

Establishing a Governance Framework for AI-Powered Applications

Artificial intelligence (AI) is advancing rapidly, with organizations across myriad industries deploying AI-powered applications at an unprecedented pace. In Prisma Cloud's *The State of Cloud-Native Security Report 2024*¹, for example, 100% of survey respondents said they are embracing AI-assisted application development—an astonishing result for a technology many considered science fiction just a few years ago. But while AI systems offer significant benefits, they also introduce novel security risks and governance challenges that traditional cybersecurity approaches are ill-equipped to handle.

As a security leader, it is imperative to understand the current state of AI, its potential implications, and the unique risks it poses to your organization. In this whitepaper, we provide an overview of the AI landscape, highlight key security considerations, and propose a governance framework to help you navigate this complex terrain. You will get a clear understanding of the key AI technologies your organization might be exploring and their security implications, as well as the potential guardrails and mitigations you'll need to consider. We'll also help you understand the factors you'll need to take into account when building your AI security strategy.

¹ <https://www.paloaltonetworks.com/state-of-cloud-native-security>

Current AI Landscape and Its Security Implications

The AI landscape has changed dramatically in recent years, with significant advancements in both general and specialized systems as organizations evaluate and implement a slew of new technologies.

To manage AI's associated risks and opportunities, security leaders need to keep their finger on the pulse of this evolving ecosystem. Risk can manifest in myriad ways, as seen in our recent research, which found that 47% of organizations are concerned with security risks associated with AI-generated code².

In this paper, we will focus on the cloud infrastructure risks that are manifested when deploying AI powered applications.

As of today, here are the basic contours of the AI landscape.

“Traditional” AI and ML

Organizations for many years have been leveraging domain-specific AI and machine learning (ML) inference tools to drive efficiency and automation in manufacturing, supply chain optimization, fraud detection, online marketing, and other areas. These narrow systems are trained on particular datasets to perform well-defined tasks. While they have become integral to many business processes, each application is typically siloed and limited in scope.

While these tools have been around for a long time and may have not experienced breakthrough developments in recent years, they are impacted by the “rising tide that lifts all boats.” The general excitement around AI is accelerating budgets and appetites for deploying these and newer LLM-based tools.

Security Implications

- Ensuring the integrity and security of the data used to train and operate AI/ML models
- Monitoring AI/ML systems for anomalous behavior, unexpected outputs, or performance degradation that can indicate security issues like data poisoning attacks, model evasion, and system compromise
- Validating that AI/ML models are not introducing unintended biases, fairness issues, or discriminatory outcomes that could create legal and reputational risks, which requires models to be regularly audited for fairness
- Providing appropriate access controls and governance regarding who can develop, modify, or use AI/ML systems



Traditional AI and Machine Learning

Use Cases

- Manufacturing optimization, supply chain, fraud detection, marketing

Security Risks


- Data integrity, model performance monitoring, bias auditing, access controls

² <https://www.paloaltonetworks.com/state-of-cloud-native-security>

Large Language Models

LLMs have emerged as highly versatile, general-purpose AI systems capable of engaging in open-ended dialogue, generating coherent text, and even writing code. While consumer-facing chatbots like ChatGPT have garnered widespread attention, they are just the tip of the iceberg. The larger impact of LLMs is their ability to power a wide range of enterprise applications—both internal and customer-facing—in ways that often are not immediately transparent to end users.

Within this broader context, there are several categories of implementation patterns, each relying on a different set of tools and related security implications.



LLM

	Use Cases	Security Risks
Pre-trained foundation models	<ul style="list-style-type: none">• Content generation, chatbots, sentiment analysis, translation, code assistants	<ul style="list-style-type: none">• Shadow AI projects, data privacy, model governance, output control
Fine-tuning and RAG	<ul style="list-style-type: none">• Specialized AI assistants (support, HR, IT), Q&A apps (docs, code, training)	<ul style="list-style-type: none">• Exposure of sensitive data during fine-tuning, data governance
Custom model training	<ul style="list-style-type: none">• Advanced applications (drug discovery, materials science, autonomous systems)	<ul style="list-style-type: none">• Data poisoning attacks, compute resource isolation, and model accountability & auditability

Use of Pre-Trained LLMs (Proprietary or Open Source)

Cloud providers like OpenAI and Anthropic offer API access to powerful LLMs that they manage and secure. Organizations can leverage these APIs to incorporate LLM capabilities into their applications without having to manage the underlying infrastructure.

