description** you would like to map. The Description is optional, you can add details if required. Click **Continue**.

API:

1. The application is created and now you can add a connection. Click **Add Connection** on the right top corner of the Applications page.
2. In the Add connection pane, enter connection name, click **Add connection**. Click **Add API Key**.
3. In the Add API Key pane, enter the **Token name**, select **Expiration date** and click **Generate API token**.
4. Once the token is generated, make a note of the API key, and click **Add API Key**.

Note For more information on API key, see our AI Defense API Document.

Gateway:

1. The application is created and now you can add a connection. Click **Add Connection** on the right top corner of the **Applications** page. Enter **Connection name** and select the dropdown for **Endpoint.** In case, no endpoints are added. Click **Add endpoint**.
2. Click **Add connection**.

Note The connection guide provides a cURL command to update the proxy URL, ensuring that all AI interactions are effectively monitored.

Filter Application List

You can filter the applications list view by clicking the settings icon on the right top corner of the table. You can select from:

- Name
- Description
- Connection type
- Connections

Click **Apply.** This changes the columns displayed in the table.

Add Endpoint

An endpoint is a crucial connection point that allows AI Defense to interact with, monitor, and secure the AI models your applications rely on.

By adding an endpoint, you provide AI Defense with the necessary details to communicate with the AI model.

To add a new endpoint:

1. On the Add/manage Endpoints dialogbox, click **Add**.
2. In the **Add endpoint** dialogbox:
   a. Select **Model provider** from the dropdown list.
   b. Select **Endpoint** from the dropdown list.
3. Click **Add** at the bottom of the dialogbox.

Connections

The Connections section focuses on managing the individual endpoints associated with your listed applications. It provides tools to configure and oversee these endpoints, ensuring seamless and secure communication between your applications and their connected AI models.

Filter Connections List

You can filter the connections list view by clicking the settings icon on the right top corner of the table. You can select from:

- Name
- Connection type
- Endpoints
- Models
- Application
- Policy
- Status
- Last active

Click **Apply.** This changes the columns displayed in the table.

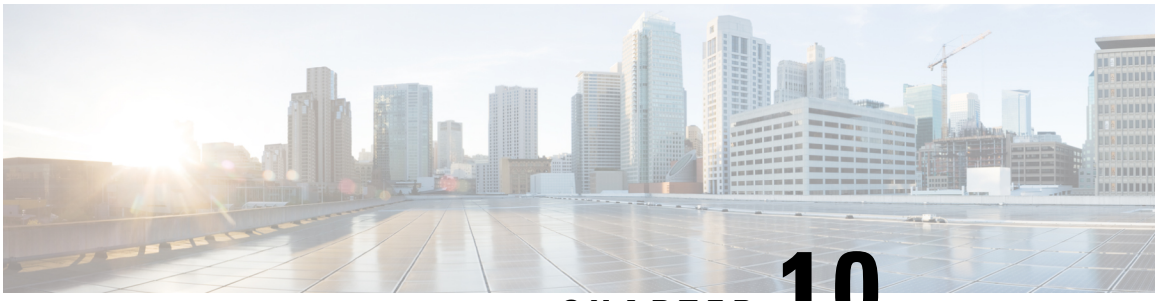
Connection Guide

You can access the **Connection Guide** by clicking on the three dots in the same row as the **Name** of the application.

AI Defense provides two primary methods for integrating applications: **API Method** and **Gateway Method**. These methods enable developers to seamlessly secure and monitor AI interactions based on organizational requirements.

The connection guide provides step-by-step instructions to implement both methods:

- **API Method**: Developers can follow detailed API documentation, including endpoint definitions, authentication steps, and examples of integrating Guardrails and monitoring policies. The connection guide provides you with details of the added API key.
- **Gateway Method**: The guide outlines network or proxy configurations required to route traffic through the AI Gateway, ensuring seamless and secure application integration. The connection guide provides you with all the details of the connection and endpoint. You can edit the cURL code to add to your proxy server.



