



Secure AI innovation

Securing the
agentic enterprise
on Google Cloud.





Executive summary

Cyber attackers are no longer operating at human speed—the era of machine speed has well and truly arrived.

Sophisticated criminal groups are deploying fully agentic, multi-modal attacks, collapsing the exploit window before manual security can respond. Against this backdrop, it's vital that enterprises protect their full AI stack—including infrastructure, shadow AI, and unmapped supply chains.

As attackers accelerate their autonomous attacks from the outside, enterprises are racing to deploy their own agentic systems to execute complex tasks, trigger APIs, and interact directly with sensitive data. The issue? Innovation is outpacing security governance, creating a significant gap.

But here's the thing. AI security governance isn't an obstacle for enterprises to overcome—it's actually the catalyst for better, faster innovation. Organizations that establish clear AI governance move faster, secure executive buy-in, and successfully protect their IP. They're 2x more likely to adopt agentic AI, and 2x more confident in their ability to protect AI systems.¹

This guide will show you how to build a resilient, multi-agent enterprise by embedding identity boundaries, runtime guardrails, and autonomous remediation directly into your core infrastructure—allowing you to scale AI safely and effectively.

-7 days

is now the estimated mean time to exploit vulnerabilities—exploits are happening before patches are even released²

60+

React2Shell exploitation campaigns were identified by Wiz within a single week of a public proof of concept³

\$4.99M

is the average global cost of a data breach⁴

“

To transition successfully from experimental AI to enterprise-wide operations, security can no longer be an afterthought. True protection requires a holistic approach that seamlessly connects identity boundaries, runtime guardrails, and continuous threat posture monitoring.”



Dr. Grace Trinidad

Research Director,
AI Security & Trust, IDC

¹ Cloud Security Alliance (CSA), [The State of AI Security & Governance](#), 2025

² Mandiant and Google Cloud Security, [M-Trends 2026: Executive Edition](#), 2026

³ Wiz Research, [Wiz Cloud Threat Retrospective](#), 2025–2026

⁴ IBM Security and Ponemon Institute, [Cost of a Data Breach Report](#), 2026

Table of contents



Securing the baseline	4
Governing the AI supply chain	9
Enforcing the guardrails	13
Fighting AI threats with AI defense	17
Achieving digital sovereignty	22
Bridging the gap between AI velocity and governance	26



Securing the baseline





You can't protect a system if you don't know exactly who—or what—is operating inside it.

Static service accounts and broad access keys are fundamentally unable to secure autonomous agentic workloads. To entrust agents to query databases and execute APIs, they must possess unique, verifiable identities.

The exposure risk

82x

more non-human workload identities exist in modern cloud environments than human identities⁵

42%

of active AI agents currently possess highly sensitive, over-permissive actions⁵

57%

of organizations have already deployed self-hosted AI agent technologies⁶

Closing this exposure gap requires a dynamic, identity-first framework built specifically for autonomous intelligence. Across Google Cloud, AI agents operate under unique, verifiable digital IDs that exist for the duration of their task, ensuring every action is strictly scoped, authenticated, and audited.

This identity-first approach is built on three core controls:

1 Agent Identity

A dedicated digital identity created specifically for AI, rather than reusing generic service accounts.

2 Least-privilege controls

Restricts agent capabilities using Privileged Access Manager (PAM) for just-in-time access.

3 Principal access boundary (PAB)

Acts like a digital guardrail—setting strict limits on how far an agent can travel so it never touches systems outside its scope.

⁵ Abhishek A. Hemrajani, Google Cloud Next '26, [What's next in IAM: Security, governance, and runtime defense for AI agents](#), 2026