Alternatively, open-source LLMs such as Meta's LLaMa can be run on an organization's own infrastructure. This provides more control and customization options, but requires significant compute resources and AI expertise to implement and maintain securely.

LLMs are available through various deployment models:

- **API-based SaaS:** The infrastructure is provided and managed by the LLM developer (e.g., OpenAI) and provisioned via a public API.
- **CSP-managed:** The LLM is deployed on infrastructure provided by cloud hyperscalers and can run in a private or public cloud, such as Azure, OpenAI, and Amazon Bedrock.
- **Self-managed:** The LLM is deployed on the company's own infrastructure, which is relevant only for open-source or homegrown models.

Typical Use Cases

LLM use cases include content generation, chatbots, sentiment analysis, language translation, and code assistants. An e-commerce company might use an LLM to generate product descriptions, while a software development firm could leverage an LLM-powered coding assistant to boost programmer productivity.

Security Implications

The availability of easily accessible cloud APIs and open-source models has dramatically lowered the barriers to adding advanced AI language capabilities to applications. Developers can now plug LLMs into their software without maintaining deep expertise in AI and ML. While this accelerates innovation, it increases the risk of shadow AI projects that lack proper security and compliance oversight. Development teams, meanwhile, may be experimenting with LLMs without fully considering data privacy, model governance, and output control issues.

Fine-Tuning and Retrieval-Augmented Generation (RAG)

To customize LLMs for specific applications, organizations can fine-tune them on smaller datasets related to the desired task or implement RAG, which involves integrating LLMs with knowledge bases for question-answering and content summarization.

Typical Use Cases

These include specialized AI assistants with access to internal data (e.g., for customer support, HR, or IT help desk) and Q&A apps (e.g., for documentation, code repositories, or training materials). For example, a telecommunication company's customer service chatbot could be fine-tuned on product documentation, FAQs, or past support interactions to better assist customers with technical issues and account management.

Security Implications

Fine-tuning and RAG allow organizations to adapt LLMs to their specific domain and data, enabling more targeted and accurate outputs. But this customization process often involves exposing the model to sensitive internal information during training. Strong data governance practices are required to ensure that only authorized data is used for fine-tuning and that the resulting models are secured.

Model Training

Some large technology companies and research institutions are investing in training their own LLMs from scratch. This is a highly resource-intensive process that requires massive compute power and datasets. It does, however, allow companies to have full control over the model architecture, training data, and optimization process, as well as to maintain full intellectual property rights over the resulting models.

Typical Use Cases

Use cases for proprietary LLMs include highly specialized applications like drug discovery, materials science, or autonomous systems. For example, a healthcare organization could develop a model to help diagnose diseases from medical records and imaging data.

The availability of cloud APIs and open-source models increases the risk of shadow AI projects

Model training by fine-tuning and RAG can lead to exposing the model to sensitive internal information

Security Implications

Training custom LLMs requires carefully curating massive datasets and building high-performance computing infrastructure, which can introduce new security challenges. The training data must be thoroughly vetted for sensitive information and personal data that the model can ingest. There is also concern over data poisoning attacks, when an adversary intentionally injects malicious examples into the training data to manipulate model behavior.

The training process consumes enormous compute resources, necessitating strong isolation and access controls around the training environment to prevent abuse or interference. There are also difficult questions regarding how to maintain accountability and auditability of model behavior when dealing with complex black-box models. And these questions may be even more pressing when the model is homegrown.