CHAPTER 10

AI Defense Administration

Provide the details of your organization's AI-powered applications and their associated AI model endpoints to enable seamless integration within the system. By entering these details, you establish a connection between your applications and the AI Defense platform, allowing for effective monitoring, compliance, and protection.

These integrations streamline the interaction between diverse systems and applications, ensuring smooth data flow and enriched functionality across platforms. These integrations enable consistent application of policies, guardrail monitoring, and runtime protection, allowing your organization to maintain compliance and mitigate risks associated with AI-powered operations. Additionally, by linking model endpoints, you can track usage patterns, assess risks, and receive actionable insights tailored to your organization's unique needs.

Connect Multicloud Defense

You must connect to Cisco Multicloud Defense before you can use the AI Assets feature to detect AI workloads in your environment.

To connect AI Defense to Multicloud Defense, you must provide your:

- Multicloud Defense account name
- Multicloud Defense API key
- Multicloud Defense API secret

Connect Secure Access

You must connect to Cisco Secure Access before you can use the AI Access feature to show third-party AI applications that your users connect to.

To connect AI Defense to Secure Access, you must provide your:

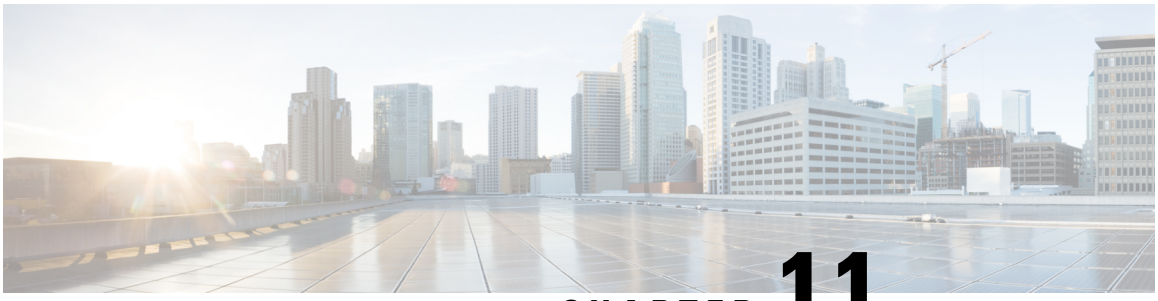
- Secure Access account name
- Secure Access API key
- Secure Access API secret

Connect AWS Bedrock

You must connect to AWS Bedrock in order to scan AI models and applications hosted there.

To connect:

1. Find or create an IAM role with permission to invoke models in AWS Bedrock. Make sure this role has access to all models you plan to scan. Copy the role ARN.
2. In AI Defense, navigate to **Administration**, **Integrations**, **AWS Bedrock**. Click **Edit**.
3. In the Connect AWS Bedrock pane, Paste the role ARN into the **Bedrock Inference IAM Role ARN** field and click **Connect** at the bottom of the pane.



CHAPTER 11

AI Defense FAQ

Q. What is AI Defense?

A. AI Defense is a comprehensive security solution designed to empower organizations to confidently adopt and integrate generative AI into their operations. It provides a cutting-edge, user-centric, and transparent suite of tools focused on ensuring the highest standards of data protection, compliance, and ethical integrity in an evolving AI landscape. With AI Defense, organizations can secure their AI operations, maintain compliance with industry regulations, and uphold ethical standards in their use of generative AI.

Q. How does AI defense protect us?

A. Addressing the risks of adopting generative AI requires a comprehensive strategy that integrates strong security protocols, well-defined policies, and cutting-edge technology. AI Defense offers an all-in-one solution that enables organizations to identify potential risks and safeguard their operations effectively.

Q. How do I get started with AI Defense?

A. Use your Cisco SSO credentials to log in to your AI Defense account. Once logged in, simply add an application to get started.

Q. How does AI Defense use my data? What type of data is captured by AI Defense

A. AI Defense does not capture or store any personal data. Instead, it monitors AI interactions to ensure they comply with established rules and regulations. When a prompt or AI response violates these guidelines, an event is generated. This event contains relevant details to help administrators review and address potential issues but does not involve the direct collection of user data.