⁶ Wiz, [State of AI in the Cloud](#), 2026



“

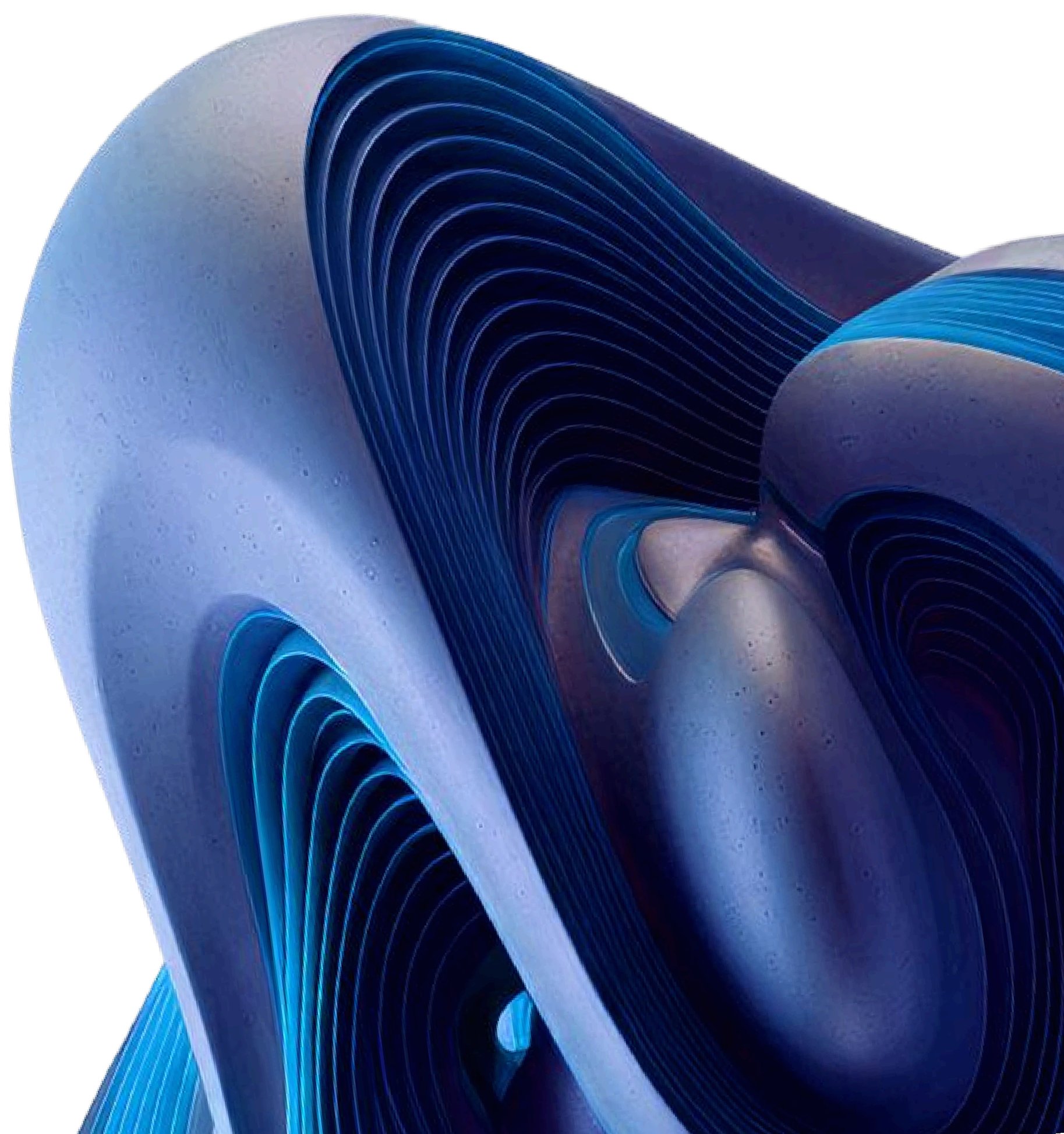
We built Agent Identity as a unique first-class citizen inside Google Cloud. It is cryptographic, strongly attested, and lifecycle tied.”



Ravi Shah

Product Management Lead for Identity,
Google Cloud

Google Cloud Next '26, [Agent Identity as the backbone for secure AI innovation](#)



Supporting AI with a secure-by-default foundation

While identity secures the agent, overall security is only as strong as its underlying platform. Google Cloud's Enterprise Foundations Blueprint provides the baseline configurations required to support enterprise workloads at scale.

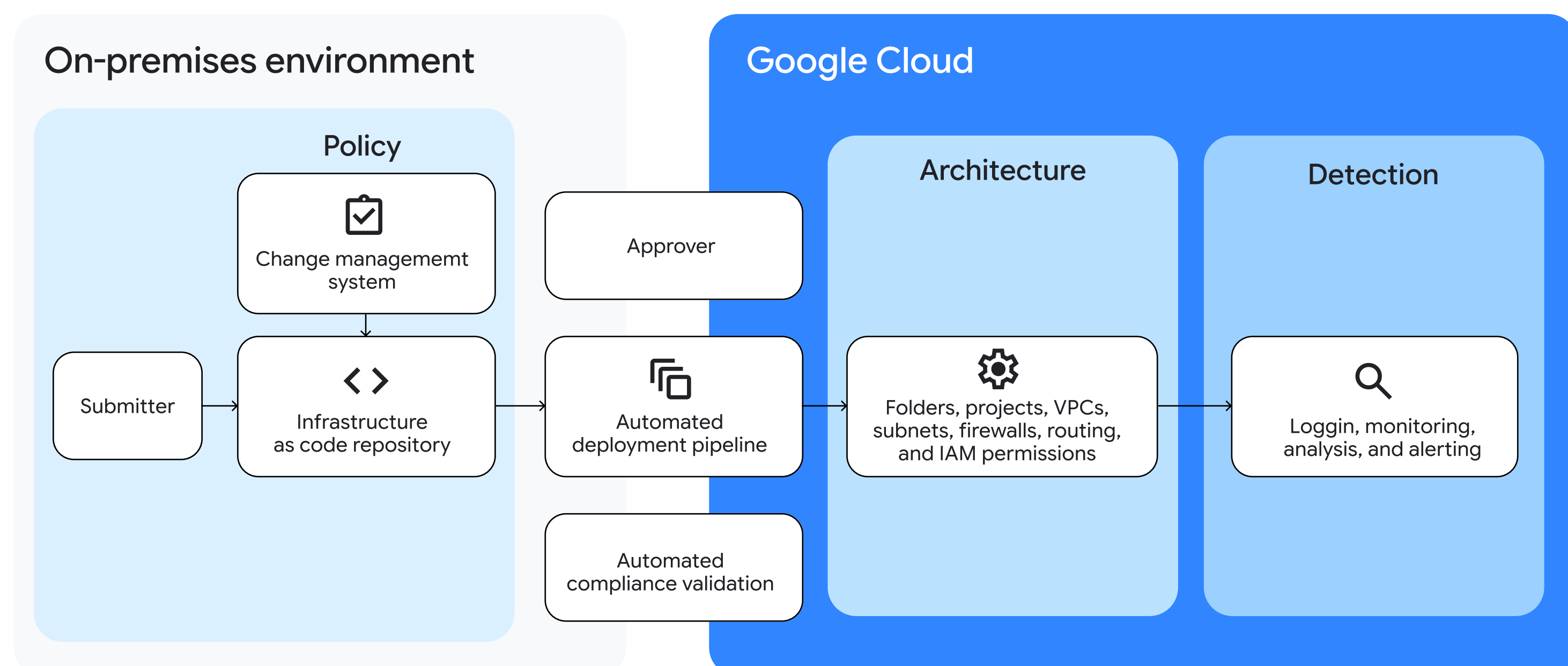
Core infrastructure controls

GitOps and declarative pipelines

Deployments run through automated Cloud Build pipelines that check code quality before launch—creating a tamper-proof digital paper trail (SLSA Level 3 provenance) verified by Binary Authorization.

