Google Cloud

A technical paper

Implementing Secure AI Framework Controls in Google Cloud

Tom Curry, Devon Beanish

Table of contents

Table of contents (cont.)

01 Introduction

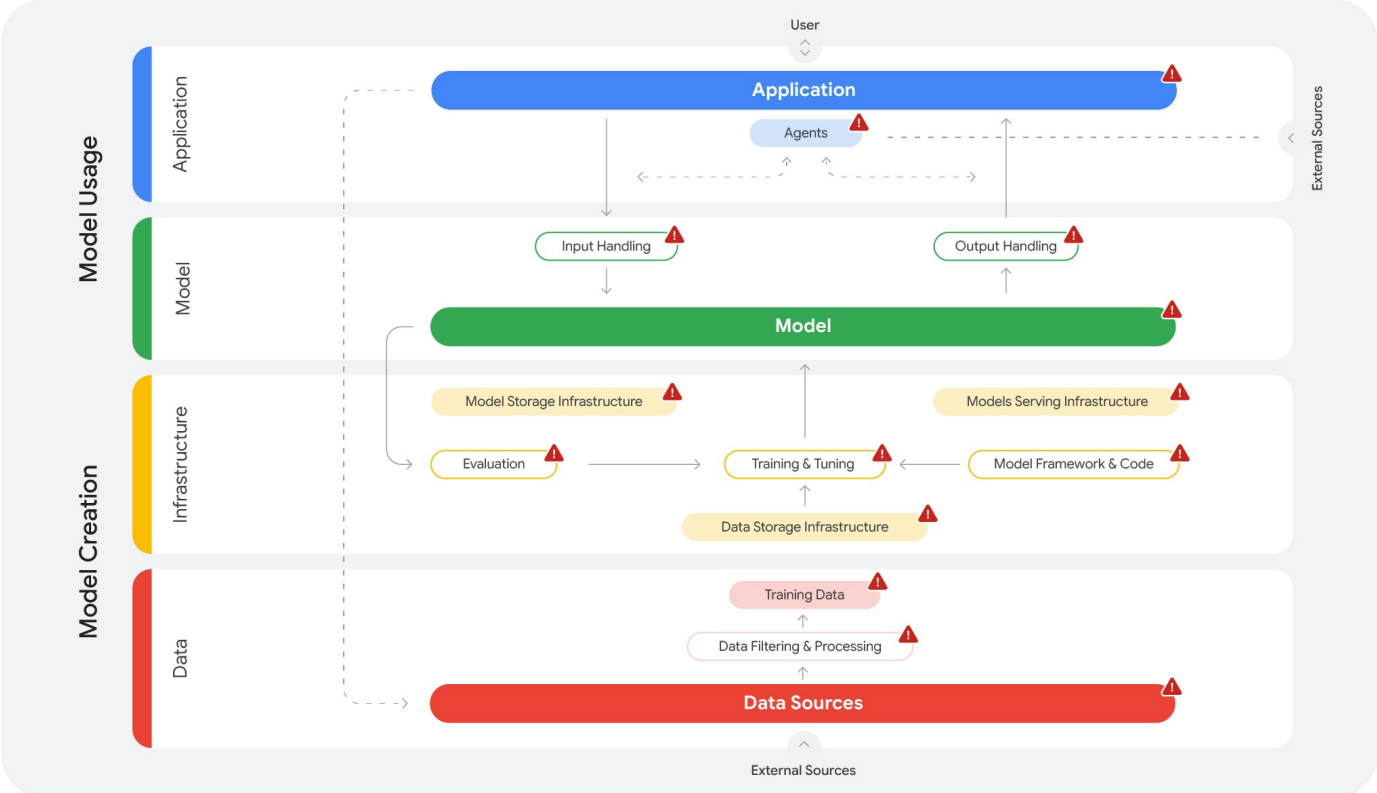
Securing enterprise AI

Generative AI is driving a profound transformation, unlocking unprecedented opportunities for innovation, efficiency, and value creation. At the same time, AI introduces complex new risks to the landscape. AI systems present unique vulnerabilities, such as Data Poisoning, Model Evasion, and Prompt Injection, with increased risk of Sensitive Data Disclosure. Securing these systems is a prerequisite for building trust, ensuring regulatory compliance, and realizing the full potential of AI – all to help provide the confidence needed to deploy AI solutions at scale.

The role Google's Secure AI Framework plays

Google's Secure AI Framework (SAIF) is a framework for securing AI systems throughout their lifecycles. SAIF is designed for practitioners – the security professionals, developers, and data scientists on the front lines – to ensure AI models and applications are secure by design.

SAIF includes the SAIF Risk Map (shown here), a conceptual system architecture for AI based on four components: Data, Infrastructure, Model, and Application. SAIF identifies 15 AI risks (for example, Insecure Model Output and Model Reverse Engineering), highlights where these risks occur, and maps each one against a number of AI controls (for example, Input/Output Validation and Adversarial Training).



SAIF 2.0 extends this framework with guidance on agentic AI risks and controls. Google has contributed SAIF components to the Coalition for Secure AI (CoSAI).

Purpose and structure of this paper

This paper provides a technical guide for implementing technology-agnostic SAIF security controls on Google Cloud, mapping each SAIF control to the specific security capabilities in Google's foundation models (such as Gemini) and the services available in Google Cloud.

Each SAIF control is analyzed in two key areas:

⚙ What we do

We cover the protections that are intrinsically built into Google Cloud's infrastructure, its AI platform services like Vertex AI, and Google's foundation models. You benefit from these safeguards out of the box, with no additional configurations required.

✓ What you should consider

We outline optional services, tools, and configurations that you should implement within your Google Cloud environment to strengthen your security posture in alignment with a specific SAIF control. This approach is based on the shared responsibility model for AI security and provides the actionable guidance needed to build secure and compliant AI systems on Google Cloud.

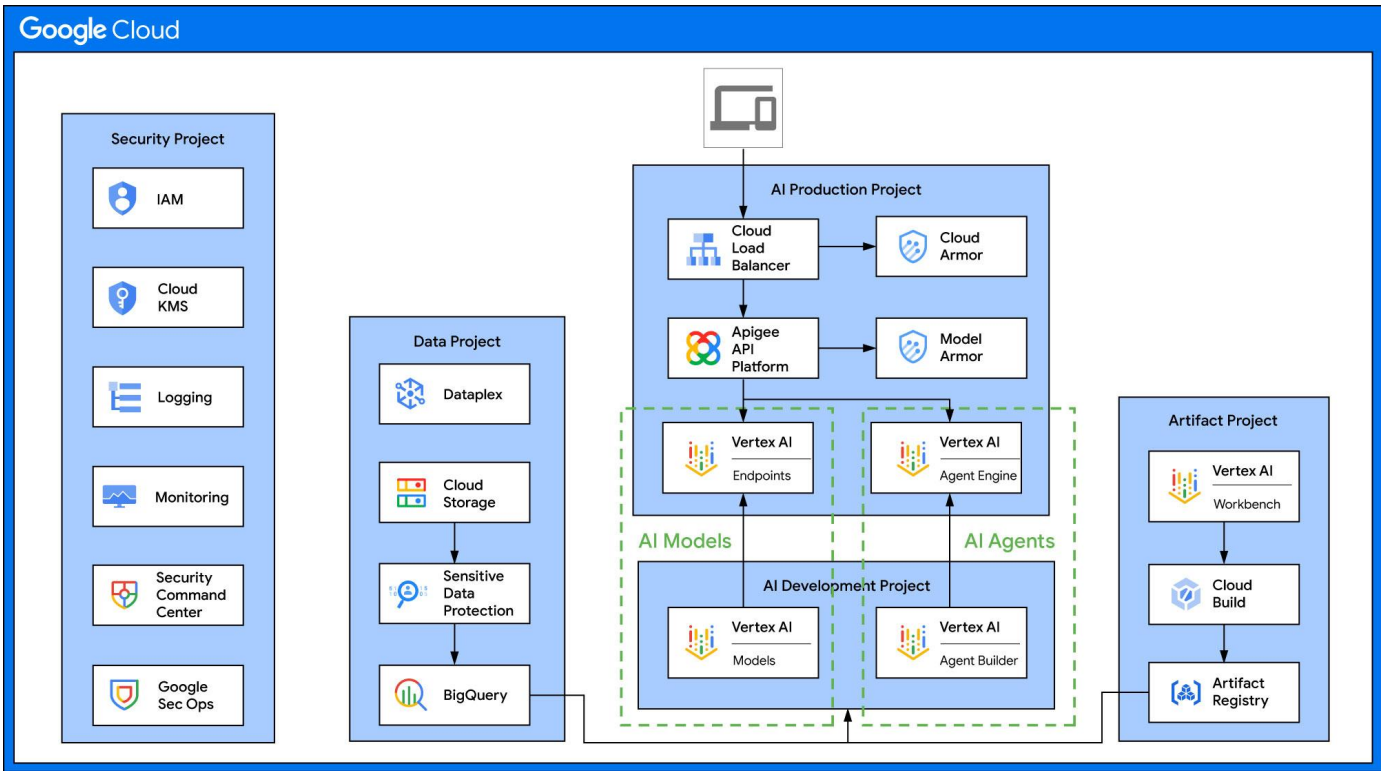
This paper is the latest in a series of Google Cloud publications on AI security, including:

