CHECK POINT™

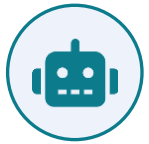
Von reaktiv zu kontinuierlich

Wie man AI-Applikationen, Agenten und Modelle schützt

Christian Voss-Freund | Sales Engineer | 16.07.2026

DAS PROBLEM

Warum reaktive Signaturen an ihre Grenzen stoßen



64%

der Unternehmen betreiben KI-Agenten bereits im Pilotbetrieb oder produktiv

12%

haben Agenten bereits privilegierten Zugriff auf Kernsysteme gewährt

77% / 26%

haben die Sicherheitsstrategie wegen KI angepasst – aber nur 26% haben die Architektur, sie durchzusetzen

2,3 J. → 10 Std.

so stark ist die Zeit von CVE-Offenlegung bis Exploit seit 2018 geschrumpft

PARADIGMENWECHSEL

Reaktiv & statisch → kontinuierlich & verhaltensbasiert

SIGNATURBASIERT

Bekannte Muster & Blacklists

Point-in-Time-Scans

Statische Severity-Scores (CVSS)

Reaktion erst nach dem Vorfall

→

KONTINUIERLICH & VERHALTENSBASIERT

Laufzeit-Analyse von Prompts, Tool-Calls & Aktionen

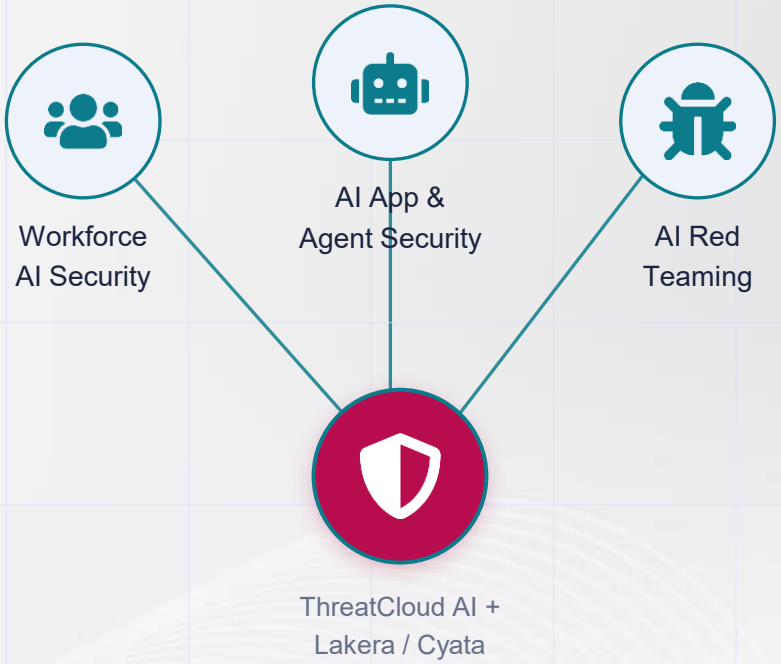
Kontinuierliches, automatisiertes Red Teaming

Evidenzbasierte Validierung statt Schätzung

Adaptive Kontrolle in Echtzeit (< 50 ms)

DIE KONTROLLEBENE

Check Point AI Defense Plane



Workforce AI Security

Sichtbarkeit & Governance für Mitarbeitende, die KI-Tools und Copilots nutzen.

AI Application & Agent Security

Discovery, Posture-Bewertung und Runtime-Kontrolle für Anwendungen und autonome Agenten (Cyata-Technologie).

AI Red Teaming

Kontinuierliches adversariales Testen von Prompts, Reasoning-Pfaden, Tool-Nutzung und Agentenverhalten.

KONTINUIERLICHE ERKENNUNG IN DER PRAXIS

Guardrails, Red Teaming & globale Threat Intelligence

AI Guardrails (ehem. Lakera Guard)

Laufzeitschutz gegen Prompt Injection, Jailbreaks und Datenlecks – verhaltensbasiert statt signaturbasiert, mehrsprachig, Reaktion < 50 ms.

Lakera Red / Check Point AI Red

Kontinuierliches, automatisiertes Red Teaming von Modellen und Agenten – deckt Schwachstellen vor der Produktivsetzung auf.

ThreatCloud AI

Globale Bedrohungsintelligenz aus Milliarden täglicher Ereignisse als Grundlage für die Bewertung von Verhaltensmustern.

Integration in bestehende Stacks

Z. B. Amazon Bedrock AgentCore: probabilistische KI-Erkennung von Check Point kombiniert mit deterministischer Gateway-Durchsetzung.

AI Guardrails: technische Architektur

Inline-Proxy vor dem Modell

Ein einzelner API-Endpunkt prüft jede Ein- und Ausgabe, bevor sie das Modell bzw. den Nutzer erreicht – kein Eingriff in Modell oder Prompts nötig.

