

Governance für agentenbasierte Aktivitäten

Ein Architekturkonzept für den sicheren Einsatz von Agenten in Unternehmen

Wo sind meine Agenten?

Was können meine Agenten tun?

Was tun meine Agenten tatsächlich?

Wie reagiere ich?

Erstellt von den Gründungsmitgliedern der Blueprint Alliance

Eine branchenübergreifende Referenzarchitektur für CISOs, CIOs, CTOs und Enterprise Security Architects

Inhalt

1	Executive Summary
2	Einführung
3	Prinzipien des Konzepts
4	Referenzarchitektur der Blueprint Alliance
5	Operationalisierung mit globalen Systemintegratoren (GSIs)
6	Fazit: Gestaltung der Zukunft der KI-Sicherheit
7	Konzept in der Praxis
8	Best Practices für den Einstieg

Executive Summary

Für den traditionellen Sicherheitsperimeter von Unternehmen stellen KI-Agenten eine Herausforderung dar. Experimentelle LLM-Chatbots werden nach nur zwei Jahren von autonomen, mehrstufigen agentenbasierten Workflows verdrängt. Während Software früher nur tat, was man ihr sagte, entscheidet, agiert und vernetzt sie sich heute eigenständig. Diese Autonomie bietet einen echten geschäftlichen Mehrwert, bringt jedoch eine Governance-Herausforderung mit sich, für die die meisten Security-Teams noch nicht gerüstet sind.

[Gartner prognostiziert](#), dass ein durchschnittliches Unternehmen aus den weltweiten Fortune 500 bis 2028 über 150.000 Agenten im Einsatz haben wird. 2025 waren es noch weniger als 15. Dies wird zu einem erheblichen Wildwuchs bei Agenten, IT-Komplexität und Management-Herausforderungen führen. Moderne KI-Agenten zeigen aber nicht nur Daten an. Sie denken logisch, leiten weiter, rufen nachgelagerte APIs auf, erstellen Unteragenten und führen mehrstufige Operationen aus, die sich mitunter über zahlreiche Unternehmenssysteme erstrecken.

Ohne einen einheitlichen Architekturstandard wird sich die **Lücke bei den Identity-Management-, Sicherheits- und Governance-Maßnahmen** von Unternehmen vergrößern: Es wird mehr Schatten-Agenten geben, Anmeldedaten werden über vertrauenswürdige Grenzen hinweg vererbt und die Ausführung kann nicht mehr kontrolliert werden. Die Referenzarchitektur der Blueprint Alliance ist ein offenes Multi-Vendor-Framework und wurde für die offene Nutzung durch die Community veröffentlicht. Damit wird die Sicherheit vom Netzwerkperimeter auf eine einheitliche, agentenbasierte Zero-Trust-Kontrollebene verlagert. Außerdem erhalten Technologieverantwortliche eine konkrete Möglichkeit, agentenbasierte Bereitstellungen ohne Kontrollverlust zu skalieren.

Einführung

Die Unternehmenssicherheit basierte jahrzehntelang auf einer klaren Trennung zwischen menschlichen Identitäten und Maschinenidentitäten, wobei erstere per IAM, MFA, SSO und letztere über statische Service-Accounts, API-Keys oder Cloud-basierte Computing-Dienste verwaltet wurden. Autonome KI-Agenten heben diese Dichotomie auf. Wenn KI-Agenten im Namen menschlicher Mitarbeiter:innen agieren, Zugriffsberechtigungen erben, Vektordatenbanken abfragen und Finanztransaktionen auslösen, werden sie zu aktiven Sicherheits-Principals und Akteuren, die explizite Identity-, Zugriffs- und Governance-Kontrollen benötigen.

Sicherheit kann nicht allein auf der Modell- oder Netzwerkebene nachgerüstet werden. Vielmehr muss sie in einer Kontrollebene verankert werden, die die Transparenz, Konnektivität und Ausführung von Agenten kontinuierlich steuert.