Private topologies

We group AI systems into secure, centrally managed areas (Shared VPCs). VPC Service Controls (VPC-SC) then act like a digital vault, creating a protective perimeter to keep sensitive data inside.





Security in action

Eliminating static keys in multi-agent workflows

When a customer submits a warranty claim, the customer-facing agent must securely communicate with backend data agents to verify the claim. By assigning unique agent identities to authenticate this agent-to-agent (A2A) communication, organizations can permanently eliminate static, shared keys from their database environments.



How many of your agents actually have a badge to get into your database? If agents are going to do work for you autonomously, they need to have a unique identity.”

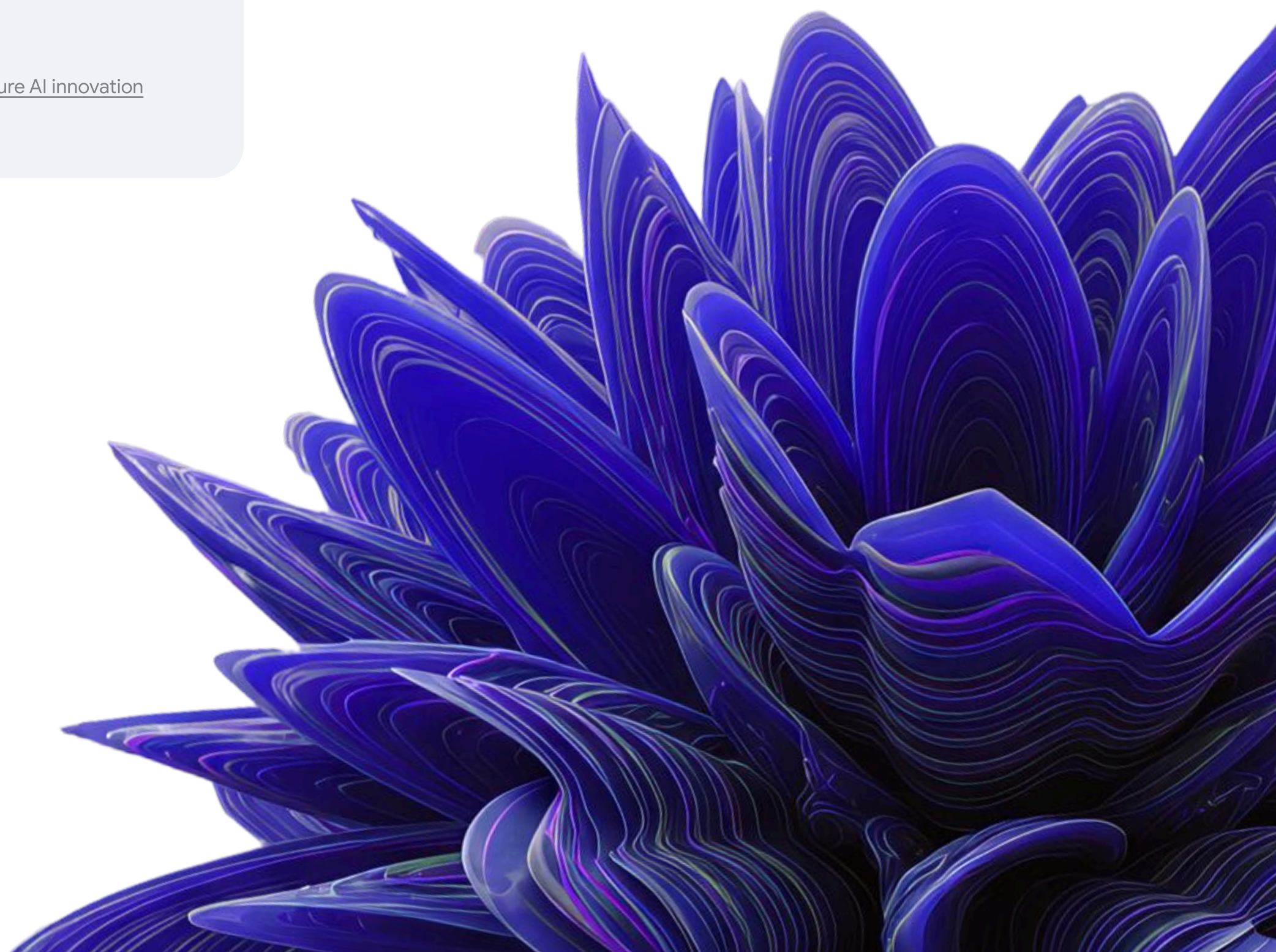
**Ravi Shah**

PM Lead for Identity, Google Cloud

Google Cloud Next '26, [Agent Identity as the backbone for secure AI innovation](#)

