

Agentic AI Security: The Enterprise Playbook

Governing AI Systems Across Applications, Agents, and Employees

Table of contents

Executive Summary

PART I: THE PROBLEM

1. AI Has Crossed the Line from Use to Action
 2. Where Sensitive Data Is Actually Going
 3. Why Traditional Security Controls Fall Short
-

PART II: THE SOLUTION

4. The Three New AI Exposure Surfaces
 5. MCPs and the New Execution Surface
 6. The AI Defense Plane
 7. One Platform, Three Layers of Protection
-

PART III: IN PRACTICE

8. Proof, Production, and Scale
-

Conclusion

+67%

SURGE IN INCIDENTS

Reported AI security incidents jumped 67% year-over-year (from 9% to 15%) as organizations moved from pilots to production.

Source: Lakera 2025 GenAI Security Readiness Report

“Your data isn’t leaking. Your employees—and agents—are handing it over.”

AI Security Is No Longer a Feature. It’s a Category.

Artificial intelligence has crossed a critical threshold.

What began as isolated experimentation quickly evolved into AI assistants embedded in daily workflows. Those assistants became AI-powered applications that interact with customers and employees. Today, organizations are deploying autonomous AI agents that retrieve data, invoke tools, and take action across enterprise systems.

Each step increases business impact—and expands risk.

AI systems no longer simply process requests. They interpret language, infer intent from context, and act on behalf of users. In doing so, they have turned **language itself into an execution layer**.

Security failures now emerge not only from misconfiguration or software vulnerabilities, but from interaction: what users say, what systems retrieve, how models reason, and what agents do next.

Traditional security tools were not designed for this shift. Most cannot see prompts, dynamically assembled context, reasoning chains, or tool decisions. They cannot reliably infer intent. And they cannot control behavior as it unfolds in real time.

As a result, enterprises are scaling AI adoption faster than their ability to govern it.

A New Requirement: A Unified AI Defense Plane

To secure AI at enterprise scale, organizations need a new approach. AI security cannot be bolted onto existing tools or handled as isolated point problems. It must operate across how AI is accessed, how it is embedded, and how it executes actions.

DISCOVER

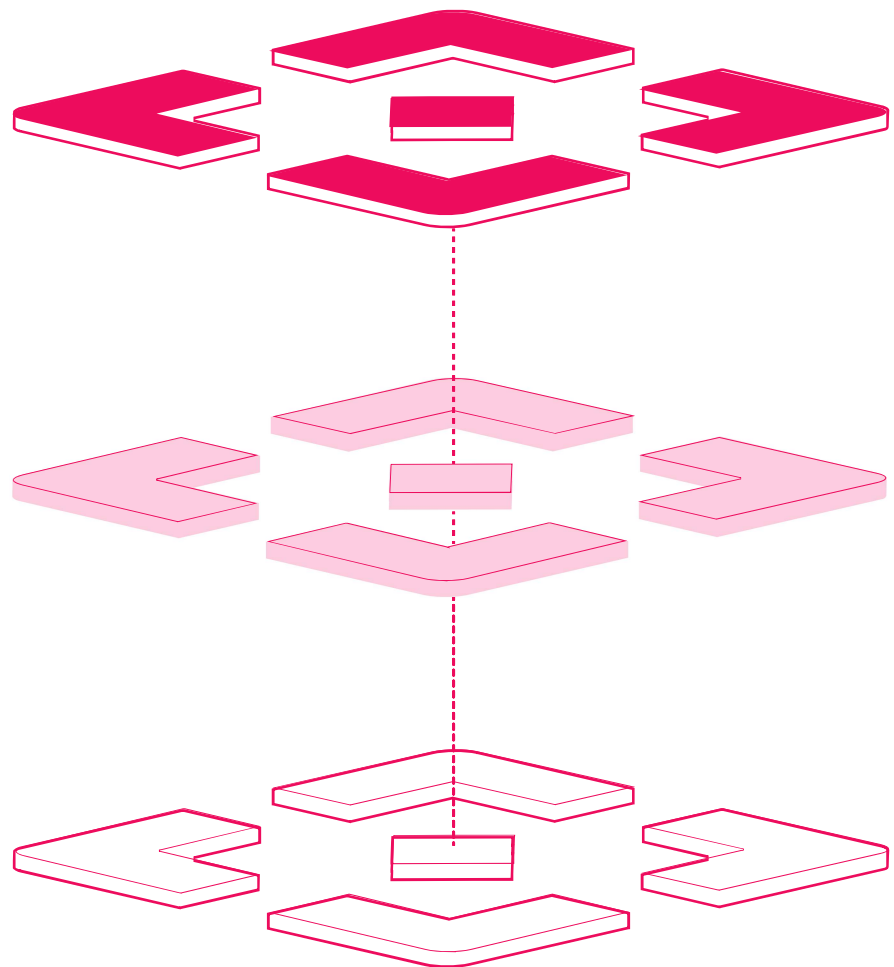
Where AI is used and how it behaves across the organization.