Klassische Software (deterministisch)	KI-Agent (nicht-deterministisch)
Führt explizite schrittweise Codepfade aus	Wird allgemein über die menschliche Absicht informiert und formuliert mehrstufige Ausführungspläne
Nutzt statische, vordefinierte Berechtigungen und Service-Accounts	Wählt dynamisch APIs und Anwendungen aus, nutzt Token Vaults, Anmeldedaten-Broker und delegierte Sessions
Strikt beschränkt auf festcodierte Integrationspfade	Agiert autonom, nutzt Tools mithilfe des Model Context Protocol (MCP) und erzeugt Unteragenten
Agiert innerhalb authentifizierter Benutzer-Sessions	Arbeitet asynchron, häufig auch lange nach der Abmeldung der menschlichen verantwortlichen Person

Die Governance-Lücke:

Fünf Fragen, die jeder Vorstand stellt

KI-Agenten vermehren sich so schnell, dass die Security-Teams kaum den Überblick behalten können. [Gartner prognostiziert](#), dass ein durchschnittliches Unternehmen aus den weltweiten Fortune 500 bis 2028 über 150.000 Agenten im Einsatz haben wird. 2025 waren es noch weniger als 15. Dies wird zu einem erheblichen Wildwuchs bei Agenten, IT-Komplexität und Management-Herausforderungen führen. Nur 13 % der Unternehmen glauben, dass sie über ausreichend Governance-Maßnahmen für ihre KI-Agenten verfügen. Diese Lücke ist der Grund dafür, dass Ausschüsse für Cybersicherheit und Compliance immer wieder auf dieselben fünf Fragen zurückkommen.

1. Wo befinden sich unsere KI-Agenten und wer trägt die Verantwortung für ihre Aktivitäten?
2. Wie schränken wir ein, was Agenten ausführen können, und wie stellen wir sicher, dass sie ausschließlich im Rahmen ihrer Aufgaben agieren?
3. Können wir autonome Entscheidungen im Nachhinein auditieren und erklären?
4. Wie groß sind unser Datenkompromittierungsrisiko und unser Drittanbieterrisiko in der gesamten Agenten-Lieferkette?
5. Haben wir einen verifizierten Not-Aus-Schalter und einen Rollback-Plan für den Fall, dass ein Agent außer Kontrolle gerät?

Diese fünf Fragen lassen sich den vier Säulen der Referenzarchitektur der Blueprint Alliance in Abschnitt 3 zuordnen: „Erkennung und Identity-Registrierung“ beantwortet Frage 1, „Zugriffsmanagement“ beantwortet Frage 2, „Laufzeitautorisierung und -überwachung“ beantwortet die Fragen 3 und 4 und „Aktive Eindämmung“ beantwortet Frage 5. Indem CISOs auf diese fünf Fragen mit Fakten, statt lediglich Zusicherungen antworten, können sie die sichere Einführung von Agenten beschleunigen und dabei ihre Kontrollmaßnahmen an die regulatorischen und betrieblichen Risiken anpassen. Wer das nicht kann, bleibt für immer mehr und immer gravierendere Sicherheitsvorfälle anfällig.

Diese Fragen dienen als Diagnose-Tool für Risikobewertungen, nicht als einheitliche Checkliste. Sicherheitsanforderungen variieren naturgemäß je nach operativem Einsatzbereich, da ein Produktivitätsassistent, der ausschließlich über Leserechte verfügt, nicht dieselben Eindämmungskontrollen erfordert wie ein Agent, der Finanztransaktionen ausführen kann. Durch die Analyse der tatsächlich bestehenden Risiken können Security-Teams ihre Kontrollen an die reale Gefährdungslage anpassen. Dies bewahrt Unternehmen davor, risikoarme Bereitstellungen übermäßig komplex anzulegen und wegen Genehmigungen mehrmonatige Verzögerungen zu riskieren, und stellt gleichzeitig sicher, dass risikoreichen Agenten angemessene operative Grenzen auferlegt werden.