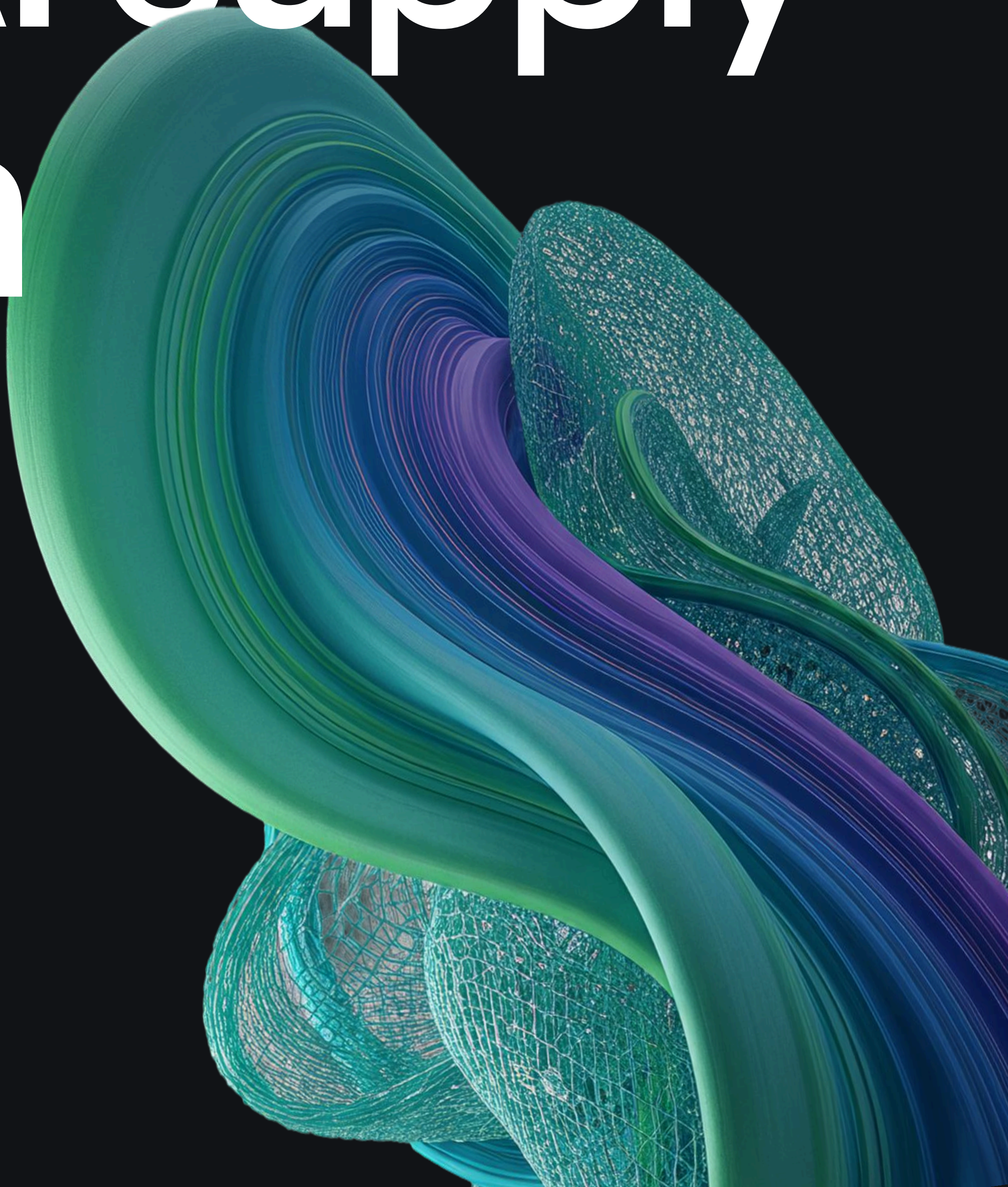
Key takeaway

You can't govern what you can't identify—establishing a unique, verifiable identity for every agent is the first step towards secure AI.





Governing the AI supply chain





Even with a solid platform baseline, enterprises can still face issues with shadow AI.

Up against strict deadlines, developers often bypass IT to plug in untested third-party agent skills, introducing potentially significant vulnerabilities and unmanaged API endpoints directly into production.

The supply chain threat

Over 15%

of agent skills in public marketplaces are found to be malicious⁷

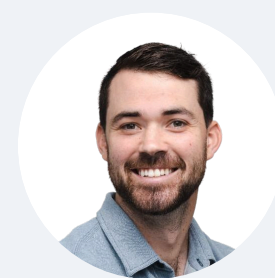
68%

of organizations are running self-hosted models bundled deep inside third-party software—often without realizing it⁸

To regain control over this sprawling, unmanaged supply chain, organizations need a centralized hub. Gemini acts as a control plane, becoming the primary secure execution host for business and security agents alike.

“

From a supply chain perspective, we're seeing the same mistakes as in open source... You must restrict ingestion, verify skills, scan manifests, and pin dependencies.”



Knox Anderson

Lead for Securing AI, Google Cloud

⁷ Knox Anderson, Google Cloud Next '26, [Secure AI agents from development to runtime](#), 2026

⁸ Wiz, [State of AI in the Cloud](#), 2026



Gemini key security and governance capabilities

Agent Registry

Centralized catalog for all running agents, external APIs, and secure tool connections.

Agent Development Kit (ADK)

Enforces strict user boundaries to prevent data leaks, ensuring one user's private session never bleeds into another's.

Google Cloud MCP Server

Standardizes how agents connect to data sources, eliminating risky, patchwork integrations.

Certificate manager

A single pane of glass for all agent-related security certificates.

Gemini Threat Detection

Uses continuous telemetry to instantly detect anomalous agent activity, operational misuse, or active exploits at machine speed.

Wiz AI-BOM integration

Wiz's dynamic AI-Bill of Materials (AI-BOM) provides continuous, automated discovery of every AI component across your cloud environment.

Agent Gateway

Acts as the "air traffic control" for all agentic interactions, using Identity-Aware Proxy (IAP) to verify digital IDs and enforce strict access boundaries before routing connections.

Identity auth manager

Securely handles complex user logins (OAuth) behind the scenes, ensuring the AI agent never actually sees or holds the sensitive passwords or access tokens.