- **Securing AI: Similar or Different?**
- **Best Practices for Securely Deploying AI on Google Cloud**
- **SAIF in the Real World**
- **Delivering Trusted and Secure AI**

Scope and example architecture

This paper is aimed at Google Cloud enterprise customers using Vertex AI to "consume" Google's foundation models (e.g., Gemini) or "create" new models (via fine-tuning or ML training). AI agents created and deployed via Vertex AI Agent Builder, Agent Development Kit, and Vertex AI Agent Engine are considered. AI Studio, Gemini Enterprise, and Gemini consumer applications are not covered in this paper.

While every implementation is unique, the following architecture provides a simplified example of how you can implement SAIF within Google Cloud, highlighting some of the key services and typical data flows. The rest of this paper explores these services in more detail (mapped against each SAIF control).



02 Data Controls

As per SAIF, we start with the data component. Data is the foundation of any AI system, making the security and governance of data pipelines a paramount concern. The SAIF Data Controls address risks in data sourcing, management, and use for model training and user interaction, ensuring privacy, integrity, and authorized use throughout the AI lifecycle.

SAIF control	Description	What we do	What you should consider
Training Data Management	Ensure that all data used to train and evaluate models is authorized for the intended purposes.	- Training data selection for Gemini - Google's privacy principles - Cloud Data Processing Addendum	- IAM - Cloud Storage access controls - BigQuery access controls - Dataplex for data governance - Vertex AI managed datasets
Training Data Sanitization	Detect and remove or remediate poisoned or sensitive data in training and evaluation.	Training data sanitization for Gemini	- Sensitive Data Protection - Dataplex data quality checks
Privacy Enhancing Technologies	Use technologies that minimize, de-identify, or restrict use of PII data in training or evaluating models.	- Federated learning and differential privacy - VaultGemma	- Differential privacy in BigQuery - Federated learning with TensorFlow Federated - Confidential Space
User Data Management	Store, process, and use all user data (e.g., prompts and logs) from AI applications in compliance with user consent.	AI service terms	- IAM - Vertex / Model Armor audit logging - Cloud Storage Object Lifecycle Management - BigQuery table management

Training Data Management

Control definition

This control ensures that all data utilized for training and evaluating models is authorized for its intended purposes. It is implemented by Model Creators and addresses the risks of Inferred Sensitive Data and Unauthorized Training Data.

What we do

Google's AI models (such as the Gemini 2.5 model series) are trained on a large-scale, diverse collection of data, including publicly available web documents, code, images, audio, and video. Google's AI models are developed on our privacy principles. Google Cloud customer data is not used to train our foundation models without the customer's prior permission. The Cloud Data Processing Addendum contractually commits Google to process your customer data only according to your agreement and instructions.

For more information, refer to this paper: Generative AI, Privacy, and Google Cloud.

What you should consider

For customers that are training or fine-tuning AI models, the following features help retain control of training data.

Identity and Access Management: Identity and Access Management (IAM) is the primary authorization mechanism within Google Cloud. Customers should configure granular IAM policies on resources storing training data, such as Cloud Storage buckets and BigQuery datasets, and in Vertex AI. Assigning specific roles to accounts ensures only authorized principals access data for model training.

Cloud Storage access controls: You should use uniform bucket-level IAM controls to protect training data in Cloud Storage. Additional access controls such as signed URLs, public access prevention, and bucket IP filtering can also be considered.

BigQuery access controls: If you are using BigQuery to manage training data, you should, at a minimum, implement IAM access controls at the BigQuery dataset and table level. For more granular access controls, consider row-level security and column-level access control.

Dataplex for data governance: In large enterprises, distributed data makes manual tracking and policy enforcement challenging. Dataplex automates metadata discovery, classification, and enrichment for distributed assets, providing a unified view for governance and a centralized control plane for data-usage policies.

Vertex AI managed datasets: Vertex AI managed datasets offer a centralized, controlled environment for managing ML data lifecycles, facilitating governance, and lineage tracking. Datasets and trained models are versioned, creating an auditable trail for compliance and debugging. Vertex AI also streamlines data preparation and labeling, ensuring consistent, predefined annotations to reduce biases and inaccuracies during training.

Training Data Sanitization

Control definition

This control focuses on detecting and removing or remediating poisoned or sensitive data within training and evaluation datasets. It is implemented by Model Creators and addresses the risks of Data Poisoning and Unauthorized Training Data.

What we do

Google's internal data pipelines for training its foundation models (such as the Gemini 2.5 model series) employ extensive filtering, cleaning, and sanitization techniques to remove harmful, biased, toxic, and sensitive information. This preprocessing provides a sanitized baseline for customers using Google's pretrained models for fine-tuning or inference.

For more information, refer to this Google whitepaper: Our Approach to Protecting AI Training Data.

What you should consider

If you are training or fine-tuning AI models, the following features help with sanitizing training data.

Sensitive Data Protection: Customers can use Sensitive Data Protection to discover, classify, and de-identify sensitive information, such as personally identifiable information (PII), financial, and health data, within training data via de-identification techniques such as redaction, masking, and tokenization. Automating sensitive data removal from large datasets prevents models from memorizing and disclosing private information, aiding privacy compliance while preserving data utility.

Dataplex data quality checks: Data poisoning often appears as statistical anomalies. You can configure automated data quality checks in Dataplex to profile datasets and monitor deviations. For example, rules can alert you if feature value distribution shifts or null values spike. This proactively defends against data poisoning by monitoring training data's statistical health, allowing early detection and remediation before corrupted data compromises a model.

Privacy Enhancing Technologies

Control definition

This control involves using technologies that minimize, de-identify, or restrict the use of personally identifiable information (PII) during the training or evaluation of models. It is implemented by Model Creators and addresses the risk of Sensitive Data Disclosure.

What we do

Google's foundation models are trained using a variety of advanced Privacy Enhancing Technologies. As part of our commitment to Responsible AI, Google has pioneered techniques such as federated learning and differential privacy to train and improve models on distributed datasets while protecting user privacy. For example, Google has recently announced VaultGemma, the largest open model trained from scratch with differential privacy.

What you should consider

If you are training or fine-tuning AI models, Google Cloud offers several advanced Privacy Enhancing Technologies.

Differential privacy in BigQuery: You can apply differential privacy directly to your BigQuery datasets. This feature adds mathematically calibrated statistical noise to aggregation results during data preparation for custom model training. This allows data scientists to extract valuable patterns and statistics about a population without identifying specific individuals, unlocking sensitive datasets for AI model training that might otherwise be inaccessible due to privacy regulations. It provides a quantifiable and provable privacy guarantee, balancing data utility and privacy risk.

Federated learning with TensorFlow Federated: For organizations with decentralized data (e.g., mobile device providers, healthcare consortiums), Google Cloud enables federated learning using the open-source TensorFlow Federated (TFF) framework. In this paradigm, the ML model is sent to the data, not vice versa. The model is trained locally on each client's device or server, and only the resulting model updates (encrypted weights or gradients) are sent back to a central server for aggregation.

Confidential Space: As an advanced implementation of Confidential Computing, Confidential Space provides a trusted execution environment (TEE) where multiple parties can collaboratively analyze data or train an ML model. Each party contributes its encrypted data to the secure enclave, where it is decrypted and processed. Critically, no party can view the raw data of any other participant, unlocking high-value, multiparty AI scenarios that were previously impossible due to trust and confidentiality barriers.

User Data Management

Control definition

This control mandates that all user data, such as prompts and logs from AI applications, is stored, processed, and used in compliance with user consent. It addresses the risks of Sensitive Data Disclosure and Excessive Data Handling.

What we do

Google Cloud's generative AI service terms protect your data, including prompts, responses, and training data. Google Cloud does not use this data to train its own models or for any purpose other than to provide the generative AI service to you, unless you explicitly give your permission. Google Cloud also commits not to store customer prompts for longer than is reasonably necessary to generate the model output.