PROTECT

Against unsafe or malicious interactions in real time.

GOVERN

AI usage consistently across users, applications, and agents.



Together, Check Point delivers a unified AI security architecture that spans workforce AI usage, AI applications, and agent runtime autonomy.

PART I

The Problem

Why AI Changes the Security Equation

15%

AUTONOMOUS DECISIONS

By 2028, 15% of day-to-day work decisions will be made autonomously by agentic AI—shifting the burden of trust from humans to machines.

[Source: Gartner](#)

AI Has Crossed the Line from Use to Action

Initially, AI systems were largely observational.

They summarized text, answered questions, and generated recommendations. Even in enterprise tools, their role was advisory. Humans remained the final decision-makers, and traditional security assumptions still held.

That is no longer the case. **Modern AI systems can act.**

They retrieve internal data, modify records, trigger workflows, and invoke APIs. In many organizations, these actions happen automatically or semi-autonomously, based on natural-language instructions rather than explicit code paths. This transition marks a fundamental shift in the security model.

From Assistive AI to Operational AI

The progression has been subtle but rapid. Copilots evolved from suggestion engines into workflow accelerators. Applications moved from static prompts to dynamic context retrieval. Agents gained memory, tool access, and the ability to plan multi-step actions.

What changed was not just capability, but **authority**. AI systems are now entrusted with access to internal knowledge bases, customer data, and operational tooling. They operate across multiple systems, often without a single, clearly defined execution boundary.

Why “Correct Output” Is No Longer the Goal

Traditional AI evaluation focuses on output quality: Is the response accurate? Is it aligned with policy? Does the model refuse unsafe requests? Those questions still matter, but they are no longer sufficient.

When AI systems can take action, the primary risk is not whether an output looks reasonable. The risk is whether the system's behavior remains within intended bounds. A response can be linguistically correct and still unsafe if it exposes sensitive data through indirect reasoning, triggers an unintended tool call, or escalates privileges.

Language as an Execution Layer

Prompts are no longer just inputs. They shape plans, determine tool selection, and influence downstream actions. In this environment, inputs are unstructured, intent is inferred, and execution paths emerge at runtime.

Autonomy Expands the Blast Radius

As autonomy increases, so does impact. Agents that can plan and execute multi-step tasks introduce compounding risk. A single manipulation at the prompt or context level can cascade across tools and systems.

This is why AI security incidents increasingly resemble operational failures rather than simple misuse:

- Data is modified rather than leaked
- Actions are taken rather than suggested
- Systems behave correctly according to logic, but incorrectly according to intent

As autonomy increases, so does impact. Agents that can plan and execute multi-step tasks introduce In these scenarios, accountability becomes blurred. Was the failure caused by the user, the model, the application, or the system design? Without visibility into AI behavior at runtime, organizations cannot answer that question with confidence. risk. A single manipulation at the prompt or context level can cascade across tools and systems.

60%

TARGET SYSTEM DATA

~60% of observed attack traffic attempts to leak system prompts—targeting the application's internal instructions and IP.

Source: Laker Q4 2025 Agent Security Trends

Where Sensitive Data Is Actually Going

Most conversations about AI data risk start with a familiar question: “Is our data being used to train the model?”

That question matters, but it often distracts from the larger, more immediate issue. In enterprise settings, sensitive data rarely leaves because a model provider is training on it. It leaves because AI systems create new paths for legitimate-looking disclosure, often inside workflows that existing controls were never designed to inspect.

Sensitive data does not only leak through files and databases. It moves through interactions.

The New Data Paths in AI Systems

1. PROMPTS & INSTRUCTIONS

Users paste customer records or code into AI tools. Even legitimate intent creates a data transfer mechanism.

2. RETRIEVED CONTEXT

Apps pull internal docs into the context window. Manipulation can expose this retrieved content unexpectedly.

3. OUTPUTS

Summaries and transformations can reconstruct private data even when users do not explicitly ask for it.

4. AGENT ACTIONS

Agents pass private data through tool calls. Data moves across systems without appearing as a file transfer.