Prinzipien des Konzepts

Das Blueprint Alliance-Framework wendet Zero Trust auch bei Agenten an. Dabei werden Agenten als vollwertige Identitäten und nicht als Code, Anmeldedaten oder nachträgliche Ergänzung des menschlichen Identity-Stacks behandelt. Diese sechs Prinzipien bilden die Grundlage für alles Folgende.

Jeder gehostete Agent ist eine eigenständige Sicherheitsidentität, die von Anfang an mitgeplant werden muss. Agenten werden mit derselben Sorgfalt bereitgestellt, authentifiziert und deprovisioniert wie Mitarbeiter:innen und Service-Accounts. Alle vom Unternehmen gehosteten Agenten, einschließlich Pilotprojekte und interne Tools, sollten innerhalb eines zulässigen Governance-Frameworks agieren, wobei die Kontrollmaßnahmen ihrem Risiko und dem operativem Einsatzbereich entsprechen.

Zugriffe sind auf die jeweilige Aufgabe beschränkt und werden nicht dauerhaft gewährt. Während aufgabenbezogene dynamische Zuweisungen das Idealziel darstellen, kombinieren praktikable Implementierungen kurzlebige Aufgabenzuweisungen mit robusten grundlegenden Berechtigungen. Auf diese Weise wird der Verwaltungsaufwand in Grenzen gehalten.

Delegierungen sind durchgängig nachverfolgbar. Wenn ein Mensch Aufgaben an einen Agenten delegiert und dieser Agent einen Unteragenten erstellt, muss die Verantwortlichkeitskette bei jedem Schritt nachvollziehbar bleiben. Jede Aktion muss zu der Person oder dem Prozess zurückverfolgt werden können, die bzw. der sie autorisiert hat. Angesichts der heutigen Tools bleiben End-to-End-Delegierungen über mehrere Hops hinweg ein ehrgeiziges Ziel. Dennoch definiert dieses Konzept einen Referenzpfad, während sich die Standards noch weiterentwickeln.

Die Agentenlaufzeit wird isoliert und überwacht, nicht nur provisioniert. Zugriffsentscheidungen, die bei der Einrichtung getroffen werden, sind notwendig, aber nicht ausreichend. Das Konzept erfordert nicht nur die Überprüfung aller gewährten Berechtigungen, sondern auch eine kontinuierliche Beobachtung über Ausführungsläufe hinweg.

Die Isolierung erfolgt sofort und ist reversibel. Unabhängig von Anbieter und Plattform benötigt jeder Agent einen zielgerichteten Not-Aus-Schalter, der ihn in Echtzeit pausieren, unter Quarantäne stellen oder beenden kann, mit einem klaren Pfad für das Rückgängigmachen seiner Aktionen. Die Kombination aus kontinuierlichem verhaltensbezogenem Monitoring und einer strukturellen Ausführungsgrenze schließt die Lücke zwischen „autorisiert“ und „sicher“, da sie den Wirkungsradius für den Fall eindämmt, dass ein Agent von seiner Baseline abweicht.

Governance passt sich dem Tempo der KI an. Statische Sicherheitsrichtlinien können mit den sich schnell entwickelnden Fähigkeiten von Agenten nicht Schritt halten und die Governance muss einfangen, was Richtlinien allein nicht vorhersehen können. Dieses Konzept fungiert als anpassbare Architektur, die sich entsprechend den neuen Verhaltensweisen von Agenten skalieren lässt. Dabei werden statische Berechtigungsobergrenzen mit dynamischem Echtzeit-Scoping kombiniert, sodass Unternehmen strikte Leitplanken definieren und den Zugriff an den Kontext anpassen können.

In Summe wenden diese sechs Prinzipien Zero Trust auf eine neue Klasse von Identitäten an: Agenten, die konstant, dauerhaft und autonom im gesamten Unternehmen agieren. Wenn Sie diese Prinzipien richtig umsetzen, wird die Lücke bei den Identity-Management-, Sicherheits- und Governance-Maßnahmen geschlossen.