Agent security dashboard

Delivers real-time visibility into your AI ecosystem, mapping misconfigurations, toxic permission combinations, and overall security posture in a single view.



Best Buy's BYOC isolation and human-in-the-loop architecture

While Gemini acts as the control plane, high-risk agents often require isolated execution environments. Best Buy isolates agents (like a Logistics Liaison) on Cloud Run using Bring Your Own Container (BYOC). High-risk actions are physically paused, requiring a human-in-the-loop (HITL) approval step in the ADK before APIs can execute.



Our core security considerations were a centralized registry, restricted tooling (so a support agent can't access payment databases), and strictly separating deploy-time permissions from runtime permissions.”



Kishor Patil

Senior Engineering Manager, Best Buy

Google Cloud Next '26, [Secure AI agents from development to runtime](#)

💡 Key takeaway

Safe agent deployment requires strict boundaries—maintaining human-in-the-loop approval ensures high-risk actions remain under human control.





Enforcing the guardrails

An abstract graphic in the bottom right corner consisting of several overlapping, wavy, ribbed shapes in shades of orange and yellow, resembling a stylized flame or a modern architectural element.



As multi-agent systems interact in complex, unmonitored chains, centralized auditability and operational control become more challenging.

While 52% of organizations cite sensitive data exposure as their top concern, only 21% actively prioritize model-level risks.⁹ For enterprises to secure their data, they first need to secure the model interactions.

Google Cloud enforces two distinct runtime policy layers to control both where an agent can go and what it intends to do.

01 Controlling access (IAM boundaries)

Acts as a strict guard at the Agent Gateway—only allowing agents into systems they are explicitly approved for.

02 Controlling intent (semantic boundaries)

Smart filters dynamically analyze the meaning of a user's prompt in real-time to ensure the agent isn't being tricked into doing something harmful.

⁹ Cloud Security Alliance (CSA), *The State of AI Security & Governance*, 2025



Inline prompt sanitization and data validation

Acting as a protective shield for your AI models, Model Armor provides real-time, inline protection to block prompt injections, prevent jailbreaks, and redact sensitive data without adding latency.

Consider an autonomous warranty claim system. When a user submits a prompt, Model Armor scrubs it before it reaches the orchestrator. To stop attackers from sneaking malicious code into database commands (indirect SQL injections), agents check user inputs against strict rules—for example, verifying a serial number contains only numbers and letters before letting it query BigQuery.



Agents are the new insider threats; they don't need malicious intent to cause harm."



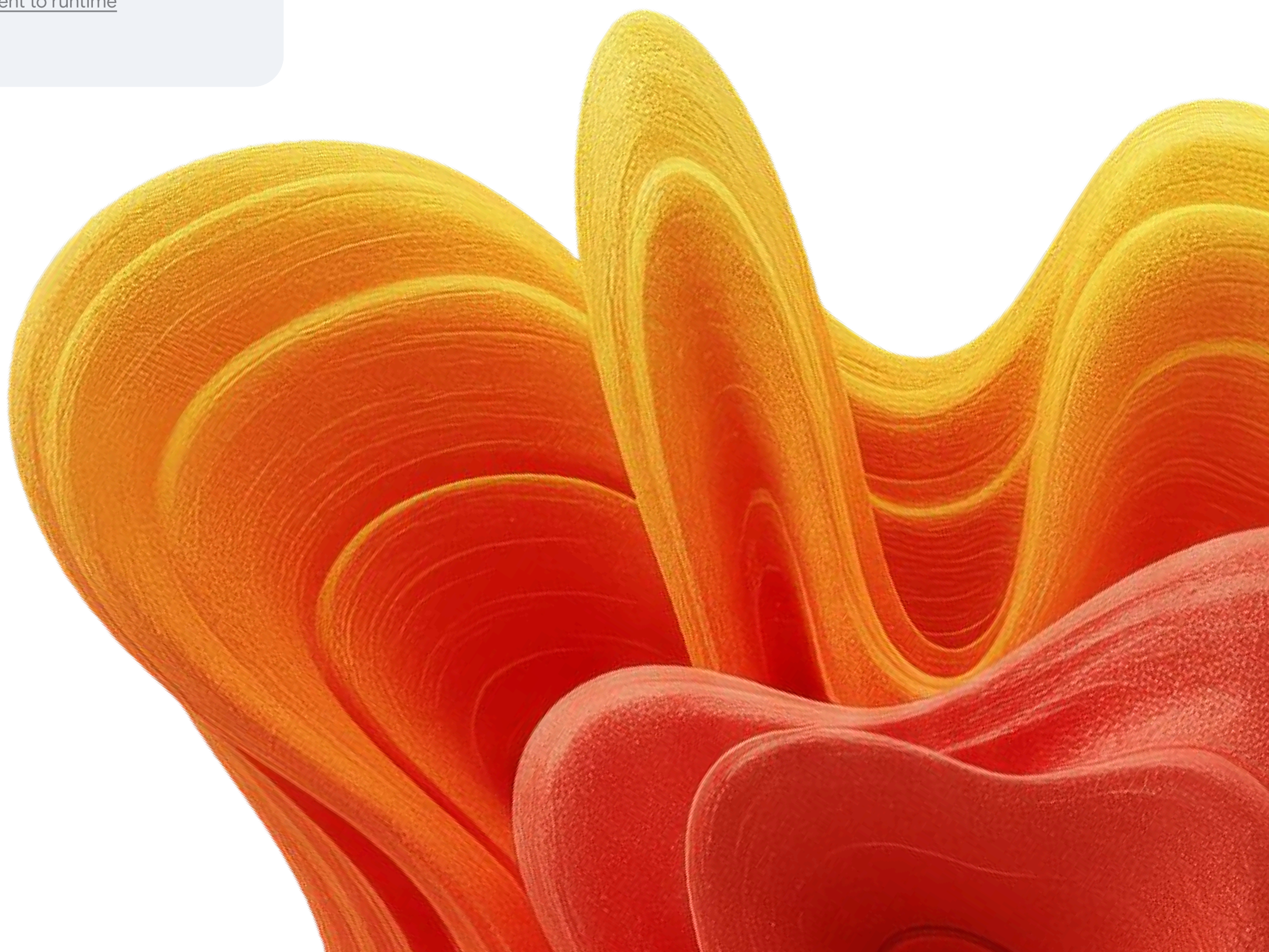
Abhishek A. Hemrajani

Senior Director, Product Management,
Google Cloud Security

Google Cloud Next '26, [Secure AI agents from development to runtime](#)