What you should consider

If you are providing AI services to end users who are customers or employees, you also have a responsibility to ensure any end user data is managed in accordance with your terms and conditions, and any applicable data laws. The following features can help you meet these requirements.

Identity and Access Management provides a foundational capability (as described earlier) to ensure only authorized principals can access user data.

Vertex AI audit logging: Vertex AI does not log user prompts or model responses by default (except for abuse monitoring, which is also configurable). Requests and responses can be logged in Vertex AI for Gemini and supported partner models. Requests and responses can also be sampled to help control log volume.

Model Armor audit logging: Model Armor does not log user prompts or model responses by default. By configuring Model Armor logging, logs are generated when templates are created, deleted, or updated. Prompts and sanitized responses can also be logged.

You can also use features such as Object Lifecycle Management in Cloud Storage or table management in BigQuery to automatically delete data of a specific age.

03 Infrastructure Controls

The security of AI systems is fundamentally dependent on the security of the underlying infrastructure where data is stored, models are trained, and inferences are served. The SAIF Infrastructure Controls address the need to inventory, control access to, ensure the integrity of, and harden the entire technology stack that supports the AI lifecycle.

SAIF control	Description	What we do	What you should consider
Model and Data Inventory Management	Ensure that all data, code, models, and transformation tools used in AI applications are inventoried and tracked.	Google Cloud inventory management	- Cloud Asset Inventory - AI Protection in SCC - Dataplex Universal Catalog - Vertex AI Model Registry - Artifact Registry
Model and Data Access Controls	Minimize internal access to models, weights, datasets, etc. in storage and in production use.	Default encryption (rest/transit)	- IAM - Organization Policy Service - VPC Service Controls - Confidential Computing - Access Transparency and Access Approval
Model and Data Integrity Management	Ensure that all data, models, and code used to produce AI models are verifiably integrity-protected during development and deployment.	Google's secure software development lifecycle	- Cloud Storage data validation - Cloud Storage Object Versioning / soft delete - Assured OSS - Artifact Analysis - Binary Authorization
Secure-by-Default ML Tooling	Use secure-by-default frameworks, libraries, software systems, and hardware components for AI development or deployment to protect confidentiality and integrity of AI assets and outputs.	- Google's secure-by-default infrastructure - Cloud Audit Logs	- Organization Policy Service - Shielded VMs - CMEK / EKM - Secure AI posture templates - Security Health Analytics in SCC

Model and Data Inventory Management

Control definition

This control ensures that all data, code, models, and transformation tools used in AI applications are inventoried and tracked. It is implemented by Model Creators and Model Consumers, and addresses the risks of Data Poisoning, Model Source Tampering, and Model Exfiltration.

What we do

The Google Cloud platform is resource-oriented, with each entity (e.g., Cloud Storage bucket, BigQuery dataset, Compute Engine VM, or Vertex AI custom model) as a distinct asset within a resource hierarchy. All Google Cloud resources are visible and can be monitored and managed via the Google Cloud console, gcloud CLI, and APIs.

This approach provides a basic level of inventory management for all Google Cloud customers.

What you should consider

Google Cloud also provides a number of optional tools to enhance your control over your cloud inventory, including AI assets.

Cloud Asset Inventory: Cloud Asset Inventory is a global metadata inventory service that lets you view, search, export, monitor, and analyze your Google Cloud asset inventory (including resources, policies, and relationships), with up to 35 days of create, update, and delete history, providing a centralized way to track and manage all your cloud resources.

AI Protection in Security Command Center: AI Protection enables customers to discover and secure all AI assets from a central dashboard, helping you to manage your AI security posture by identifying and managing risks, verifying compliance with standards, identifying vulnerabilities, detecting threats, and applying AI security policies at the organization level.

Dataplex Universal Catalog: Customers can leverage Dataplex to create a unified data catalog across distributed data assets. It automatically discovers, classifies, and tags data, providing a comprehensive inventory of all data used by AI systems across your organization (including BigQuery datasets, Cloud Storage filesets, and Vertex AI models). Dataplex creates a unified, searchable inventory that can be enriched with business context, such as data owners, sensitivity labels, and usage guidelines.

Vertex AI Model Registry: Vertex AI Model Registry establishes a formal, governed inventory of all ML models, tracking their versions, evaluation metrics, and deployment status, thereby providing the visibility necessary for effective governance and risk management. Model Registry integrates with Dataplex Universal Catalog (described earlier).

Artifact Registry: Artifact Registry creates a secure, version-controlled inventory for all software components and dependencies used to build AI models and AI agents. This enables consistent access control, vulnerability scanning, and reproducible builds, which are essential for mitigating supply chain risks like Model Source Tampering.

Model and Data Access Controls

Control definition

This control aims to minimize internal access to models, weights, datasets, and similar assets, both in storage and in production use. It addresses the risks of Data Poisoning, Model Source Tampering, and Model Exfiltration.

What we do

Encryption at rest and in transit: All of your data in Google Cloud (e.g., Cloud Storage, BigQuery, and model artifacts in Vertex AI) is encrypted at rest by default using strong cryptographic standards like AES-256. Similarly, all data moving between Google's data centers over our private global network is automatically encrypted in transit. This provides a fundamental and pervasive layer of confidentiality, protecting data from unauthorized access at the physical storage layer and during internal network transit, without requiring any customer action.

What you should consider

Traditional perimeter-based security is insufficient for modern cloud and AI environments. A zero-trust approach based on identity and access management, security policies as code, API security, and audit logging is essential.

Identity and Access Management: You should implement the principle of least privilege using IAM for Vertex AI (and other Google Cloud services, as described previously) to ensure that compromised accounts or applications cannot perform unauthorized actions.

Organization Policy Service: Organization policies implement security guardrails across your entire resource hierarchy in Google Cloud (including Vertex AI). You can enforce policies that restrict available models, limit deployment locations and public access, enforce customer-managed encryption keys (CMEK), specify grounding sources, or create custom constraints.

VPC Service Controls: With VPC Service Controls, customers can define a virtual security perimeter around Google Cloud projects and services handling sensitive AI models and data, blocking any unauthorized data movement to services outside the perimeter. VPC Service Controls add a context-based layer of security on top of IAM, helping to defend against malicious insiders or a compromised VM attempting model or data exfiltration.

Confidential Computing for AI: For highly sensitive AI workloads, Google Cloud's Confidential Computing allows training and inference jobs to run within Confidential VMs or in Confidential GKE (including use of NVIDIA H100 GPUs). These deployments use hardware-level encryption to protect the data and model in memory from access by the host hypervisor or any cloud administrators, providing the strongest level of isolation, confidentiality, and compliance for AI assets.

Access Transparency and Access Approval: Access Transparency and Access Approval provide you with visibility and control over Google personnel actions, enabling customers to audit privileged actions and enforce approval workflows. Access Transparency logs every instance of Google access to customer data for support or operational reasons. Access Approval allows customers to provide explicit approval before access is granted.

Model and Data Integrity Management

Control definition

This control ensures that all data, models, and code used to produce AI models are verifiably integrity-protected during development and deployment. It addresses the risks of Data Poisoning and Model Source Tampering.

What we do

Google Cloud's multilayered security protections provide a strong defense against many possible threats to model and data integrity. For more information, refer to **Secure-by-Default ML Tooling** below.

Additionally, Google Cloud's **secure software development lifecycle** helps to ensure the integrity of Google's own AI platforms and models while also providing a secure foundation for the integrity management of customer models and data. For more information, refer to the following resources:

- **Google's safe software development process**
- **Google's software build and deployment process**

What you should consider

There are a number of security features that you can use to provide additional integrity management over your AI resources.

Cloud Storage data validation: Cloud Storage automatically ensures data integrity when it is copied or rewritten. Customers can supply an expected MD5 or CRC32C hash during upload for verification. Similarly, Google can provide server-side hashes for downloaded verification.

Cloud Storage Object Versioning and soft delete: Object Versioning and soft delete provide additional safeguards against tampering for any Cloud Storage buckets that contain critical AI assets. Object Versioning creates a new version of any object that's modified or deleted, instead of overwriting the original. And soft delete provides additional protection against accidental or malicious deletion of a bucket. If a dataset is poisoned or a model file corrupted, administrators can easily revert to a known-good version, ensuring AI pipeline integrity.

Assured OSS: Assured OSS provides a Google Cloud curated and vetted repository of common open-source packages. For AI applications using open source, Assured OSS protects component integrity through provenance verification, security scanning, secure building, and signing, helping to prevent vulnerabilities and unauthorized modifications.