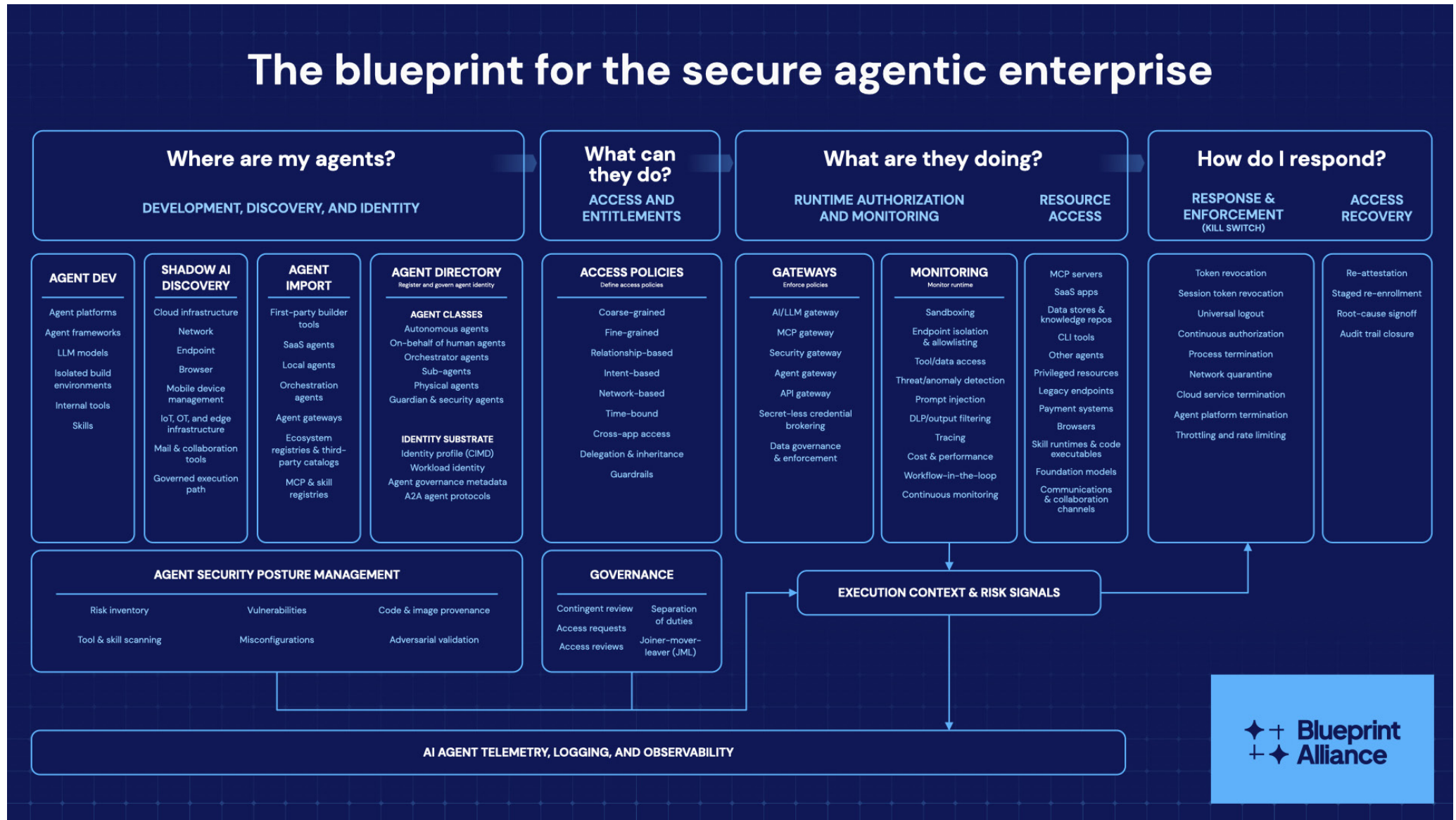
Referenzarchitektur der Blueprint Alliance