Training custom LLMs requires thorough vetting for sensitive information and personal data

How Security Leaders Should Conceptualize AI Risk

Cybersecurity teams are used to playing catch-up. More often than not, they have to adapt their strategies and controls to keep pace with the adoption of new technologies—e.g., cloud computing, containerization, and serverless architectures. The rise of AI, however, presents new obstacles for cybersecurity—many of which are qualitatively different from those in the past.

Change Is Faster and Often More Extreme

AI is causing a seismic shift in organizations, and it is happening far faster than previous transformations:

- New models, techniques, and applications are emerging at breakneck speed, with major breakthroughs occurring on a monthly, or even weekly, basis.
- Organizations are feeling immense pressure to quickly adopt and integrate these technologies to stay competitive and drive innovation.
- The availability of tools through simple APIs and the emergence of a supporting ecosystem of tools and frameworks have accelerated adoption by removing roadblocks caused by skill shortages.

This combination of rapid technological change and urgent business demand can pose unique security challenges. Compressed timelines make it difficult to assess risks thoroughly and to implement appropriate controls before AI systems are deployed. Security often becomes an afterthought in the rush to market. Best practices and regulatory guidance struggle to keep pace, which means security teams must adapt on the fly and make judgments without the benefit of industry consensus or clear standards.

However, this also presents an opportunity to secure AI by design. By embedding security and governance considerations into the AI development process from the outset, organizations can proactively mitigate risks and build more resilient AI systems. It requires close collaboration between security, legal, and AI development teams to align with best practices and integrate necessary controls into the AI lifecycle.

Broader Implications

The potential impact of AI systems is far-reaching and not always well understood. AI models can automate high-stake decisions, generate content with legal implications (e.g., use of copyright-protected material), and access vast amounts of sensitive data. The risks associated with AI—such as biased outcomes, privacy violations, intellectual property exposure, and malicious use—require a fundamentally different risk management paradigm.

Addressing these broader implications requires cybersecurity leaders to engage with stakeholders across legal, ethical, and business functions. Collaborative governance structures need to be established to align with risk tolerance, develop policies and guidelines, and implement ongoing monitoring and auditing processes. Cybersecurity teams will have to work closely with data science and engineering teams to embed security and risk management into the AI development life cycle.

Cybersecurity teams will have to work closely with data science and engineering teams to embed security and risk management into the AI development life cycle

New Types of Security and Compliance Oversight Are Required

Traditional cybersecurity frameworks are often focused on protecting data confidentiality, integrity, and availability. With AI, however, additional dimensions, such as fairness, transparency, and accountability, come into play.

Emerging regulatory frameworks like the EU AI Act are placing new demands on organizations regarding oversight and governance mechanisms for AI systems. These regulations require companies to assess and mitigate the risks associated with AI applications, particularly in high-stake domains such as hiring, credit scoring, and law enforcement. Compliance obligations may vary based on the specific use case and the level of risk involved. For instance, AI systems used for hiring decisions are likely to be subject to more stringent auditing and transparency requirements to ensure they are not perpetuating biases or discrimination.

This means that security teams need to go beyond their traditional focus on access controls and data protection. They must work with legal and compliance teams to establish mechanisms for monitoring and validating the actual outputs and decisions made by AI models. This may involve implementing explainable AI techniques, conducting regular bias audits, or maintaining detailed documentation of model inputs, outputs, and decision logic to support compliance reporting and investigations.

Novel Challenges in Threat Detection and Remediation

The technical challenges of securing AI systems are novel and complex and apply to both detection and remediation.

New Threat Categories Require New Detection Approaches

As mentioned, security teams need to monitor not only the underlying data and model artifacts, but also the actual outputs and behaviors of the AI system in production. This requires analyzing vast troves of unstructured data, and detecting novel threats such as data poisoning attacks, model evasion techniques, and hidden biases that could manipulate the AI's outputs in harmful ways. Existing security tools are often ill-equipped to handle these AI-specific threats.