Artifact Analysis: When packaging AI models in containers, enabling Artifact Analysis on Artifact Registry repositories will automatically scan uploaded container images for known vulnerabilities in their OS packages and application language dependencies. By detecting vulnerable libraries before deployment, this helps prevent the introduction of components that could be exploited to tamper with the model or exfiltrate data.

Binary Authorization: Customers can use Binary Authorization to enforce deployment integrity. This service integrates with Google Kubernetes Engine (GKE) to create a deployment-time policy, only allowing the deployment of container images that have been signed by a trusted authority and scanned by Artifact Analysis with no critical vulnerabilities found. This helps create a strong, enforceable chain of trust from development to production.

For more information about **Model Signing**, you can also refer to this CoSAI paper (co-authored by Google): Signing ML Artifacts.

Secure-by-Default ML Tooling

Control definition

This control advocates for the use of secure-by-default frameworks, libraries, software systems, and hardware components for AI development or deployment to protect the confidentiality and integrity of AI assets and outputs.

What we do

Google's secure-by-default infrastructure: The entire Google Cloud platform is built on a foundation of security. This foundation includes physically secure data centers, custom-designed server hardware with security chips like Titan, a hardened and minimal base OS, and default encryption at rest and in transit. Users of any Google Cloud service, including Vertex AI, inherit the security benefits of this vertically integrated and hardened stack.

Cloud Audit Logs: Audit Logs are enabled by default[1] throughout Google Cloud, providing a secure audit trail of who did what, where, and when.

✓ What you should consider

Google Cloud provides a range of tooling to further enhance secure defaults.

Organization Policy Service: As described earlier, organization policies help you enforce secure defaults across your Google Cloud environment.

Shielded VMs: Shielded VMs extend Google Cloud's inherent infrastructure security to individual VMs used for Compute Engine instances or as Shielded GKE Nodes. Shielded VMs use a Virtual Trusted Platform Module derived from Google's Titan security chip to establish a secure root of trust to verify VM identity and help to ensure VM integrity via Secure Boot and Measured Boot. This can be enforced via organization policies.

Customer-managed encryption keys: Customer-managed encryption keys give you full control over the cryptographic keys used to protect your AI assets, helping to meet compliance requirements and providing a powerful tool for revoking access if necessary. Customers can choose to store these keys in a software-based Cloud KMS, a hardware based Cloud HSM, or outside of Google Cloud via Cloud External Key Manager. This can be enforced via organization policies.

Secure AI posture templates: Security Command Center supports secure AI posture templates (essentials and extended) that define a set of preventative and detective security policies that apply to Vertex AI workloads (including instances of Vertex AI Workbench, a managed Jupyter notebook environment for data scientists).

Security Health Analytics in Security Command Center: Security Health Analytics is a managed service that scans your Google Cloud environments for over 100 types of common misconfigurations and vulnerabilities, helping you to stay secure by default.

1. Admin Activity, System Event, and Policy Denied logs enabled by default. Data Access logs are optional.

04 Model Controls

Once data and infrastructure are secured, the focus shifts to the model itself. Models can be attacked directly through malicious inputs (prompts), or their outputs can be manipulated to cause harm. The SAIF Model Controls are designed to build resilience into the model and sanitize its inputs and outputs to protect against these emerging threats.

SAIF control	Description	What we do	What you should consider
Input Validation and Sanitization	Block or restrict adversarial queries to AI models.	Gemini default input validation	- Model Armor - Gemini as a guard model - Apigee API management
Output Validation and Sanitization	Block, nullify, or sanitize insecure output from AI models before passing it to applications, extensions, or users.	- Default safety filters - Citation filters - Indemnification	- Vertex AI configurable safety filters - Vertex AI system instructions - Model Armor - Grounding in Vertex AI
Adversarial Training and Testing	Use techniques to make AI models robust to adversarial inputs (i.e., prompts) in the context of their use in applications.	- Internal adversarial testing and training - AI safety research	- Responsible Generative AI Toolkit - Model evaluation in Vertex AI - Vertex AI Pipelines

Input Validation and Sanitization

Control definition

This control involves blocking or restricting adversarial queries to AI models. It can be implemented by Model Creators and Model Consumers, and it addresses the risk of Prompt Injection.

What we do

Google's foundation models, like Gemini, have undergone extensive safety tuning and alignment. This process includes training the model to be more robust against adversarial inputs, and to recognize and refuse to act on prompts that violate safety policies. This provides a baseline level of resilience against common prompt injection and jailbreaking techniques. For more information, refer to the Gemini 2.5 technical report.

What you should consider

Google Cloud provides a number of optional features to help you implement input validation and sanitization.

Model Armor is a Google Cloud service designed to act as an LLM firewall for both first-party and third-party models, providing highly configurable runtime protection for AI applications. It screens both incoming prompts and outgoing responses for various threats. For input validation, its key feature is **LLM-AI model threat detection**, which identifies and blocks sophisticated prompt injection and jailbreaking techniques. It can also detect and block malicious URLs and malware embedded in prompts, and provides fine-grained control of harmful, unethical, or undesirable content, such as hate speech, harassment, sexually explicit material, and dangerous topics. Model Armor is integrated with Sensitive Data Protection to screen personally identifiable information and other sensitive data types, such as financial information.

Gemini as a guard model: You can also use Gemini as a guard model to implement input filtering. Using a second, faster model like Gemini Flash-Lite, an application can first evaluate a user's prompt for safety policy violations or malicious intent before passing it to the primary, more powerful model. This dual-model approach allows for highly customizable and robust filtering of user input that's tailored to the specific risks of the application.

Apigee API management: When exposing an AI model via an API, Apigee can serve as a sophisticated API gateway to inspect incoming API requests, including the prompt data within the payload. Policies can be applied to check for known malicious patterns, limit request sizes and rates to prevent resource-exhaustion attacks, and enforce authentication and authorization before the request ever reaches the model endpoint.

Output Validation and Sanitization

Control definition

This control focuses on blocking, nullifying, or sanitizing insecure output from AI models before it is passed to applications, extensions, or users. It addresses the risks of Prompt Injection, Rogue Actions, Sensitive Data Disclosure, Inferred Sensitive Data, and Insecure Model Output.

What we do

Non-configurable safety filters in Gemini: Google's Gemini models have built-in, non-configurable safety filters that are always active. These filters are designed to block the generation of the most harmful types of content and prevent the disclosure of certain types of Sensitive Personally Identifiable Information.

Citation filters: Gemini models also include citation filters to limit the replication of existing content at length. If a Gemini feature does make an extensive quotation from a web page, Gemini cites that page.

Indemnification: As per Google Cloud's service terms, Google offers indemnification (under defined conditions) against allegations that either training data or generated output from a Generative AI Indemnified Service infringes upon a third-party's intellectual property rights.

What you should consider

As with input filtering, Google Cloud provides a number of optional features to help you implement output validation and sanitization, in addition to the baseline described earlier.

Vertex AI configurable safety filters: Vertex AI gives you granular control over the safety policies applied to first-party model responses. You can set blocking thresholds for several harm categories (including hate speech, harassment, sexually explicit content, and dangerous content) with configurable thresholds. This allows your organization to align the model's output safety with your specific safety standards, user base, and risk tolerance.

Vertex AI system instructions: Vertex AI also supports system instructions, which directly steer the model's behavior, to augment or replace safety filters. By providing clear and specific instructions, you can help the model generate responses that are safe and aligned with your policies.

Model Armor inspects all model-generated responses before they are sent to the user. It applies granular content-safety filters, a Sensitive Data Protection engine, and malicious URL detection to the output. This provides a second layer of defense for first-party models and can also be used with third-party models, catching harmful or sensitive content that might have been generated despite the model's safety training.

Grounding in Vertex AI: Vertex AI provides various options to ground model output based on Google Search, Google Maps, or other enterprise data. Grounding reduces model hallucination and anchors model responses to wider data sources, providing additional auditability and explainability for model outputs.

Adversarial Training and Testing

Control definition

This control uses techniques to make AI models robust to adversarial inputs (prompts) in the context of their use in applications. It addresses risks like Model Evasion, Prompt Injection, and Sensitive Data Disclosure.

What we do

Internal adversarial testing and training: The development process for Google's foundation models includes extensive adversarial testing and training. This involves dedicated AI Red Teams and researchers who systematically attempt to "break" the model by crafting inputs designed to elicit harmful, biased, or unsafe responses. The insights from these tests are then used to further train and fine-tune the model, making it inherently more robust against such attacks. This continuous feedback loop is a key component of Google's approach to AI safety. For more details, refer to the Gemini 2.5 technical report.