Die Referenzarchitektur der Blueprint Alliance stützt sich beim sicheren Einsatz von Agenten in Unternehmen auf vier operative Fragen, die für die Governance agentenbasierter Aktivitäten beantwortet werden müssen und jeweils zu einer Säule der Referenzarchitektur gehören:

1. Wo sind meine Agenten?
2. Was können meine Agenten tun?
3. Was tun meine Agenten tatsächlich?
4. Wie reagiere ich?

Über all diesen Fragen steht die Notwendigkeit eines soliden Fundaments aus Ausführungskontext und Risikoindikatoren, das kontinuierlich durch Laufzeitlemetrie, Logging und Beobachtbarkeit gespeist wird. Zusammen bilden sie das Bindeglied, mit dem das System Risikomuster erkennen und reagieren kann.

Referenzarchitektur der Blueprint Alliance



3.1 Säule 1:

Wo sind meine Agenten? (Entwicklung, Erkennung und Identity-Management)

Die grundlegende Herausforderung bei der Absicherung der Agentennutzung in Unternehmen ist Transparenz. Unternehmen können nichts verwalten oder schützen, von dem sie nicht wissen, dass es existiert. Diese Säule implementiert eine umfassende Tracking-Funktion, die in allen Unternehmensumgebungen jeden KI-Agenten identifiziert, katalogisiert und bewertet, der sich in der Unternehmensinfrastruktur bewegt, unabhängig davon, ob es sich um einen genehmigten oder nicht verwalteten Agenten handelt. Diese Säule endet mit der Registrierung validierter Agenten als verwaltete Identitäten mit zugewiesenen verantwortlichen Personen. Dadurch entsteht das maßgebliche Inventar, für das Säule 2 die Zugriffsrichtlinien festlegt.

3.1.1 Erkennung von KI-Agenten

Die Erkennung von KI-Agenten dient als kontinuierlicher Erkennungsmechanismus für alle im Unternehmen genutzten Agenten. Sie bildet den gesamten Lebenszyklus von Agenten ab – von internen Entwicklungspipelines bis hin zu externen Drittanbieter-Integrationen – und stellt sicher, dass kein aktives Modell und kein autonomer Prozess unbemerkt operiert.

3.1.1.1 Entwicklung von Agenten

Diese Komponente konzentriert sich auf interne Entwicklungsumgebungen, in denen proprietäre Agenten aktiv entwickelt, trainiert und optimiert werden. Die Absicherung der Entwicklungspipeline trägt dazu bei, die inhärente Sicherheit der Agenten zu gewährleisten, bevor sie zum ersten Mal mit Produktionsdaten interagieren.

- **Agenten-Plattformen:** Vom Unternehmen verwaltete Orchestrierungsumgebungen, in denen Agenten gehostet, trainiert und systematisch bereitgestellt werden.
- **Agenten-Frameworks:** Die zugrunde liegenden Softwarebibliotheken und Architekturen (z. B. LangChain, LlamaIndex oder AutoGen), mit denen Entwicklungsteams agentenbasierte Logik entwickeln.
- **LLM-Modelle:** Die grundlegenden und feinabgestimmten LLMs, die als kognitive Engines fungieren und die Agenten steuern.
- **Isolierte Erstellungs- und Ausführungsumgebungen:** Sichere, rechtechnisch isolierte CI/CD-Grenzen, die Agentencode während der Test- und Evaluierungsphase ausführen, damit Security-Teams kurzlebige Testausführungen erkennen, statt sich ausschließlich auf statische Registry-Deklarationen zu verlassen.
- **Zusätzliche interne Tools:** Benutzerdefinierte Programmcode-Repositorys, Code-Assistenten und interne Modell-Registries, die in die Entwicklungspipeline integriert sind.
- **Skills:** Definierte prozedurale Module, Code-Snippets oder Tools, die entwickelt wurden, um den Funktionsumfang von Agenten zu erweitern.