These data paths are difficult to manage because they are unstructured, dynamic, and ambiguous. Data appears inside natural language, not fields. Context changes at runtime. Intent must be inferred, not declared.

Most importantly, they blur responsibility. Data can originate with a user, be pulled from internal sources, be transformed by a model, and then be acted on by an agent. Security teams need visibility into that

Lower

BARRIER TO PASS

Indirect prompt injection attacks (via files or websites) succeed with significantly fewer attempts than direct attacks, bypassing standard filters.

Source: Laker Q4 2025 Agent Security Trends

Why Traditional Security Controls Fall Short

Most enterprise security architectures are built around a clear assumption: systems execute code, and users trigger actions through well-defined interfaces.

For decades, this assumption held. Inputs were structured. APIs had schemas. Security controls could inspect traffic, validate requests, and enforce access at predictable boundaries. AI systems break this model.

They ingest natural language. They infer intent. They combine user input with retrieved context and system instructions. The most security-critical decisions happen inside the reasoning process, not at the network boundary or API call.

Why Perimeter Controls Are Necessary — and Insufficient

Tool	What It Sees	What It Misses in AI
Network / WAF	Packets, HTTP requests	Prompt intent, reasoning paths
App Security	Schemas, validation	Indirect manipulation, context
DLP	Files, regex patterns	Semantic leakage in output
IAM	Roles, permissions	Unsafe autonomous sequences

Every layer is doing its job. None of them are watching behavior.

Model-Level Safeguards Cannot See the System

Model providers increasingly ship safety features. These safeguards can reduce certain categories of misuse, but they operate in isolation from enterprise context.

They do not know who the user is, what permissions they hold, which data the application has access to, or what downstream actions a response may trigger.

As a result, model-level protections cannot account for business logic, authorization, or operational impact. AI security failures often occur precisely where model behavior intersects with application context and system access.

The Common Failure Pattern

Across real-world incidents, a consistent pattern emerges. No single control fails outright. Instead, risk accumulates between layers. A prompt bypasses application logic. Retrieved context is subtly manipulated. An agent takes an action that appears reasonable in isolation.

Each step is technically valid. The outcome is not. What's missing is a way to observe, interpret, and constrain AI behavior as it unfolds.

PART II

The Solution

The Practical Model for Securing AI

14%

LIVE IN PRODUCTION

Agentic AI has moved beyond hype. 14% of organizations have already deployed autonomous agents into live production environments.

Source: Lakera 2025 GenAI Security Readiness Report

The Three New AI Exposure Surfaces

As AI systems move from assistance to action, the nature of exposure changes.

Risk is no longer confined to a single model, application, or user interaction. Instead, it emerges across distinct but interconnected surfaces, each with its own threat patterns and control requirements.



1. Workforce AI Interactions

Employees interacting with public GenAI, copilots, and assistants. Risk emerges from normal usage at scale: sensitive pastes, incomplete context, and implicit intent.



2. AI Applications

AI embedded directly into business logic. Models assemble prompts from dynamic sources. Risk shifts to prompt injection, indirect manipulation, and unintended disclosure via summarization.



3. Autonomous Agents

Agents with memory, planning, and tool access. Failures resemble operational incidents: records changed, permissions exercised incorrectly, actions taken beyond scope.

The Architectural Consequence

Controls can no longer focus on a single boundary. They must operate across interactions, applications, and agents. Nowhere is this shift more visible than in how agents execute actions across systems—a new critical security boundary.

4%

SECURITY CONFIDENCE

Only 4% of organizations rate their AI security confidence at the highest level, despite widespread adoption.

Source: Lakeria 2025 GenAI Security Readiness Report

MCPs and the New Agent Execution Surface

As AI systems move toward autonomy, the most important security shift is not happening at the model layer. It is happening at the execution layer.

Model Context Protocols (MCPs) define how language-driven systems discover capabilities, select tools, and translate intent into action. In doing so, they introduce a new and critical security surface.

MCPs as the “USB-C” of Agentic AI

Just as USB-C standardized how devices connect to power and peripherals, MCPs standardize how agents connect to tools, data, and systems. This brings faster integration, but also concentrates risk. When a single interface becomes the default path from reasoning to execution, it becomes the place where control either exists or disappears.

Why MCPs Matter Architecturally

MCPs sit at the junction where reasoning becomes execution, language becomes action, and autonomy becomes impact. Securing MCP-based systems does not mean removing autonomy. It means enforcing guardrails at the moment intent turns into action, based on context, permissions, and expected behavior.