AI safety research: Google actively researches and publishes its findings on AI safety topics, such as adversarial testing, defense strategies, and AI security principles.

What you should consider

Google provides tools and features to help you implement your own adversarial testing – for example, if you are training or fine-tuning your own AI models.

Responsible Generative AI Toolkit: Google provides a public Responsible Generative AI Toolkit that includes guidance and best practices for adversarial testing. This toolkit helps you to identify potential AI failure modes, create diverse and representative test datasets, evaluate model outputs, and mitigate AI failures through fine-tuning or model safeguards.

Model evaluation in Vertex AI: The Vertex AI platform provides tools for generative AI model evaluation that can be used to assess a model's performance against static or adaptive rubrics, or computation-based metrics. These evaluation metrics can be tracked over time and across different model versions in the Vertex AI Model Registry. Vertex also supports automatic side-by-side evaluation of generative AI models in the Vertex AI Model Registry and tooling to evaluate potential bias in AI models.

Vertex AI Pipelines: Vertex AI Pipelines allows you to define your machine learning workflow as a graph of components, helping you establish automated, repeatable, and auditable workflows. These include test generation using adversarial attack libraries and pretrained models as inputs, test execution, and test evaluation, which logs results to provide safety metrics. Based on the evaluation results, the pipeline can trigger a retraining process to make the model more resilient, including learning from adversarial examples.

05 Application Controls

Application controls secure the interface between the end user and the AI model. They ensure that only authorized users can access the AI application, that users are aware of the AI's capabilities and limitations, and that any actions taken by AI agents on a user's behalf are properly controlled and permissioned.

SAIF control	Description	What we do	What you should consider
Application Access Management	Ensure that only authorized users and endpoints can access specific resources for authorized actions.	- Default firewall rules - Default rate limits - Default DoS protection	- Identity-Aware Proxy - Apigee Advanced API Security - Cloud NGFW - Cloud Armor - ReCAPTCHA
User Transparency and Controls	Inform users of relevant AI risks with disclosures, and provide transparency and control experiences for use of their data in AI applications.	- Model cards - Technical reports - AI service terms	- Google Cloud Privacy Notice - OAuth consent screens - Gemini thinking models - Vertex Explainable AI
Agent Permissions	Use least-privilege principle as the upper bound on agentic system permissions to minimize the number of tools that an agent is permitted to interact with and the actions it is allowed to take.	Vertex AI Agent Builder / Agent Engine (default security)	- IAM for agent service accounts - Connector security - MCP/A2A security best practices
Agent User Control	Ensure user approval for any actions performed by agents/plugins that alter user data or act on the user's behalf.	Vertex AI Agent Builder / Agent Engine (default security)	- Front-end authorization - Identity propagation - MCP/A2A authorization - IAM for Google Cloud services
Agent Observability	Ensure an agent's actions, tool use, and reasoning are transparent and auditable through logging, allowing for debugging, security oversight, and user insights into agent activity.	Vertex AI Agent Builder / Agent Engine (native integration with Cloud Operations)	- Cloud Monitoring in Agent Engine - Cloud Logging in Agent Engine - Cloud Trace in Agent Engine

Application Access Management

Control definition

This control ensures that only authorized users and endpoints can access specific resources for authorized actions. It is implemented by Model Consumers and addresses the risks of Denial of ML Service and Model Reverse Engineering.

What we do

Google Cloud's foundational software-defined network provides robust, built-in defenses against network-level attacks.

Default firewall rules: All networked resources within Google Cloud benefit from an SDN-based Cloud Firewall, which, by default, blocks all ingress traffic[2] and allows all egress traffic. You can enhance your security posture by adding a deny rule for all egress traffic, then allowing only what's required.

Default rate limits: Google limits the packet rate and bandwidth that can reach any VM based on machine type.

Default DoS protection: Inherent DoS protection is always active for services deployed behind Google's global external load balancers. This default protection primarily targets network and transport layer attacks, such as SYN floods, UDP floods, and IP fragmentation attacks.

What you should consider

We recommend you consider the following optional features for enhanced application access security.

Identity-Aware Proxy (IAP) provides centralized authentication and authorization for user access to web applications, enabling a zero-trust access model. Instead of relying on network-level firewalls, IAP enforces access policies based on user identity and context. IAP first authenticates the user via their Google identity (or external identity), and then checks if they have the necessary IAM role. Access levels, created within the Access Context Manager, specify additional device requirements like source IP range and geographic location.[3]

Apigee (including Advanced API Security): For AI applications exposed via APIs, Apigee can be used to manage and secure access. Apigee acts as the front door for all API calls, enforcing authentication (e.g., via API keys, OAuth 2.0, or JWTs), rate limiting to prevent abuse, and routing requests to the appropriate backend AI model. Beyond basic API management, Apigee Advanced API Security continually monitors your APIs to protect them from security threats, including attacks from malicious clients and abuse. Customers can identify, block, or flag suspicious API requests.

2. In the default network of each VPC, default firewall rules allow SSH, RDP, and ICMP ingress.
3. With Chrome Enterprise Premium, you can extend this capability to require specific device attributes, such as OS version, screen lock, storage encryption, and certificate.

Cloud NGFW: Beyond the default capability described earlier, Cloud Next Generation Firewall provides a highly flexible suite of firewall capabilities that can be customized according to your needs.

- NGFW Essentials supports VPC firewall rules, address groups, tag integration, and global and regional firewall policies.
- NGFW Standard adds FQDN support, geolocation filtering, and the integration of Google Threat Intelligence.
- NGFW Enterprise adds intrusion detection and prevention, as well as Transport Layer Security decryption.

Cloud Armor provides enterprise-grade DDoS protection and web application firewall capabilities to protect the availability and integrity of the AI application. It includes preconfigured rules to defend against common web attacks like SQL injection and cross-site scripting. It can also filter traffic based on IP addresses, geographies, threat intelligence, request headers, rate limits, and more. Cloud Armor also integrates with reCAPTCHA for bot management and includes Adaptive Protection backed by machine-learning models to help protect against L7 DDoS attacks.

ReCAPTCHA: ReCAPTCHA Enterprise acts as an intelligent gatekeeper that protects your AI applications from fraudulent and automated access. It uses an adaptive risk analysis engine to differentiate between human users and bots in the background, helping to ensure only legitimate human users are interacting with your models and applications.

User Transparency and Controls

Control definition

This control involves informing users of relevant AI risks through disclosures and providing transparency and control experiences for the use of their data in AI applications. It addresses the risks of Sensitive Data Disclosure and Excessive Data Handling.

What we do

Google is committed to responsible AI development and provides extensive public documentation on its AI Principles, privacy policies, and the capabilities and limitations of its models. This provides a foundation of transparency upon which you can build.

Model cards: Model cards are simple, structured overviews of how an advanced AI model was designed and evaluated, providing a standardized way to communicate a model's characteristics to various stakeholders. They may also contain guidance on model usage and limitations, terms of use, and information about the data used to train the model. Google provides model cards for its own first-party models. Google Cloud also provides model cards for third-party models hosted in the Vertex AI Model Garden.

Technical reports: In addition to model cards, Google publishes detailed technical reports about its first-party models (for example, this Gemini 2.5 technical report) covering topics such as model architecture, training, datasets, evaluation, safety, security, and responsibility.

AI service terms: Google also publishes service terms and conditions for its generative AI services in Google Cloud (note that these terms are different from those that apply to consumer use of Google's generative AI, for example via the Gemini App).

✓ What you should consider

Any application using AI must have a clear, easily accessible privacy policy that includes an explanation of what user data is collected, how it is used by the AI system, and how long it is retained. The application must also obtain explicit user consent and/or provide disclosures about the use of AI where required by regulations. You can refer to the Google Cloud Privacy Notice for an example of how Google provides these notifications and consider using OAuth consent screens to display your privacy policy when using OAuth in Google Cloud for user authorization.

Application developers should also consider how they provide transparency and explainability of the AI capabilities they are exposing. There are several features that can help here, including grounding (described earlier), thinking models, and Explainable AI.

Gemini thinking models: The Gemini 2.5 series models use an internal "thinking process" that improves their reasoning and multistep planning abilities. It also improves transparency because thought summaries can be included in the model output to offer insights into the model's internal reasoning.

Vertex Explainable AI: For predictive models, Vertex Explainable AI can provide feature attribution capabilities – to understand *why* the model made a prediction – by highlighting how much each input feature contributed to the final result. This helps demystify the "black box" nature of some ML models. Providing these explanations to users builds trust, aids in debugging, and helps ensure the model is making decisions based on the correct factors.