Remediation Is Far From Straightforward

Unlike traditional software vulnerabilities, which often can be patched with a few lines of code, AI model issues may require retraining the entire model from scratch to fix them. AI models learn from the data they are trained on, which becomes deeply embedded in the model's parameters. If the training data is found to contain sensitive information, biases, or malicious examples, it is not possible to simply remove or correct the specific data points. Rather, the model must be completely retrained on a sanitized dataset, which can take weeks or months and cost hundreds of thousands of dollars in compute resources and human effort.

Moreover, retraining a model is not a guaranteed solution, since it may degrade performance or introduce new issues. While there is ongoing research into machine unlearning, data removal, and other techniques, they are still nascent and not widely applicable. As a result, prevention and early detection of AI vulnerabilities is paramount, as reactive remediation can prove costly and time-consuming.


Suggested Governance Framework for AI-Powered Applications

To manage the risks and opportunities presented by AI-powered applications effectively, we advise organizations to adopt new governance frameworks that focus on two key aspects—visibility and control.

Visibility is about gaining a clear understanding of how AI is being used across the organization. It includes maintaining an inventory of all deployed AI models, tracking what data is being used to train and operate these models, and documenting the capabilities and access permissions of each model. Without this foundational visibility, it is impossible to assess risk or enforce policies.

Control refers to the policies, processes, and technical safeguards that need to be put in place to ensure that AI is being used responsibly and in alignment with organizational values. It spans from data governance policies that dictate what information can be used for AI, through access controls that restrict who can develop and deploy models, to ongoing monitoring and auditing to validate model behavior and performance.

The goal is to provide a structured approach for security leaders to collaborate with stakeholders across the organization—including security, compliance, and engineering teams—to design and implement appropriate governance mechanisms for AI. Specific implementation details and policies can vary according to the organization's requirements, priorities, and local regulations.

	 Visibility	 Control (policies)
 Models	Which models are being used?	01 Sanctioned and unsanctioned models 02 Approval chain for new model training / usage deployment
 Data	Which data is being used for training / inference / fine-tuning?	01 Accepted sensitive data usage policies 02 Oversight of storage processing, data flows
 Use cases	How is AI being used?	01 Approved and unapproved use cases 02 Policies for use of AI agents
 Access	Who is using AI in the organization?	01 Stricter oversight of public-facing AI applications 02 Control policies
 Compliance	What are the relevant compliance frameworks?	01 Ongoing consideration of current and future risk 02 Ownership and accountability for violations

Models

Visibility Into Model Inventory

First, establish a comprehensive inventory of all models deployed across the organization. The model inventory provides a single source of truth for understanding the organization's AI footprint and forms the foundation for risk assessment and policy enforcement. This inventory should include key metadata such as the model provider, purpose, and intended use case, as well as model type (e.g., LLM), computer vision, tabular data, training data sources, access permissions and restrictions, and data flow diagrams.

Automated discovery tools can help identify models deployed in production environments, but manual processes may be needed to capture models in development or hosted externally.

Policies for Sanctioned and Unsanctioned Models

Organizations need to determine which models are approved for usage, evaluation, and customer-facing deployments. Policies should align with the organization's overall risk tolerance and compliance obligations.

Consider the following factors when sanctioning models:

- Model provenance and pedigree (e.g., reputable vendor vs. open source)
- Level of testing and validation
- Alignment with organizational values and principles for responsible AI
- Compliance with relevant regulations and industry standards
- Integration with security and governance controls

Models that do not meet these criteria should be blocked from deployment or be subject to far more stringent approval processes. For sanctioned models, guidelines should be established around appropriate use cases, required security and compliance controls, and ongoing monitoring and maintenance expectations. The guidelines help ensure consistency and risk management across deployments.

Policies to Vet and Approve New AI Models

Organizations need structured processes for security and compliance teams to vet new or pre-deployed AI models. This vetting process should assess the model against sanctioning criteria and identify specific risks or required controls.