99.8%

BLOCK RATE

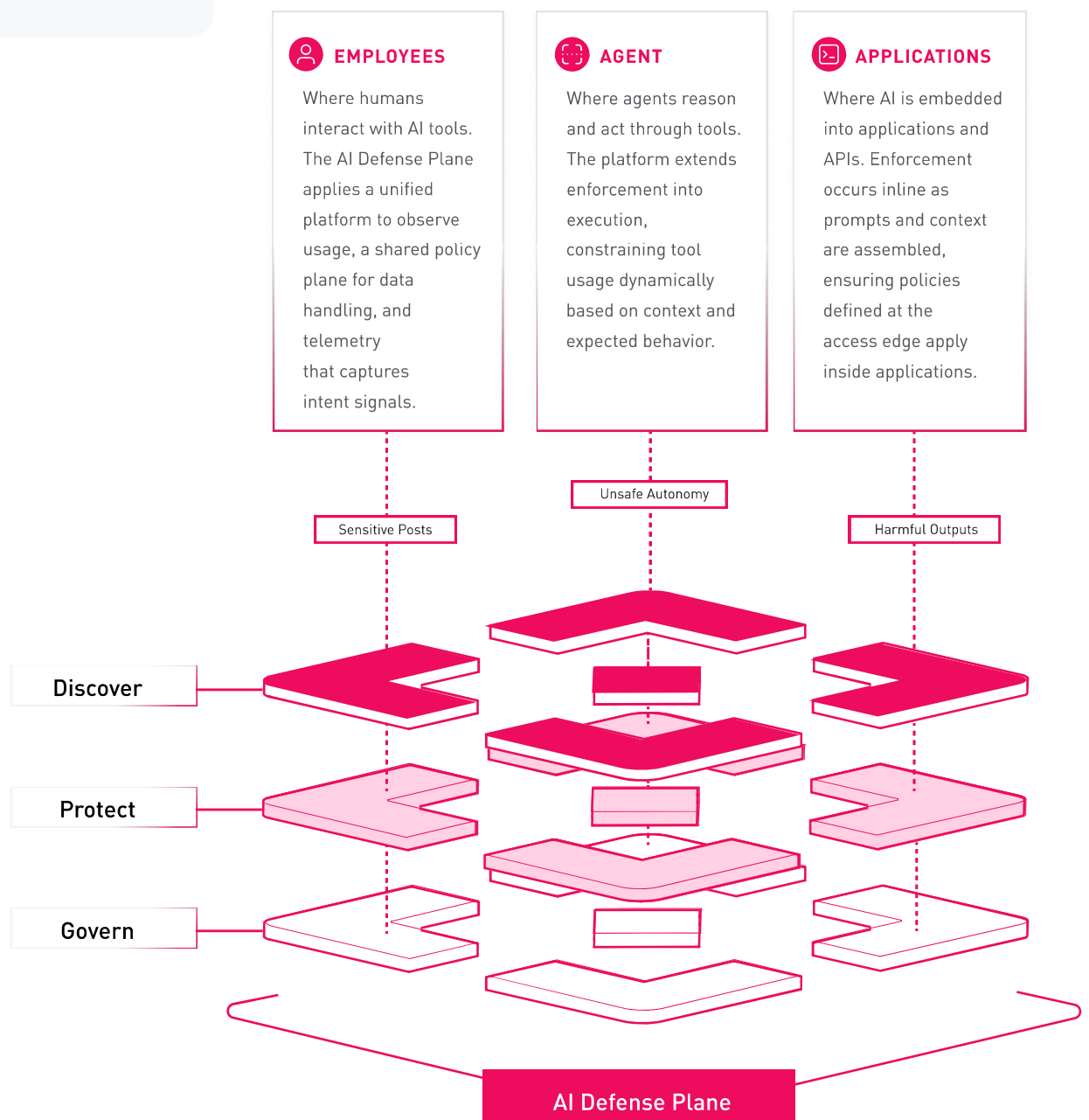
Unified defense planes that correlate signals across edges achieve a 99.8% block rate against new malware vectors.

[Source: Check Point](#)

The AI Defense Plane

AI systems that act cannot be secured with isolated controls.

What's required is not another point solution, but a unified security plane. One that coordinates visibility, policy, and enforcement across every place AI operates. This is the role of the AI Defense Plane.



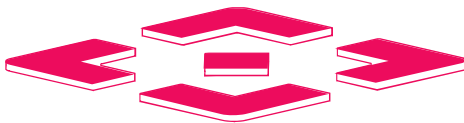
One platform. One lens. From employees to applications to agents.