💡 Key takeaway

Traditional access controls are no longer enough—to protect your sensitive data, you need to secure the interaction using real-time content sanitization.



“

Agent Gateway is the control plane for all ingress and egress... Model Armor sits in front of any AI interaction to screen for prompt injections, sensitive data leaks, and hate speech.”

**Knox Anderson**

Lead for Securing AI, Google Cloud

Google Cloud Next '26, [What's next in IAM: Security, governance, and runtime defense for AI agents](#)



Fighting AI threats with AI defense



With cyber attacks taking place at machine speed, vulnerabilities are often exploited before a software patch is even published.

Traditional siloed security teams—such as application security, DevOps, and cloud security—lead to high operational friction and severe alert fatigue, stalling AI adoption.

To achieve true machine-speed defense, enterprises must move from manual triage to autonomous remediation. By pairing Wiz’s vulnerability findings—enriched with cloud context to pinpoint high-impact risks—with CodeMender’s autonomous patching, organizations create an immediate find-and-fix loop. Delivered natively through Wiz Green Agent, these capabilities leverage Gemini guardrails to autonomously generate tested, production-ready code fixes at scale.

“

At Wiz, our goal is to help customers secure every layer of AI applications... Wiz breaks these silos by mapping vulnerabilities, misconfigurations, and active exploits onto the unified Wiz Security Graph.”



Jiong Liu

VP of Product Marketing, Wiz

Google Cloud Next '26, [Beyond the buzz: How AI is actually transforming modern security](#)

Unified posture visibility

Mapping vulnerabilities across your entire AI stack

As the driving force behind Google Cloud’s multi-cloud defense, Wiz breaks down security silos by mapping vulnerabilities, misconfigurations, and active exploits across your entire AI stack onto a single, cohesive Wiz Security Graph.

For GCP workloads

Google Security Command Center delivers built-in, preemptive protection tailored for GCP environments.

For complex multi-cloud

Wiz AI-APP is the highly recommended standard (spanning Gemini, AWS Bedrock, and unmanaged open-source models).

Active exploit verification

Traditional scanners can overwhelm teams with false positives. To ensure enterprises prioritize real security tasks, CodeMender proves whether a vulnerability is genuinely exploitable before flagging it to engineers.

Once an exploit is verified, remediation timelines must accelerate from weeks to minutes. CodeMender acts as an autonomous defense agent to generate tested, ready-to-merge patches directly inside the developer’s workspace.

How CodeMender and Wiz work together

Context-enriched findings

CodeMender ingests raw findings and exposure context directly from the Wiz Security Graph to target the highest-risk attack paths.

External exposure testing

Wiz’s Red Agent identifies and tests exploitable application endpoints from the outside, feeding external exposure context directly into CodeMender’s automated remediation pipeline.

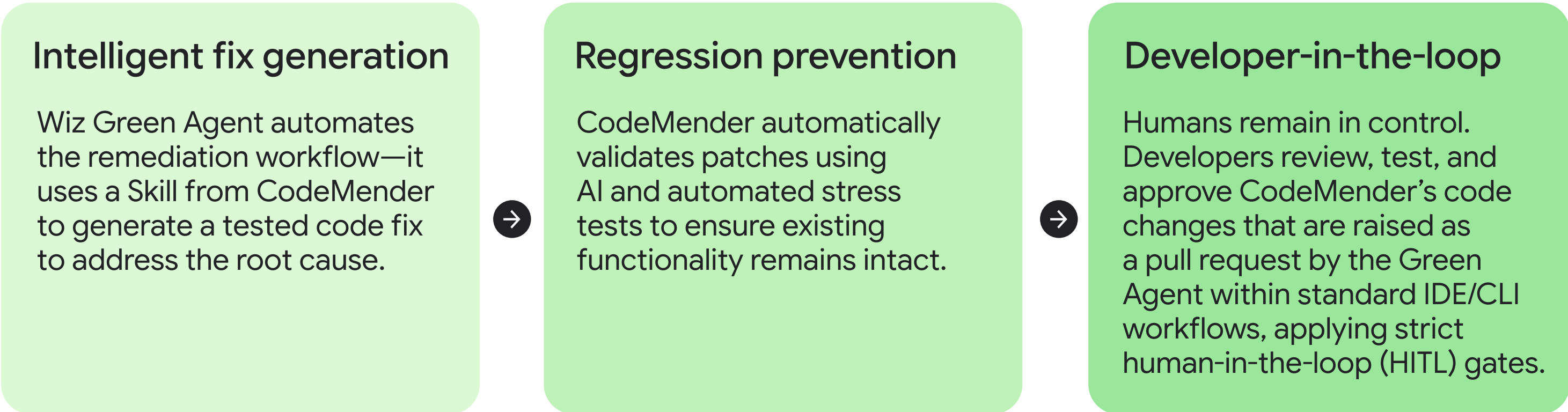
DeepMind-engineered reasoning

CodeMender safely builds the application in an isolated sandbox and attempts to execute the attack. By proving the vulnerability is real before alerting engineers, it eliminates false alarms.