Clear processes are necessary for promoting approved models through the software development life cycle, from development through staging to production, with testing and signoff requirements at each stage. Organizations should also develop processes to monitor deployed models continuously and to trigger additional reviews when significant changes, such as model retraining, data drift, or performance degradation, occur.

Data

Discovery and Classification Data Used for AI Model Training and Deployment

Data is the lifeblood of AI models. It is therefore critical to have clear visibility into what data is being used across the AI life cycle. This includes data used for training models, for inference or predictions in production, and for fine-tuning or retraining of models over time. Organizations should maintain an updated AI inventory that details all datasets and classifies them based on sensitivity level, regulatory requirements, and approved use cases. This inventory should integrate with the model inventory to provide traceability between models and their underlying data.

Policies to Prevent Data Poisoning

Governance of training data is particularly important in the context of data poisoning attacks. If an adversary is able to manipulate the training data, it can introduce hidden backdoors or biases that may be exploited to subvert the model's behavior in production. Strong access controls and continuous data flow monitoring are needed to mitigate this risk.

Policies Regarding the Use of Sensitive Data for Training, Inference, and Fine-Tuning

Organizations should establish clear policies governing how different types of data can be used for AI. The policies should be based on the sensitivity level of the data, regulatory requirements, and ethical considerations.

For example, policies may stipulate that personally identifiable information (PII) or protected health information (PHI) cannot be used for model training without explicit consent and anonymization. Similarly, there may be restrictions on using sensitive intellectual property or third-party data for AI without appropriate licensing and usage rights. Policies can also define required security and privacy controls for different types of data, such as encryption, access controls, and retention limits. Compliance teams should be involved in crafting these policies to ensure they align with relevant regulations like GDPR, HIPAA, and CCPA.

Use Cases

Visibility Into What AI Is Being Used For

Different AI use cases can have vastly different implications for security and compliance. For example, an AI-powered chatbot for customer support may require strict controls around data privacy and content filtering, while an internal AI tool for optimizing supply chain logistics might not raise the same concerns.

Use case documentation should capture key details such as:

- Business purpose and intended outcomes
- End users and stakeholders
- Data inputs and outputs
- Decision-making scope and autonomy
- Human oversight and intervention points
- Performance metrics and success criteria
- Risk assessment and mitigation plans

Policies for Approved and Unapproved Use Cases

Organizations must define policies that clearly delineate which, and under which conditions, are AI use cases permitted. Factors to consider when approving use cases include:

- Alignment with organizational values and principles for responsible AI
- Compliance with relevant laws, regulations, and industry standards
- Potential for harm or unintended consequences
- Level of human oversight and control
- Transparency and explainability requirements
- Reputational and brand risks

High-stake use cases involving decisions about individuals, such as lending, hiring, or healthcare diagnoses, warrant heightened scrutiny, and may require dedicated review boards or oversight committees. Low-risk use cases like internal productivity tools can follow a more streamlined approval process.

Define Policies for Approved Use of AI Agents

AI agents are a special class of AI systems that can make autonomous decisions and take actions based on their own previous outputs, without human intervention at each step. For example, an AI agent might be used to write, test, and optimize code. The use of such agents introduces additional governance challenges, as the potential risks and unintended consequences can be harder to predict and control. Organizations need to establish clear policies regarding when and how AI agents can be used, including:

- Defining the scope of the agent's decision-making authority
- Setting performance boundaries and fail-safe mechanisms
- Implementing robust monitoring and alerting for anomalous behavior

AI agents may require dedicated risk assessments and approval processes given their higher level of autonomy and potential impact.