The system focuses solely on ensuring that AI usage remains within safe, compliant boundaries without compromising privacy.

Q. What are policies?

A. Policies are customizable set of guardrails and rules designed to meet the unique security, privacy, and relevancy requirements of organizations. Each policy contains three types of guardrails—security, privacy, and relevancy—offering a flexible way to tailor and assign protective measures to different associations based on their specific needs. These policies are assigned to connections and each connection can have one policy.

Q. What are guardrails?

A. Guardrails in AI refers to predefined rules, or mechanisms that ensures AI-adapted organizations operate within safe and secure boundaries. These guardrails are configured as a part of policy and they help prevent unintended actions, security vulnerabilities, and compliance violations.

AI Defense guardrails keep traffic secure, ensure privacy is maintained, and avoid exposing sensitive data.

Q. How are the guardrails protecting my traffic?

A. Guardrails scan your traffic for security, privacy and safety by ensuring it flows in secure, controlled, and compliant ways:

- Cybersecurity and Hacking: Obtain or provide assistance to conduct cybersecurity attacks or deliberate misuse of systems.
- Model Vulnerabilities: Exploit weaknesses in a model with the intent to compromise its security, integrity, or functionality.
- PII (Personally Identifiable Information): Obtain or provide people's private and sensitive information, including phone numbers, addresses, emails, and any other personal information.
- Intellectual Property Theft: Steal or misuse any form of intellectual property from the victim organization, including copyrighted material, patent violations, trade secrets, competitive ideas, and protected software, with the intent to cause economic harm or competitive disadvantage to the victim organization.
- Financial Harm: Financial harm involves the loss of wealth, property, or other monetary assets due to theft, arson, vandalism, fraud, or forgery, or pressure to provide financial resources to the adversary.
- Reputational Harm: Reputational harm involves a degradation of public perception and trust in organizations. Examples of reputation-harming incidents include scandals or false impersonations.
- Societal Harm: Societal harms might generate harmful outcomes that affect the public or specific vulnerable groups.
- User Harm: User harms may encompass various harm types, including financial and reputational, that are directed at or felt by individual victims of the attack rather than at the organization level. Responses may contain specialized financial, medical, or legal advice, or indicate dangerous activities or objects as safe.

Q. What are the different types of attack prompts that AI Defense detects?

A. AI Defense identifies various types of adversarial prompts, including:

- Direct Request: A prompt directly asking for inappropriate or toxic output without any attempt to disguise the intent.
- Indirect Request: A prompt that provides access to a third-party data source containing adversarial content.
- Instruction Injection: A prompt that instructs the model to ignore or bypass previous instructions or guidelines.
- Obfuscation: A prompt that appears harmless but subtly shifts into harmful or inappropriate content.

- Fictionalization: A prompt that hides an inappropriate request within a fictional or role-playing context.

Q. How do I direct traffic to a specific group of users?

A. Currently, AI Defense does not support directing traffic to specific groups of users. Traffic routing is limited to application-based and model-based configurations. This means that you can control how traffic is routed through specific applications or models, but not by user groups at this time.

Q. How do I send user information for user-level reporting?

A. To enable user-level reporting, you must include user-specific information in the requests sent to the AI model. This could involve passing user identifiers, such as user IDs or roles, as part of the input.

Q. Why don’t I see any user data?

A. You are not seeing any user data because it hasn't been included as part of the requests sent to the AI model.

Q. What types of models are we leveraging to provide protection?

A. We leverage proprietary models developed specifically by Cisco to ensure the security, privacy, and safety of AI-adopting organizations. These models are designed to detect threats, enforce compliance, and provide robust protection against vulnerabilities unique to AI operations. By integrating advanced security measures into AI workflows, these models safeguard organizations from emerging risks in the AI landscape.