Privacy-first execution

CodeMender’s scanning engine works through a local CLI, processing and verifying code in a secured sandbox so your private code is protected and never used to train our models.

The remediation workflow



Continuous threat protection at machine speed

Remediation is an ongoing process. As AI agents operate at massive scale, continuous monitoring is required to detect anomalies and trigger the find-and-fix loop.

Wiz Blue Agent

Correlates real-time telemetry (invocation logs, DB queries) during live threat incidents to provide immediate context.

Agent anomaly detection

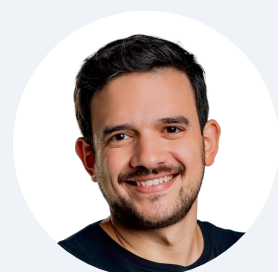
Agents are the new insider threats. Google Cloud uses behavioral analytics (UEBA) to monitor agent activity against normal baselines and detect misuse.

The scale of the threat

Frontier AI models are already capable of completing ~50% of real-world attack challenges.¹⁰

“

The attack surface is expanding due to ‘vibe coding’... agents operate at insane scale. We saw one agent make 42,000 cloud API calls and 300 sensitive IAM modifications in an hour.”



Alon Schindel

VP of AI & Threat Research, Wiz/Google

Google Cloud Next '26, [The 2026 AI threat landscape](#)

💡 Key takeaway

Machine-speed attacks demand machine-speed defenses—use AI to actively verify exploit paths and generate secure, human-approved patches that won't break your infrastructure.

“

Using Application Design Center (ADC), you can run SCC detections on app templates before deployment... For AI agents, Agent Security Posture natively provides discovery and toxic combination mapping.”



Abhishek A. Hemrajani

Senior Director, Product Management,
Google Cloud Security

Google Cloud Next '26, [Built-in defense: The next evolution of Security Command Center for AI-era](#)





Achieving digital sovereignty

An abstract digital art composition featuring a dense, textured mass of small, colorful spheres (purple, pink, blue, and red) that resemble a digital landscape or a cluster of data points. Several larger, smooth, translucent spheres in shades of purple and pink are scattered around the main mass, some appearing to float or roll. The background is a solid dark color, making the vibrant colors of the spheres stand out.



Securing data against active threats is only half the battle—organizations must also secure it against compliance and residency risks.

Highly regulated industries face increasingly rigorous data residency and operational autonomy requirements. 50% of organizations identify regulatory compliance as one of the biggest hurdles to securing AI.¹¹

Google Cloud delivers sovereignty as a standard platform feature, ensuring the seamless transition of AI workloads using centralized trust controls.

Cloud KMS and KAJ

Customers maintain absolute cryptographic key ownership through Cloud KMS. Under Key Access Justifications (KAJ), Google cannot decrypt any workload without explicit, auditable approval.

Secret Manager

Safely stores sensitive application secrets, preventing the risk of unmanaged credentials slipping into autonomous agent code.

The Google Cloud sovereign portfolio

01 Google Cloud data boundary

Delivered with Assured Workloads, it provides a sovereign data and access boundary in the public cloud, keeping encryption keys and personnel controls strictly regional.

02 Google Cloud dedicated

Complete regional data and operational sovereignty delivered by local independent operators (survivable without Google).

03 Google Distributed Cloud (GDC)

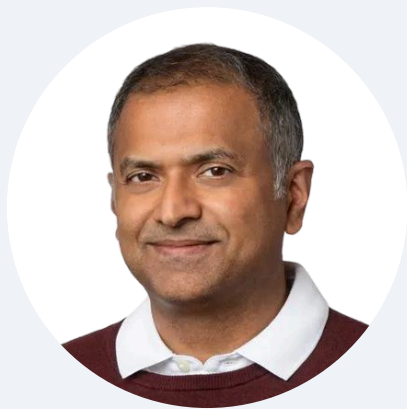
Fully air-gapped, on-premises environments that operate without any external connection to the public internet—designed for defense, intelligence, and highly regulated sectors.

¹¹ Cloud Security Alliance (CSA), *The State of AI Security & Governance*, 2025



“

Our strategy is built on providing a consistent experience, including AI solutions, across three distinct sovereign cloud platforms, so that enterprise and government organizations can innovate and meet their compliance obligations.”