Permissions and Access

Visibility Into Who Is Using AI in the Organization

Effective governance of AI requires an understanding of who is involved in developing and using AI across the organization, including:

- Roles and responsibilities (e.g., data scientist, ML engineer, product manager)
- Access permissions to AI development and deployment tools
- Machine identities, such as service accounts or API keys, that are used to access AI systems

Stricter Oversight of Public-Facing AI Applications

AI applications that interface directly with customers or the public require more governance and oversight compared to internal enterprise applications. This additional scrutiny might include:

- Rigorous bias and fairness testing across diverse demographic groups
- Adversarial testing to probe for safety risks and abuse potential
- More frequent or comprehensive compliance audits

Organizations should also consider implementing additional technical guardrails for public-facing AI, such as rate limiting, content filtering, and automated shut-off triggers based on predefined risk thresholds. Regularly scheduled audits and impact assessments are important to identify and mitigate emerging risks proactively.

Compliance

Oversight of Relevant Compliance Frameworks

AI governance does not exist in a vacuum but must be integrated with an organization's overall compliance management framework. This means aligning AI policies and controls with relevant laws, regulations, and industry standards.

Consider the following key compliance frameworks:

- Data protection regulations (e.g., GDPR, CCPA, HIPAA)
- Sector-specific regulations (e.g., FINRA for finance, FDA for healthcare)
- New and emerging AI-specific regulations (e.g., NYC AI bias law)
- Voluntary standards and certifications (e.g., IEEE, ISO, NIST)

Compliance teams should work closely with AI development and governance teams to map out applicable requirements and translate them into actionable policies and controls. Gap assessments, data protection impact assessments (DPIAs), and compliance-driven reviews at key points in the AI life cycle may be required.

Compliance oversight should also extend to third-party AI vendors and partners to ensure that their practices align with the organization's compliance obligations. Vendor risk management processes should incorporate AI-specific due diligence criteria and contractual requirements.

Ongoing Consideration of Current and Future Compliance Risk

The regulatory landscape around AI is undergoing major changes. New laws and standards are being proposed and enacted, often with significant implications regarding how organizations develop and deploy AI. Periodic reviews and audits should be instituted to ensure that compliance is maintained over time.

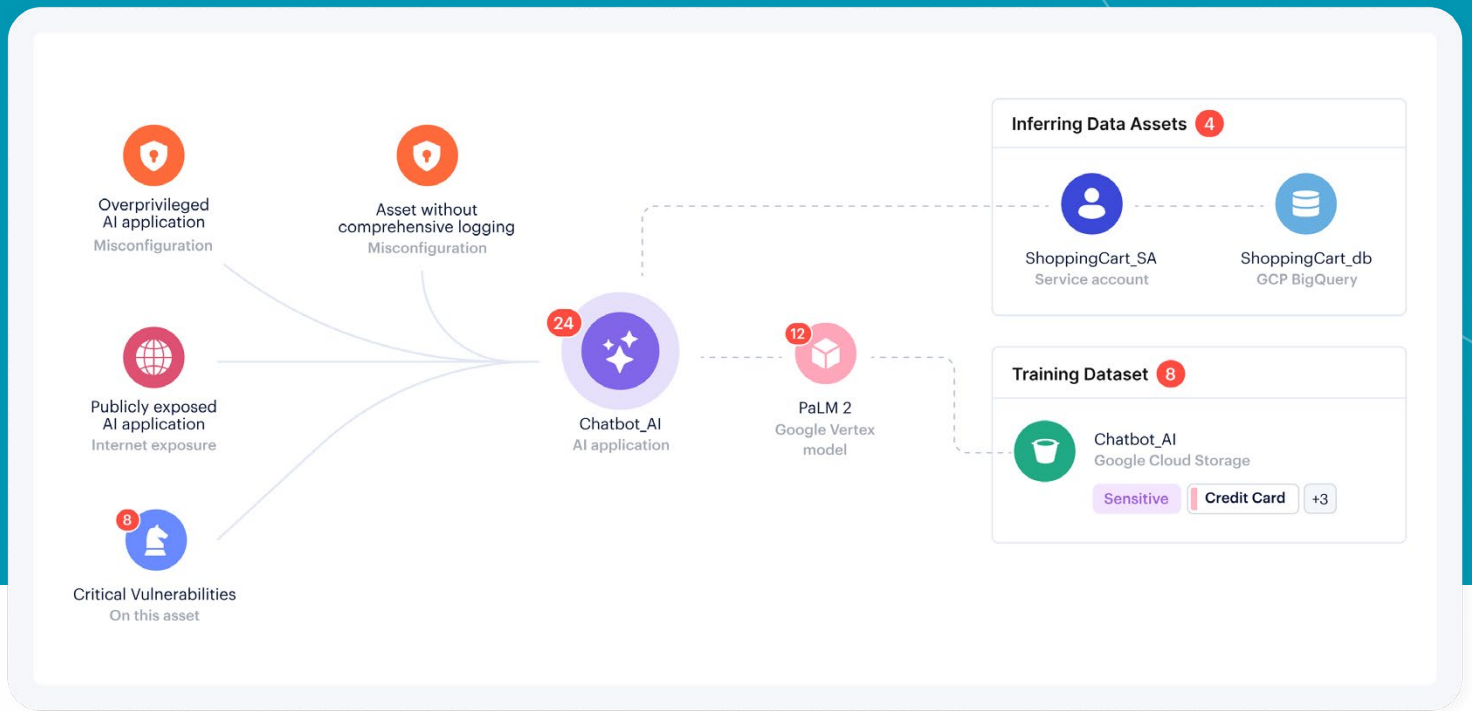
Prisma Cloud AI Security Posture Management: Visibility, Control, and Governance for AI