Agent Permissions

Control definition

Use the least-privilege principle as the upper bound on agentic system permissions to minimize the number of tools that an agent is permitted to interact with and the actions it is allowed to take. It is implemented by Model Consumers and addresses the risks of Insecure Integrated Component, Rogue Actions, and Sensitive Data Disclosure.

⚙ What we do

Vertex AI Agent Builder is a platform that helps developers build and orchestrate secure, enterprise-grade, multi-agent AI experiences. You can create sophisticated agents and multi-agent workflows using the Agent Development Kit (ADK) or leverage other popular open-source frameworks. It allows agents to connect to enterprise systems and data through prebuilt connectors, custom APIs, and retrieval-augmented generation (RAG) capabilities.

In addition to supporting model access, agent memory, and sessions management, ADK has built-in support for the Model Context Protocol (MCP) to enable agent integration with backend tools and the Agent2Agent Protocol (A2A), which enables agents built on different frameworks or by different vendors to communicate and collaborate.

Vertex AI Agent Engine provides a fully managed runtime to handle the complexities of infrastructure, scaling, security, and monitoring for your agents.

As with other Google Cloud services, these platforms inherit the benefits of Google Cloud's infrastructure security and security features, such as IAM, VPC Service Controls, customer-managed encryption keys, data residency, Access Transparency, and audit logging.

✓ What you should consider

IAM for agent service accounts: An agent with least-privilege permissions can only perform a very limited set of actions, significantly reducing the potential for damage if the agent is compromised or behaves unexpectedly.

Connector security: Vertex AI Agent Builder supports connectors to external tools and data stores. Before any data can be accessed by an agent, the connector must be **authenticated** (with stored credentials or via a Service Account) and **authorized** (to determine the permissions/scope of the connector). Google's Agent Development Kit provides support and documentation for agent authentication with backend tools, using methods such OAuth 2.0, OpenID Connect (OIDC), Google Cloud Service Accounts, and API keys. Customers can also connect to enterprise systems via Integration Connectors.

MCP/A2A security best practices: Customers should leverage the A2A enterprise-ready security features and implement MCP security best practices.

Agent User Control

Control definition

This control ensures user approval for any actions performed by agents or plugins that alter user data or act on the user's behalf. It is implemented by Model Consumers and addresses the risk of Rogue Actions and Sensitive Data Disclosure.

What we do

The frameworks for building agents on Google Cloud, such as Vertex AI Agent Builder, are designed with security in mind, but the primary responsibility for implementing user consent for actions lies with the application developer.

What you should consider

As described in Google's Agent Development Kit guidelines on safety and security, developers of AI agents must carefully consider whether interactions with backend tools should be authorized with the agent's own identity ("Agent Auth") or with the identity of the controlling user ("User Auth"). Agent Auth is appropriate for agents where all users have the same permissions, while User Auth is required for agents with multiple users that have different permissions in the backend tools.

Implementing User Auth requires the user identity to be propagated through to the backend tools that are invoked by the agent, to ensure those users can only access authorized data and can only implement authorized actions. This is implemented in several steps.

Front-end authentication: The process starts with your client application (a web app, mobile app, or chatbot) authenticating the user via standard methods like OAuth 2.0 and OIDC.

Identity propagation: The authenticated client sends the user's OAuth 2.0 access token to the agent in the Authorization Bearer header of the API call.

MCP/A2A authorization: The agent is responsible for validating this token and passing it on to backend tools. Refer to additional documentation on MCP authorization and A2A authorization.

IAM for Google Cloud services: AI Agent Builder also supports agent integrations with Google Cloud services, such as BigQuery, Spanner, Bigtable, and GKE. In this case, the final API call is made using the user's own Google Cloud credentials. Google Cloud's IAM automatically enforces that user's specific permissions.

Agent Observability

Control definition

Ensure an agent's actions, tool use, and reasoning are transparent and auditable through logging, allowing for debugging, security oversight, and user insight into agent activity. It is implemented by Model Consumers and addresses the risk of Rogue Actions and Sensitive Data Disclosure.

What we do

Vertex AI Agent Builder and Vertex AI Agent Engine are designed with observability in mind, integrating seamlessly with Google Cloud's operations suite (Cloud Monitoring, Cloud Logging, and Cloud Trace) to provide a comprehensive view of your agents' activities. Vertex AI Agent Builder also supports Audit Logs (enabled by default), to record administrative activities. The detailed monitoring, logging, and tracing capabilities provide the necessary transparency to understand and control your AI agents, ensuring they operate securely and as intended.

What you should consider

The observability features in Vertex AI Agent Builder can be customized to your needs.

Cloud Monitoring: Vertex AI Agent Engine automatically sends various metrics to **Cloud Monitoring**. This allows you to create dashboards and alerts for key indicators.

- **Request and error rates:** Track the volume of requests your agent is handling and identify any unusual spikes in errors that could indicate a security event or a performance problem.
- **Latency:** Monitor the response times of your agent to ensure it is meeting performance expectations and to detect any slowdowns that might be caused by malicious activity.
- **Resource utilization:** Monitor CPU and memory resources to optimize costs and detect any unexpected usage patterns.

You can work with built-in metrics or custom metrics generated by Vertex AI Agent Engine and define alerts in Cloud Monitoring.

Cloud Logging: Every interaction and operation of your agent can be logged in detail in Cloud Logging. This provides an invaluable audit trail for security investigations and for understanding the step-by-step behavior of your agent. Customers can enable Cloud Logging for agents in Vertex AI Agent Engine. For additional information, refer to Logging in the Agent Development Kit.

Cloud Trace: For a granular view of your agent's performance, **Cloud Trace** provides distributed tracing. This allows you to follow a request as it travels through your agent and its various components, helping you to pinpoint bottlenecks and understand the flow of execution. This is particularly useful for debugging complex agent behaviors and identifying performance degradation that could have security implications. Customers can enable Cloud Trace for agents in Vertex AI Agent Engine. For additional information, refer to Agent Observability with Cloud Trace in ADK.

06 Assurance Controls

Assurance Controls provide the mechanisms for continuously testing, monitoring, and responding to security threats against AI systems. They are universal controls that apply across the entire AI lifecycle, providing the feedback loops necessary to maintain and improve your security posture over time.

SAIF control	Description	What we do	What you should consider
Red Teaming	Identify security and privacy improvements through self-driven adversarial attacks on AI infrastructure and products.	Google AI Red Team	- Mandiant AI Red Team Assessment - Virtual red teaming in SCC
Vulnerability Management	Proactively and continually test and monitor production infrastructure and products for security and privacy regressions.	- Google internal vulnerability management - Automated patching in Google Cloud	- AI Protection in SCC - Web Security Scanner in SCC - Artifact Analysis - VM Manager - GKE auto-upgrade
Threat Detection	Detect and alert on internal or external attacks on AI assets, infrastructure, and products.	- Google internal security monitoring - Google Threat Intelligence	- Cloud Audit Logs - Threat Detection in SCC - Google Security Operations (SIEM) - Google Threat Intelligence
Incident Response Management	Detect and alert on internal or external attacks on AI assets, infrastructure, and products.	- Google incident response - Service Health dashboard	- Cloud Logging / Monitoring - Google Security Operations (SOAR) - Mandiant Incident Response Services

For more information on all the Assurance Controls, refer to Google's research paper: [Security Assurance in the Age of Generative AI](#).

Red Teaming

Control definition

This control involves identifying security and privacy improvements through self-driven "adversarial" attacks on AI infrastructure and products.

What we do

Google maintains a dedicated, internal AI Red Team composed of highly skilled offensive security experts. This team's primary mission is to proactively identify and mitigate vulnerabilities in Google's own AI systems, including foundation models like Gemini. They conduct sophisticated adversarial testing, simulating real-world attacks to uncover weaknesses related to safety, security, and privacy before products are released.

What you should consider

Google Cloud supports customer AI red teaming in the following ways.

Mandiant Red Team Assessment: If you are testing the security of your own custom AI applications, Mandiant offers expert Red Team Assessment services. These engagements simulate a realistic, persistent attack by mimicking the tactics, techniques, and procedures of known threat actors. The assessment can be tailored with custom objectives, such as attempting to exfiltrate a specific model or poison a training dataset, to test the effectiveness of your specific security controls and allow security teams to prioritize and fix the most critical vulnerabilities.