The unified architectural model spanning workforce, apps, and agents.

One Platform, Three Layers of Protection

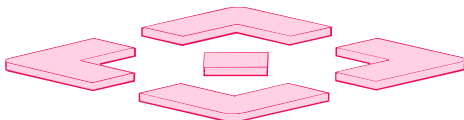
A unified security plane only works if it can be applied consistently across the environments where AI operates.

The AI Defense Plane is implemented as one coordinated platform that delivers protection across three distinct layers. Each layer addresses a different class of risk. Together, they provide end-to-end coverage.



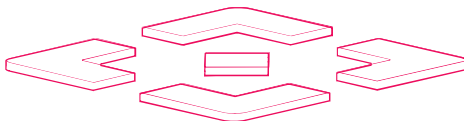
Layer 1: EMPLOYEES **Workforce AI Security**

Secures employee AI usage across browsers, devices, and enterprise tools. Provides visibility into which AI applications are in use, applies granular data handling policies inline, and detects risky interactions in real time, without slowing productivity.



Layer 2: AGENT **AI Agent Security**

Secures AI applications and autonomous agents at runtime. Inspects prompts, outputs, and agent actions inline to prevent prompt injection, data leakage, and unsafe autonomy, enforcing policy without retraining models.



Layer 3: AI SYSTEMS **AI Red Teaming**

Exposes AI failure modes before attackers do, and continuously as systems evolve. Simulates real-world attacks to uncover vulnerabilities in reasoning, workflows, and tool usage, delivering actionable remediation guidance across the AI lifecycle.

What Makes This a DEFENSE Plane

Policies are defined once. Telemetry is correlated across edges. Signals from one edge inform enforcement at another. This coordination turns separate controls into a unified defense.

PART III

In Practice

Proof, Production, and Scale

<40ms

AVERAGE LATENCY

Production benchmarks for AI Agent Security demonstrate an average latency of under 40 milliseconds, ensuring zero impact on user experience.

Source: Check Point

Proof, Production, and Scale: AI Security in Practice

Architectural models matter only if they hold up in production.

The AI Defense Plane is designed to secure systems that reason and act, but its value is measured by how it performs under real-world constraints: latency, scale, regulation, and evolving threat behavior.



Dropbox: Production Proof

Dropbox has been actively integrating AI into internal and customer-facing workflows, where reliability and safety are critical. As these systems moved from experimentation to production, Dropbox identified prompt injection and jailbreak attacks as a primary risk. Traditional moderation approaches proved insufficient.

Requirements & Results:

- Inline protection with ultra-low latency
- Centralized protection across multiple apps
- Low false positive rates on long-context prompts

Most importantly, runtime security did not slow development. Teams were able to iterate on AI features while maintaining guardrails aligned with evolving risk.



Nubank: Regulated Scale

Nubank serves more than 115 million customers across multiple geographies and operates in a highly regulated banking environment. As the company expanded its use of AI across financial services, security and compliance were foundational requirements.

Nubank selected Check Point to secure its enterprise AI deployments. As Dave Hannigan, CISO at Nubank, explained:

“We’ve chosen Check Point’s to secure our enterprise AI deployment... Check Point’s accuracy, low latency, seamless integration, scalability, and support for Portuguese and Spanish are essential for our global operations.”

This deployment highlights how AI security must extend beyond experimentation and into enterprise-wide operations, where latency, accuracy, and compliance are non-negotiable.

Check Point's Differentiated Approach

AI security is no longer about protecting isolated tools or individual models. It is about securing entire systems—systems that span employees, applications, and agents.

Most security approaches still treat AI as a collection of separate use cases: workforce AI handled at the browser, AI applications protected with ad hoc controls, and agents governed as experimental systems. The result is fragmented visibility and inconsistent enforcement.

Check Point takes a different approach. We deliver one coordinated platform that secures workforce AI interactions at the access edge, AI applications and APIs at the app edge, and autonomous agents and tools at the runtime edge. This unified model eliminates blind spots between layers and ensures that AI behavior is governed consistently as systems scale.

Integrated into the Check Point Ecosystem

By integrating AI Security capabilities into Check Point's broader security ecosystem, organizations gain unified visibility across network, application, identity, and AI layers.

The Bottom Line

AI systems that act require security systems that can see, understand, and govern action. With Check Point's AI-native intelligence, organizations gain a security model built for the next phase of AI adoption—where autonomy increases, but control is not lost.

[Assess your AI exposure](#)[See interactions in practice](#)[Red team before attackers do](#)