3.1.1.2 Import von Agenten

Moderne Unternehmen verlassen sich selten ausschließlich auf eine einzige Informationsquelle, sondern integrieren mehrere Agentenarchitekturen, die sich hinsichtlich Ursprung, Framework und Computing-Grenze unterscheiden und als Absicherung gegen die sich rasant wandelnden Fähigkeiten von Agenten dienen sollen. Beim Import von Agenten wird die Einführung der variierenden Agentenklassen im Unternehmensökosystem überwacht und gesteuert. Dabei wird stark darauf geachtet, wie sich die Daten basierend auf dem zugrunde liegenden Bereitstellungsmodell im gesamten Unternehmen bewegen.

Sie können bei jedem importierten Agenten nach seinem Typ, seiner Kategorie und seinem Kanal fragen. In der Regel liefert ein Agent alle drei Antworten auf einmal, z. B. ein (von einem Anbieter entwickelter) SaaS-Agent, der für interne Personal-Workflows (B2E) genutzt wird und über ein Agenten-Gateway eingebunden wurde.

Agententyp: Wer hat den Agenten entwickelt?

- **First-Party-Agenten:** Intern entwickelte KI-Funktionen. Diese reichen von vollständig benutzerdefinierten Implementierungen, die mit Rohcode (z. B. Python und LangChain) erstellt wurden und auf universellen Computing-Ressourcen (z. B. Kubernetes oder Serverless Functions) ausgeführt werden, bis hin zu Agenten, die mithilfe anbieterspezifischer Agenten-Frameworks entwickelt wurden und die Laufzeit, den Speicher sowie die Workflow-Orchestrierung nativ verwalten.
- **SaaS-Agenten:** Speziell entwickelte KI-Produkte von Drittanbietern, die als völlig unabhängige Einheiten innerhalb des Software-Ökosystems des Unternehmens agieren.
- **Lokaler Agent:** Client-Architektur, die direkt auf dem lokalen Gerät der Mitarbeiter:innen installiert ist, wobei jeder einzelne Logikschritt an einen Remote-Cloud-LLM-Anbieter gesendet wird. Sie fungiert als nicht-deterministische Ebene zwischen menschlicher Absicht und Tool-Aktion, wobei Daten bei jedem Inferenzaufruf die Unternehmensgrenze überschreiten.
- **Orchestrierungsagenten:** Agenten, die auf visuellen Orchestrierungsplattformen basieren und verschiedene Unternehmens-APIs miteinander verbinden.

Agentenkategorie: Für wen ist der Agent gedacht?

- **B2E-Agenten (Business-to-Employee):** Interne KI-Agenten für Mitarbeiter:innen, die im Rahmen interner Unternehmensidentitäten und rollenbasierter Zugriffskontrollen zur Automatisierung von Unternehmensworkflows, für IT-/HR-Abläufe und die tägliche Produktivität der Mitarbeiter:innen eingesetzt werden.
- **B2B-Agenten:** KI-Agenten für Unternehmen, die über Unternehmensgrenzen hinweg agieren, um Enterprise-to-Enterprise-Workflows, Lieferkettentransaktionen und Tenant-übergreifende API-Integrationen zu automatisieren.
- **B2C-Agenten:** KI-Agenten für Verbraucher:innen, die im Service, Handel oder Support eingesetzt werden, um direkt mit externen Endnutzer:innen zu interagieren, und die strenge Maßnahmen zum Schutz vor öffentlichen Prompt Injections, Gefahren für Marken und Diebstahl personenbezogener Verbraucherdaten erfordern.
- **B2B2C-Agenten:** White-Label-Lösungen oder KI-Agenten von Anbietern, die von Unternehmenskund:innen bereitgestellt werden, um die Endverbraucher:innen dieses Partners zu bedienen, und die im Rahmen einer mehrstufigen Isolierung sowie transitiver Vertrauensgrenzen agieren.
- **Personal Assistant-Agenten:** Von Benutzer:innen gesteuerte KI-Assistenten für Verbraucher:innen, die im Namen einer Einzelperson sowohl im privaten als auch im geschäftlichen Kontext agieren und Kontrollmechanismen zur Abgrenzung des Zugriffs zwischen privaten und Unternehmensdaten erfordern.