Virtual red teaming in Security Command Center: The Enterprise tier of Security Command Center includes a "virtual red teaming" capability. This feature creates a digital twin model of the customer's cloud environment and runs millions of simulated attack permutations against it. It can discover complex attack paths and toxic combinations of misconfigurations that could lead to the compromise of AI assets. For more information, refer to:

- **Toxic combinations**
- **Attack exposure scores and attack paths**

Vulnerability Management

Control definition

This control focuses on proactively and continually testing and monitoring production infrastructure and products for security and privacy regressions.

What we do

Google's internal vulnerability management program is a comprehensive, continuous process that involves a combination of in-house and commercial scanning tools, manual and automated penetration testing, and software security reviews. This program covers the entire Google Cloud infrastructure, and its findings are used to continuously harden the platform.

Google also implements automated security patching for both the underlying infrastructure of Google Cloud and for Google Cloud managed services, such as Vertex AI, Cloud Storage, and BigQuery. This is a transparent process that requires no customer action. It helps to ensure that any vulnerabilities are rapidly patched.

What you should consider

Google Cloud provides a number of tools to help you with vulnerability management.

AI Protection in Security Command Center: AI Protection enables you to manage your AI security posture by identifying vulnerabilities relevant to all AI assets within Google Cloud. It addresses those vulnerabilities by applying configuration and policy changes at the organization level.

Web Security Scanner in Security Command Center: Web Security Scanner crawls web applications hosted in Compute Engine, GKE, and App Engine environments to identify security vulnerabilities and misconfigurations, providing another layer of defense and vulnerability management for customers hosting web-based AI applications.

Artifact Analysis: As described previously, Artifact Analysis automatically scans container images in Artifact Registry for known vulnerabilities in OS and language packages. These findings are surfaced directly in Security Command Center, providing visibility into vulnerabilities within the software supply chain of AI applications.

VM Manager: For Compute Engine, management of the operating system (package management and patching) is a customer responsibility. VM Manager is a suite of tools that helps customers automate OS inventory management, configuration management (installing, removing and auto-updating software packages), and patch management of the OS itself. This tool helps customers stay secure by automating the deployment of security patches, reducing their operational burden.

GKE auto-upgrade: For GKE, patching is a shared responsibility. Google automatically patches the GKE control plane. For the worker nodes, Google provides patched images. Customers can enable node auto-upgrades through release channels to automate the process within configured maintenance windows.

Threat Detection

Control definition

This control involves detecting and alerting on internal or external attacks on AI assets, infrastructure, and products.

What we do

Google's global network is protected by sophisticated threat-detection systems that analyze massive volumes of data to identify and block malicious activity. DoS attacks are blocked at the edge, leveraging Google's massive network capacity. Anomaly detection at the machine level is linked to automated self-defences that can immediately isolate machines, remove sensitive data, and relocate workloads.

Our threat-detection capability is further augmented by Google Threat Intelligence, which we use internally to inform and guide our own security operations and to provide threat intelligence as a service to our customers.

What you should consider

Under the shared responsibility model, you are responsible for monitoring the security of your AI applications running in Google Cloud. However, Google provides a rich set of tools to help with this.

Cloud Logging: You should ensure that appropriate logs are enabled and configured to capture all relevant activity. Several types of Cloud Audit Logs are enabled by default (Admin Activity, System Event, and Policy Denied). Other configurable log types to evaluate include Data Access audit logs, firewall logs, VPC Flow Logs, and Cloud DNS Logging. These logs provide the raw data for threat detection. Logs can be routed to BigQuery or Google SecOps for further analysis.

Threat Detection in Security Command Center: Security Command Center continuously monitors your cloud environment and can identify and alert you to a wide range of threats. This is a vital component of any threat-detection strategy in Google Cloud. For more details, refer to:

- **Event Threat Detection**
- **VM Threat Detection**
- **Container Threat Detection**
- **Cloud Run Threat Detection**

Google Security Operations – SIEM: Google SecOps includes a cloud-native Security Information and Event Management (SIEM) capability. It is designed to ingest and analyze security telemetry at petabyte scale from across your environment, both in the cloud and on-premises. It uses Google Threat Intelligence and agentic AI to detect anomalous patterns and indicators of compromise that may signify an attack on your infrastructure (including your AI systems) and to recommend and automate a security response.

Google Threat Intelligence: Google's unique threat intelligence capability combines frontline intelligence derived from Mandiant's Incident Response Services and security research, over 20 years of crowdsourced threat data from VirusTotal, and wider security insights derived from Google's global business (including Search, Chrome, and Android). This provides unmatched visibility into who is targeting you, the techniques they are using, and how you can defend against the most relevant threats to your business.

Incident Response Management

Control definition

This control manages the response to AI security and privacy incidents.

What we do

Google maintains a 24/7 global team of experts to manage incidents that directly impact our infrastructure. Data incidents are managed according to a rigorous, well-defined process that covers detection, triage, coordination, resolution, and postmortem analysis to ensure that incidents are handled swiftly and effectively, and that lessons are learned to improve defenses.

Google Cloud also maintains the Service Health dashboard, which allows customers to rapidly assess the status of Google Cloud services and to view any current or recent incidents.

What you should consider

As with threat detection, incident response for your AI applications deployed to Google Cloud is your responsibility. Google Cloud offers various services to support you.

Cloud Logging and Cloud Monitoring: When a security incident is detected, the first step is to gather additional information. Google Cloud Logging provides a central repository of historical information for all cloud services, while Cloud Monitoring provides an up-to-date view on the performance, availability, and health of your applications and infrastructure. Both Cloud Logging and Cloud Monitoring are highly configurable services, therefore you should review which logs are collected and which metrics should be monitored before an incident occurs. Learning from historical incidents, including any gaps in visibility, will also ensure you are better prepared for future incidents.

Google Security Operations – SOAR: The Security Orchestration, Automation, and Response (SOAR) capabilities within Google SecOps allow you to automate and orchestrate your incident response workflows. Security teams can build playbooks that automate common response actions, such as isolating a compromised VM, revoking a user's credentials, or blocking a malicious IP address.

Mandiant's Incident Response Services: If you experience a serious security incident involving your AI systems, you can engage Mandiant's Incident Response Services. Mandiant's world-renowned team of elite cybersecurity experts will provide hands-on support to investigate the breach, contain the threat, and eradicate the attacker's presence, helping you to minimize damage from an attack and recover operations quickly.

07 Governance Controls

Governance Controls establish the policies, procedures, and oversight necessary to manage AI security and risk at an organizational level. They ensure that technical controls are aligned with business objectives so that employees and users understand their responsibilities, and that your organization maintains a proactive and compliant security posture.

SAIF control	Description	What we do	What you should consider
Product Governance	Validate that all AI models and products meet the established security and privacy requirements.	- Google AI governance - ISO/IEC 42001 / NIST AI RMF - EU AI Act - Gemini / Google Cloud certifications	- Data residency - Assured Workloads / Data Boundary - Vertex AI Model Registry - Binary Authorization
Risk Governance	Inventory, measure, and monitor residual risk to AI in your organization.	- Google AI risk management - Google Cloud Trust Center	- SAIF Risk Self-Assessment - Risk Management in SCC - Compliance resource center - Cyber Insurance Hub - Mandiant Cyber Risk Management Services
User Policies and Education	Publish easy to understand AI security and privacy policies and education for users.	- Google Cloud Acceptable Use Policy - Google Gen AI Prohibited Use Policy - Google API Services User Data Policy	- Publish application-specific policies - In-application disclosures and education - Responsible Gen AI Toolkit
Internal Policies and Education	Publish comprehensive AI security and privacy policies and education for your employees.	- Google internal training - Google AI Principles - Secure AI Framework	- AI governance framework - Internal AI security policies - AI Literacy Hub - AI courses and tools

Product Governance

Control definition

This control validates that all AI models and products meet the established security and privacy requirements.

What we do

Google's internal product governance process is multilayered, involving reviews by security, privacy, and legal teams throughout the product lifecycle. This ensures that all products, including AI services, meet Google's high standards for security and privacy before they are launched. For more information, refer to Google's Safety Center.

Google's product governance is evidenced by our strong compliance posture.

- Google Cloud Platform, Google Workspace, and Gemini (app) are certified compliant with ISO/IEC 42001, an internationally recognized standard for the responsible development and use of AI systems.
- Google Cloud AI has been assessed against the NIST AI Risk Management Framework, which provides guidance for developing, using, and evaluating AI systems.
- Google has announced a commitment to the EU AI Act Code of Practice.
- Gemini for Google Cloud has been certified against a number of global and regional security standards.
- Google Cloud as a whole supports a wide range of security standards and frameworks.