Zwei-Schichten-ML-Architektur

Schnelle Klassifikator-/Heuristik-Schicht für bekannte Muster, plus tiefere Modell-Schicht für Kontext- und Verhaltensanalyse.

Modellagnostisch

Sitzt vor beliebigen LLM-Stacks (OpenAI, Anthropic, Bedrock, selbstgehostet) – ein Kontrollpunkt für heterogene KI-Landschaften.

Kontinuierliches Lernen

> 100.000 neue adversariale Samples pro Tag, u. a. aus der Gandalf-Community (80 Mio.+ gesammelte Prompts).

>98%

Erkennungsrate

<50 ms

Latenz (p50)

<0,5%

Falsch-Positiv-Rate

100+

Sprachen & Schriftsysteme

Erkennungspipeline & Policy Engine im Detail

Detektor-Typen

`prompt_attack` — Direkte & indirekte Prompt Injection, Jailbreaks, System-Prompt-Extraction

`moderated_content` — Hassrede, Gewalt, sexuelle Inhalte, Selbstschädigung u. a.

`pii / *` — Erkennung & Redaction von Namen, Adressen, Kontodaten

`secrets` — API-Keys, Tokens, Zugangsdaten in Ein- und Ausgabe

`custom pattern` — Kundenspezifische Regeln & Datenmuster

```
{ "policy_mode": "IO", "input_detectors": [
  { "type": "prompt_attack", "threshold": "l1_confident" },
  { "type": "pii/address", "threshold": "l2_very_likely" } ] }
```

Policy-Modi

I — nur Eingabe prüfen**O** — nur Ausgabe prüfen**IO** — Ein- und Ausgabe prüfen

Schwellenwerte (Threshold-Stufen)

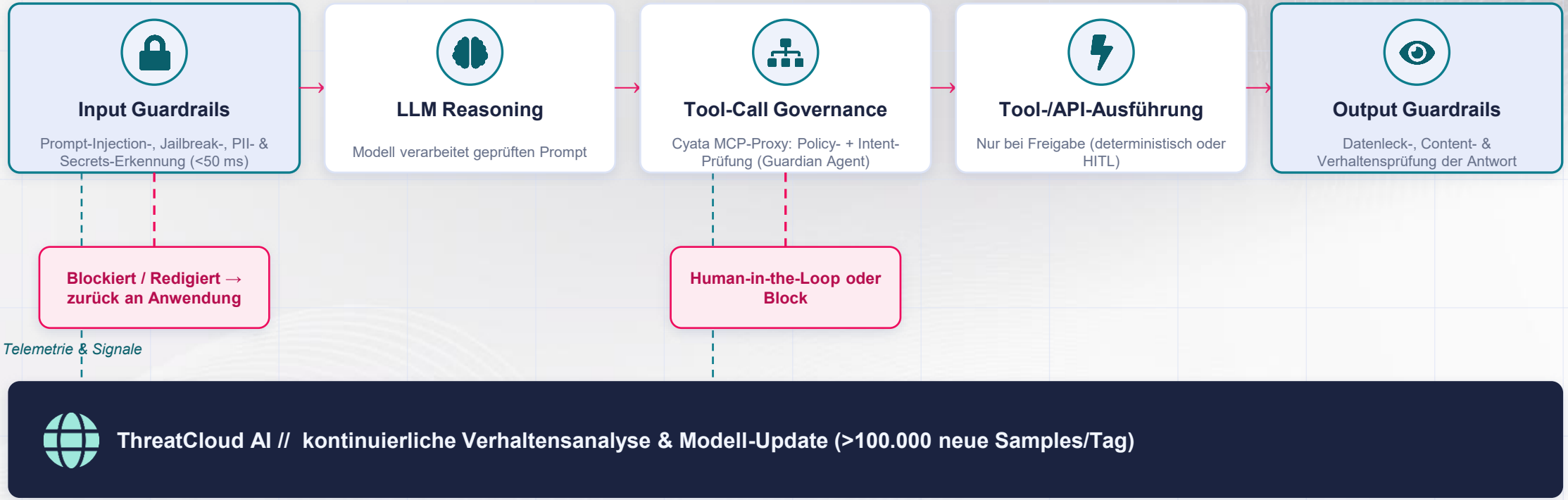
Granulare Stufen wie l1_confident oder l2_very_likely steuern das Verhältnis von Erkennungsschärfe zu Falsch-Positiv-Rate – projektspezifisch konfigurierbar.

Aktionen je Treffer

Block — Anfrage/Antwort stoppen**Flag / Warn** — durchlassen, aber protokollieren**Redact** — sensible Inhalte maskieren statt blockieren

WORKFLOW

Workflow: Request <> zu <> Aktion - Pfad



Cyata: Runtime Governance für autonome Agenten

Discovery

Non-Human-Identity über Endpoints, Browser, SaaS-Agentenplattformen und Cloud-native Deployments hinweg, mit Zuordnung zum menschlichen Owner.

Retrospektive Sichtbarkeit

Rekonstruktion von Agentenaktivität aus Telemetrie & Artefakten – auch ohne Proxy im Datenpfad, deckt bereits bestehende Agenten auf.