Given the risks and challenges outlined above, organizations need new tools and approaches to secure their rapidly expanding AI deployments. Indeed, when asked about their 2024 priorities, 100% of the organizations said they are committed to gaining visibility into the entire AI deployment pipeline (see Palo Alto Networks' *The State of Cloud Security Report 2024*³).

Prisma Cloud AI Security Posture Management (AI-SPM) provides a comprehensive solution to protect against the unique risks associated with AI, machine learning, and generative AI models.

Prisma Cloud AI-SPM delivers visibility into the full AI model ecosystem, from data ingestion and training to deployment. By analyzing model behavior, data flows, and system interactions, it identifies potential security and compliance risks that traditional tools miss.

³ <https://www.paloaltonetworks.com/state-of-cloud-native-security>



The solution's key capabilities include:

- **AI model discovery and inventory:** This involves creating an inventory of all model APIs, open-source models, and VM-deployed models in use across the organization. It helps control model sprawl and shadow AI, prevent unauthorized model use, and ensure that appropriate governance controls are in place.
- **Data exposure prevention:** AI-SPM discovers and classifies the datasets used to train and operate AI models, flagging potential exposure of sensitive data. It monitors live model interactions to detect misuse or unintended data leakage.
- **Posture and risk analysis:** By scanning the end-to-end AI deployment pipeline, AI-SPM identifies misconfigurations, weak access controls, and other vulnerabilities that could put models and data at risk. It provides a visual mapping of user access permissions and helps rightsize overly broad privileges.

Organizations can leverage these AI-SPM insights to enforce consistent security policies, proactively mitigate AI-specific threats, and maintain compliance with evolving regulations like the EU AI Act.

Prisma Cloud AI-SPM integrates seamlessly with its broader Code to Cloud™ platform, offering complete CNAPP capabilities—including CSPM and DSPM—to provide unified visibility and control across the full cloud-native stack. With a quick and easy agentless deployment model, Prisma Cloud can be up and running in minutes to help secure your critical AI assets and enable responsible innovation at scale.

To learn more, visit <https://www.paloaltonetworks.com/prisma/cloud/ai-spm>



3000 Tannery Way
 Santa Clara, CA 95054
 Main: +1.408.753.4000
 Sales: +1.866.320.4788
 Support: +1.866.898.9087
www.paloaltonetworks.com

© 2023 Palo Alto Networks, Inc. A list of our trademarks in the United States and other jurisdictions can be found at <https://www.paloaltonetworks.com/company/trademarks.html>. All other marks mentioned herein may be trademarks of their respective companies.

prisma_ds_AI Governance_July 2024