**Jai Haridas**

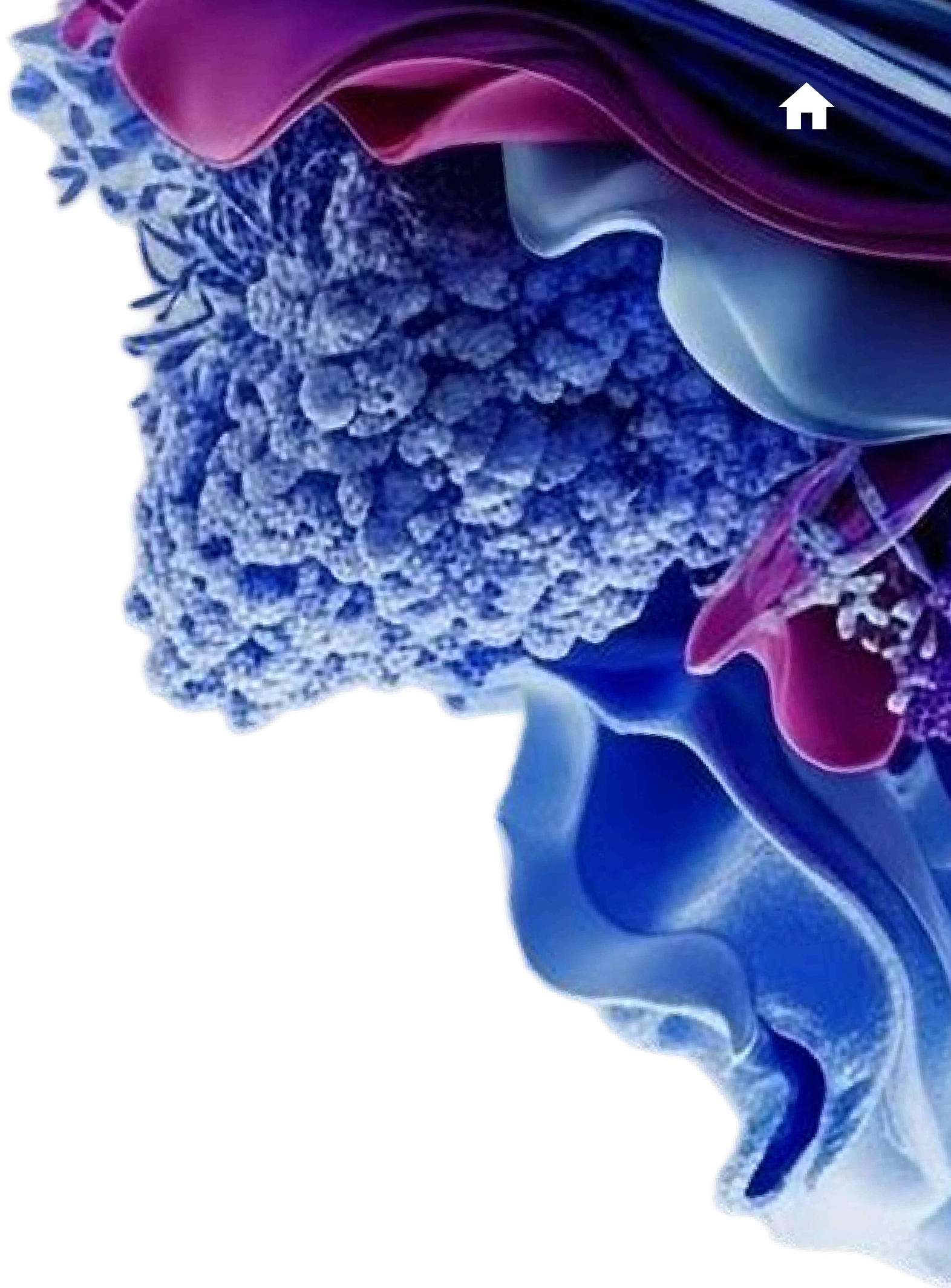
VP/GM, Regulated and Sovereign Cloud,
Google Cloud

Google Cloud Next '26, [Geopolitical resilience: Redefining sovereignty for modern enterprises](#)



Bridging the gap between AI velocity and governance





As developers rapidly adopt AI to create new innovations at unprecedented speed, the challenge for any enterprise is this: How do you secure AI without slowing down your team's momentum?

Shadow AI and vibe coding create new risks that bypass legacy security controls—and static compliance checklists simply aren't made for the new reality of machine-speed, multi-agent environments.

Closing this exposure gap requires shifting to a 4-stage agentic security lifecycle: Architect, inventory, intercept, and defend. This framework allows developers to rapidly innovate while maintaining total enterprise governance.

Stage 1

Architect: Create the foundation and identity baseline

Zero Trust for AI

Replace static service accounts with lifecycle-bound Agent Identities and control agent capabilities using Privileged Access Manager (PAM) for just-in-time access.

Perimeter strengthening

Enforce Principal Access Boundaries (PAB) and VPC Service Controls (VPC-SC) baseline perimeters.

Stage 2

Inventory: Govern the AI supply chain

Continuous discovery

Uncover shadow AI and third-party models across multi-cloud using Wiz AI-BOM.

Secure packaging and provenance

Validate build-time dependencies (SLSA Level 3 via Cloud Build) and isolate high-risk execution with BYOC on Cloud Run.

Centralized registry

Catalog vetted agents, MCP tools, and external API connectors via Gemini.

Human-in-the-Loop (HITL)

Gate sensitive downstream actions using ADK approval steps.



Stage 3

Intercept: Enforce intent-aware runtime protection**Gateway traffic control**

Authenticate and audit agent-to-agent (A2A) flows through the Agent Gateway (IAP).

Dual-layer payload defense

Implement inline prompt sanitization and PII redaction with Model Armor, combined with dynamic semantic governance.

Stage 4

Defend: Block threats at machine speed**Agent Threat Detection**

Detect anomalous agent activity, operational misuse, or active exploits using built-in threat detectors in Gemini.

Continuous behavioral telemetry

Detect anomalous API spikes and unauthorized role modifications using Agent UEBA and Wiz Blue Agent.

Autonomous find-and-fix

Validate vulnerability exploitability in isolated sandboxes and generate tested code patches with CodeMender.

Continuous sovereignty and cryptographic assurance

In tandem with the 4-stage agentic security lifecycle, it's critical to enforce verifiable customer data ownership via Cloud KMS and Key Access Justifications (KAJ). Enterprises must also maintain consistent security governance across public cloud, dedicated regional clouds, and air-gapped Google Distributed Cloud (GDC).

💡 Key takeaway

The gap between AI enthusiasm and execution is now the biggest barrier to scaling successfully. Organizations that shift security directly into their platform fabric move faster and reduce risk. Security becomes the ultimate enabler, allowing you to turn your AI ambitions into sustainable, secure business value.

Ready to secure your next AI innovation?

Contact us