What you should consider

If you are developing your own AI models and products, you should start by understanding and documenting the AI compliance requirements for the markets, jurisdictions, and industry segments in which you are operating, and the internal standards and policies that you need to follow. Ensuring your products meet your requirements is a customer responsibility, but Google Cloud provides a number of tools that can help.

Data residency: The majority of Google Cloud services (including Vertex AI, Cloud Storage, and BigQuery) support data residency. This can be further reinforced with an organization policy for resource location, which mandates data residency controls for all resources within a specific organization, folder, or project.

Data Boundary (via Assured Workloads): Assured Workloads packages a number of different Google Cloud controls to help enable compliant deployments across different jurisdictions. It combines data residency (enforced via organization policies), key management, compliance monitoring and, where relevant, in-region Assured Support, providing control packages for many different jurisdictions and use cases.

Vertex AI Model Registry: The Model Registry is the central tool for enforcing product governance for AI models. This enables customers to establish a process where all models are registered and must pass a series of checks before they can be approved for deployment. The registry can be used to track the model's version, its associated training data (via metadata), evaluation metrics, and its approval status.

Binary Authorization: The model-approval process in the Model Registry can be integrated into an automated continuous integration and deployment (CI/CD) pipeline. As part of this pipeline, Binary Authorization can be used to cryptographically attest that all containerized artifacts (including AI models and AI application code) have passed all required governance checks (e.g., vulnerability scans, fairness evaluations, security reviews) before being deployed into production.

Risk Governance

Control definition

This control involves inventorying, measuring, and monitoring residual risk to AI in an organization.

What we do

Google continuously monitors its own infrastructure for risks and provides high-level security assessments and compliance reports to customers through resources like the Google Cloud Trust Center.

What you should consider

Similar to other assurance and governance controls, implementing risk governance is a customer responsibility.

Google has published Guidance for Boards of Directors on How to Address AI Risk to help customers establish an AI risk governance framework.

Google also provides tools and resources to support customer risk assessments.

SAIF Risk Self-Assessment: Google provides a public, questionnaire-based SAIF Risk Self-Assessment tool. You can use this tool to evaluate your AI security posture against the SAIF framework. Based on the responses, the tool generates a tailored report that highlights specific risks (e.g., Data Poisoning, Prompt

Injection) and suggests relevant mitigating controls. The report can be paired with this paper to help identify specific implementations of the SAIF controls within Google Cloud.

Risk Management in Security Command Center: Security Command Center provides an overall risk dashboard, helping you to move beyond a simple list of vulnerabilities to a risk-based view of security. It allows governance teams to measure and monitor your security posture over time and focus remediation efforts on the issues that matter most.

Compliance resource center: The compliance resource center is a centralized hub that provides comprehensive documentation to help customers validate Google Cloud's security and compliance controls. It offers direct access to industry-leading certifications, third-party audit reports, and mappings to various global and industry-specific standards, enabling you to manage your compliance needs and download necessary reports

Cyber Insurance Hub: Google's Cyber Insurance Hub includes a security diagnostic tool that scans your workloads on Google Cloud and provides proactive security recommendations to minimize misconfigurations, drive down risk, and boost security readiness. Cyber Insurance Hub generates a report that helps you understand your security-risk posture on an ongoing basis. You can choose to share your report with cyber insurance providers to determine eligibility for cyber insurance policies designed exclusively for Google Cloud customers.

Mandiant Cyber Risk Management Services: Mandiant offers a number of Cyber Risk Management Services to help customers identify and manage relevant cyber risks, assess existing security programs, mitigate cyber risks related to business transactions, carry out threat modeling, or improve vulnerability management via risk-based strategies.

User Policies and Education

Control definition

This control involves publishing easy-to-understand AI security and privacy policies and education for users. It is implemented by Model Consumers, and the risk mapping will vary.

What we do

Google provides extensive public-facing documentation that can serve as a model and resource to help you develop your own user policies. This includes:

Google Cloud Acceptable Use Policy: This policy defines acceptable use for Google Cloud services in general, including AI services. For example, customers agree not to engage in illegal or fraudulent activities, distribute malware or spam, or to disrupt or impair services.

Generative AI Prohibited Use Policy: This policy clearly outlines activities that are not permitted when interacting with Google's generative AI services (whether or not they are accessed via Google Cloud), such as creating content related to illegal activities, spam, hate speech, or sexually explicit material.

Google API Services User Data Policy: This policy details requirements for applications that use Google APIs, emphasizing principles like data minimization, user consent, and transparency.

✓ What you should consider

Publish application-specific policies: The primary responsibility for this control lies with the customer (the Model Consumer). You should develop and prominently display your own Acceptable Use Policies and Privacy Policies for your AI applications. These policies should be written in clear, nontechnical language and should inform users about:

- The presence and purpose of AI in the application.
- The types of data being collected and how that data will be used.
- The rules of engagement and prohibited uses.
- The limitations of AI and potential for inaccurate or unexpected outputs.

Google has provided additional guidance via this blog post.

In-application disclosures and education: Beyond a formal policy document, effective user education involves in-application notices and tips. For example, an AI chatbot could include a disclaimer stating, "I am an AI assistant and may make mistakes. Please verify important information." Disclosures foster transparency, manage user expectations, and can help prevent misuse of the AI application. They may be a legal requirement in some jurisdictions.

Responsible Generative AI Toolkit: Google's Responsible Generative AI Toolkit includes guidelines on AI policy setting and provides templates for user-transparency artifacts, such as data cards and model cards.

Internal Policies and Education

Control definition

This control involves publishing comprehensive AI security and privacy policies, as well as education for employees.

What we do

Google internal training: Google has a robust internal security and privacy training program for all employees. This includes mandatory training during onboarding, regular security-awareness updates, and specialized technical training for engineers on topics like secure coding, threat modeling, and privacy-preserving design. Google also implements AI literacy training for all employees and provides a wide range of internal AI training resources.

Google AI Principles: Google's published AI Principles guide the development and deployment of our AI systems and provide the foundation for our AI frameworks and policies, including the Secure AI Framework discussed in this paper.

✓ What you should consider

AI governance framework: You should establish a formal internal AI governance framework. This framework should define roles and responsibilities, such as an AI steering committee for strategic oversight, an AI security team for technical controls, and data stewards for data quality and access. Google has published Guidance for Boards of Directors on How to Address AI Risk and Gen AI Governance: 10 tips to level up your AI program, both of which address this topic in more detail.

Internal AI security policies: Based on the governance framework, you should create and disseminate detailed internal policies for the secure development and deployment of AI. These policies should cover the entire lifecycle, including requirements for data handling, model training environments, vulnerability management, and incident response. Google's Secure AI Framework can serve as an excellent foundation for these internal policies.

Employee training and awareness: You should implement a training program to educate developers, data scientists, and other relevant employees about your internal AI security policies and the specific risks associated with AI. This should include AI literacy training for all users and more advanced training (such as secure coding practices for AI, data privacy principles, and how to use your approved AI tools and platforms). Google publishes substantial AI training resources, accessible from the following sites:

- **AI Literacy Hub**
- **AI Courses and Tools**

08 Conclusion

Securing the AI frontier with SAIF and Google Cloud

The Secure AI Framework provides a critical, lifecycle-aware roadmap for navigating the complex security landscape of artificial intelligence. Google Cloud offers a comprehensive and integrated suite of AI and security services that is deeply aligned with SAIF controls. This powerful alignment is no accident. SAIF codifies the battle-tested, global-scale security principles that Google developed internally to secure our own AI operations.

The alignment starts with Google Cloud's **secure-by-default infrastructure,** establishing foundational trust in the physical and network layers. Building upon this, the **Vertex AI** platform delivers a hardened, managed environment that enforces controls across the entire ML lifecycle – from managing sensitive data and model training to deploying and continuously monitoring AI models and agents. This core platform is augmented by a rich ecosystem of security services. **Dataplex** ensures strict data governance, **Security Command Center** provides unified risk visibility, and **Model Armor** offers crucial runtime protection. This comprehensive, defense-in-depth approach allows enterprises to transition from simply building AI to building **secure, trustworthy AI.**

The path to responsible innovation

Securing enterprise AI is a continuous effort, but it is one that directly enables competitive advantage. By operationalizing the structured guidance of the Secure AI Framework using the resilient capabilities of Google Cloud, you gain more than just protection. You gain the confidence and trust required to innovate faster. This robust security foundation ensures that the transformative potential of artificial intelligence is unlocked and deployed not just powerfully, but also responsibly, serving as a critical differentiator in the future of AI adoption.

Acknowledgements

Thank you to John Stone and Anton Chuvakin for contributing to this paper.