Q. How many team members do we need to bring to the deployment discussion call?

A. For the deployment discussion, you'll need the following key team members:

- A team member with details of the application, including the endpoint URLs that the application uses to communicate with the AI models. This person will be responsible for defining the application within the admin console and providing the necessary endpoint information.
- A team member who can route traffic from the application to our proxy. This person will be required to follow the connection guide to direct traffic correctly.

Q. How do I know if my traffic is going through AI Defense?

A. You can verify that your traffic is being routed through AI Defense by navigating to the **Applications** section of the AI Defense Admin Console. If traffic is successfully processed, you will see logged events, including application names, timestamps, associated models, and any policy enforcement actions.

Additionally, check the connection status. If it displays as "Connected," it indicates that traffic is successfully passing through the proxy, ensuring that AI Defense is actively monitoring and securing your AI operations.

Q. Why does my connection status show pending?

A. When the connection status displays as pending, it means that the application has been successfully added to AI Defense and is ready for use, but no traffic has yet been routed through the AI gateway. To start

directing traffic, use the provided connection guide to help configure the proxy. Once the first request passes through the proxy, the status will automatically update to connected.

Q. Where do I find the connection guide for a connection?

A. The connection guide for a specific connection can be found on the connections page under **View Connection Guide**.

Q. What are events?

A. Events are recorded instances of AI activity that are captured and logged by the AI Defense system. Each event represents an interaction or action taken within your AI environment, such as a prompt submitted to the model, a response generated, a rule violation, or any other significant activity related to your AI's operation.

Events typically include key details like:

- Time stamp: The exact time the event occurred.
- Application: The specific application or service involved in the event.
- Rule Matches: Any security or compliance rules that were triggered.
- Conversation: Provides admins an opportunity to review the conversation and the reason behind the rule match.
- Action Taken: The system’s response, such as blocking, alerting, or allowing the action.
- Model Used: The AI model that processed the interaction.

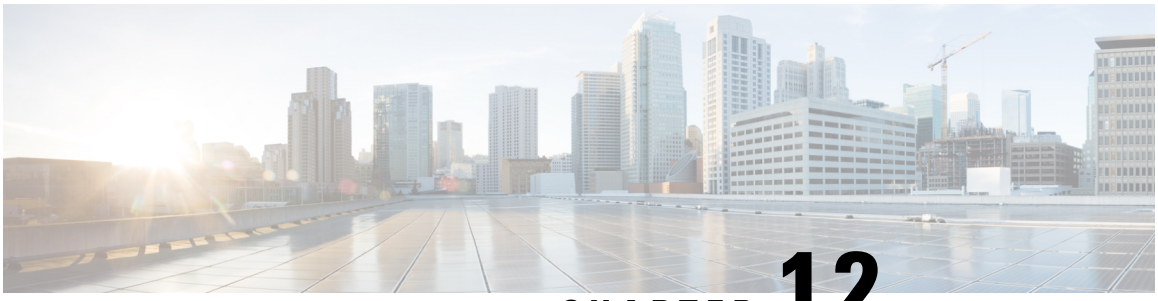
By tracking events, you can monitor AI usage, detect threats, and ensure compliance with security policies in real-time.

Q. How do I connect my application and models to AI Defense?

A. To connect your application to AI Defense, follow these steps:

1. Navigate to the Applications page in the AI Defense Admin Console.
2. Click Add Application and provide a name for your application.
3. Select the endpoint associated with your application from the list.
4. Enter Connection name and click Save

Note If the endpoint is not already defined, you can add it by selecting the appropriate provider and entering the endpoint URL.



CHAPTER 12

What's New

December 04, 2024

Enhanced Policy Support: API Policies

- AI Defense now supports **API Policies** in addition to the existing Gateway Policies, offering greater flexibility and coverage for application security.
- The **API Connections** page has been redesigned to seamlessly manage both **Gateway Applications** and **API Applications**.

Simplified API Key Management

- The process of generating API keys for API connections has been streamlined. Edit the API connection and use the **Add API Key** window to create a new key. The API key is required for your application to interact with the AI Defense **Guardrail Services**.