Deterministische Governance

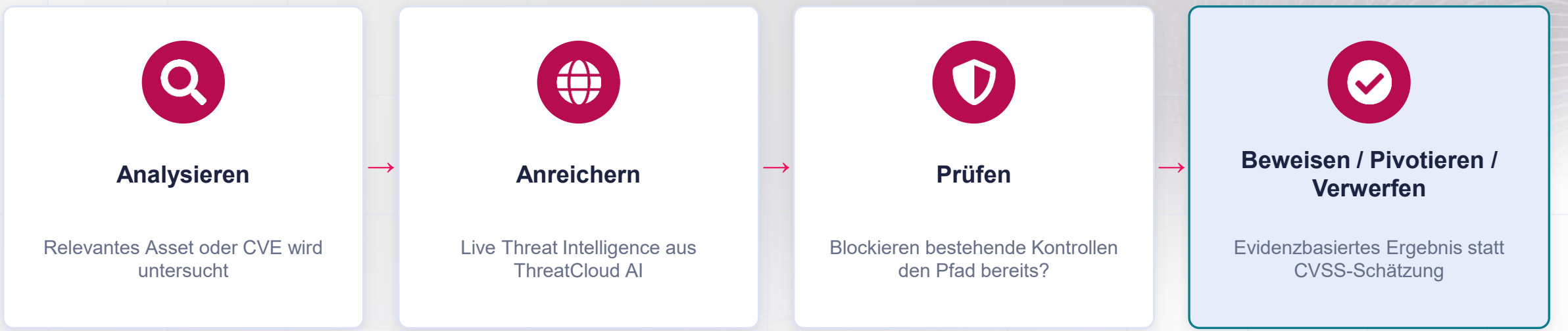
Zentrale Policy Engine legt erlaubte Tools & Scopes je Agent über alle Oberflächen hinweg fest und setzt sie durch.

Nicht-deterministische Governance

Guardian Agent verlangt vom handelnden Agenten eine Begründung für einen Tool-Aufruf und bewertet diese, bevor die Aktion freigegeben wird – alternativ Human-in-the-Loop.

MCP-Proxy: Cyata prüft Tool-Requests von Agenten vor der Ausführung über das Model Context Protocol (MCP) – ein expliziter Intent-Check statt reinem Allow/Block, wahlweise mit Human-in-the-Loop-Freigabe.

Agentic Exposure Validation (AEV): vom Fund zum Beweis



Teil von Check Point Infinity ERM / Exposure Management. KI-Agenten korrelieren Exposure-Daten, Asset-Kontext und Kontrollabdeckung, um zu bestimmen, ob eine Schwachstelle im konkreten Kundenumfeld tatsächlich ausnutzbar ist – kontinuierlich, statt einmalig gescannt.

SICHTBARKEIT ÜBER DEN LIFECYCLE

Vom Agenten bis zum Schutz der Mitarbeitenden

Cyata

Discovery & Observability für KI-Agenten: welche Agenten existieren, welche Tools/Daten sie nutzen, welche Aktionen sie ausführen (Non-Human Identity).

CloudGuard WAF

Schutz von APIs und Anwendungen, über die KI-Systeme angebunden sind, gegen Web- und API-Angriffe.

Harmony Email & Collaboration

Schutz vor KI-generiertem Phishing und Social Engineering in Postfach und Collaboration-Tools.

Browser Security

Absicherung der Interaktion von Mitarbeitenden mit externen KI-Tools direkt im Browser.

Gemeinsames Prinzip: Runtime-Kontrolle statt Point-in-Time-Prüfung – über Agenten, Anwendungen und Menschen hinweg.

ZUSAMMENFASSUNG

Check Point AI Security Portfolio im Überblick

Ebene	Check Point Produkte
Workforce AI Security	Harmony Email & Collaboration · Browser Security
AI Application & Agent Security	AI Guardrails (Lakera Guard) · Cyata · CloudGuard WAF
AI Red Teaming & Validation	Lakera Red / Check Point AI Red · Agentic Exposure Validation (Infinity ERM)
Fundament	ThreatCloud AI · Check Point AI Defense Plane

The Check Point Platform

Unified Security Management

Hybrid Mesh Network Security

- Hyperscale Data Center Gateway
- Cloud Network Security
- Web Application Firewall
- Perimeter and Enterprise Gateway
- Branch Gateway and SD-WAN
- SASE Internet Access
- SASE Private Access (ZTNA)

Workspace Security

- Email & Collaboration Security
- Endpoint Security
- Mobile Security
- Browser Security
- Extended Detection and Response (XDR)

Exposure Management

- External Attack Surface Management
- Threat Exposure Management
- Threat Intelligence Platform

AI Security

- Runtime Protection
- GenAI App Governance and DLP
- AI Data Center Security

ThreatCloud AI – Real Time Threat Prevention

Check Point Services

Herzlichen Dank!