Kanal für Datenerfassung: Wie wird der Kanal gefunden?

- **Agenten-Gateways:** Erfassungspunkte und Proxys, über die diese externen oder importierten Agententypen mit internen Unternehmens-APIs interagieren.
- **Ökosystem-Registries und Drittanbieter-Kataloge:** Automatisierte Konnektor-Pipelines und API-Adapter, die Agentendefinitionen aus externen, herstellereigenen Registries kontinuierlich überwachen, abrufen und zuordnen, damit Agenten auch ohne manuelle Eingriffe automatisch katalogisiert werden können.
- **MCP- und Skill-Registries:** Erfassungspunkte, über die Agenten Tool-Schemas, Skills und Kontextdefinitionen aus öffentlichen oder von Anbietern gehosteten MCP- und Skill-Registries abrufen. Registries sind selbst ein Injektionsvektor. Ein schädlicher oder getarnter Eintrag kann automatisch abgerufen und verarbeitet werden, sodass jedes Tool oder jeder Skill aus einer Registry als nicht vertrauenswürdige Eingabe behandelt wird und gemäß Abschnitt 3.1.2 einem Lieferketten-Scan unterzogen wird, bevor ein Agent sie aufrufen kann..

3.1.1.3 Erkennung von Schatten-KI

Zur Optimierung ihrer Workflows greifen Mitarbeiter:innen oft auf nicht genehmigte KI-Tools zurück. Dies birgt Compliance-, Datenexfiltrations- und betriebliche Sicherheitsrisiken. Bei der Erkennung von Schatten-KI werden Unternehmensperimeter aktiv gescannt, um nicht autorisierte Bereitstellungen und Nutzungen von Agenten zu erkennen. Erkennungs- und Überwachungsfunktionen sollten in Abstimmung mit den geltenden Datenschutz-, Beschäftigungs-, Arbeits-, Betriebsvertretungs- und Datenschutzvorschriften implementiert werden. Ebenso müssen angemessene Kontrollen für Transparenz, Verhältnismäßigkeit, Datenminimierung, Speicherung und Zugriffe gelten.

- **Cloud-Infrastruktur:** Cloud-Computing-Instanzen, Serverless-Umgebungen und nicht verwaltete API-Endpoints, die durch aktive Netzwerk-/API-Überprüfungen entdeckt werden.
- **Netzwerk:** Echtzeit-Datenverkehrsanalyse und Egress-Monitoring zur Identifizierung unbekannter API-Aufrufe an externe LLM-Anbieter und KI-Endpoints.
- **Endpoint:** Überwachung auf Geräteebene, um die nicht autorisierte Ausführung von KI-Binärdateien, Skripten oder lokalen Modellgewichten zu tracken.
- **Browser:** Tracking von Erweiterungen und Überwachung von Browser-Sessions, um zu erkennen, wenn Benutzer:innen sich bei nicht genehmigten webbasierten KI-Plattformen oder Chat-Systemen anmelden.

- **Mobile Device Management (MDM) und Endpoint-Management:** Kontinuierliche Audits auf Geräteebene zur Erkennung von nicht genehmigten lokalen KI-Ausführungsdateien, lokalen Modellgewichten, CLI-Agent-Tools und IDE-Erweiterungen sowie zur Durchsetzung der grundlegenden Sicherheitslage und Blockierung nicht verwalteter Agenten-Laufzeitumgebungen auf Betriebssystemebene.
- **IoT-, OT- und Edge-Infrastruktur:** Sicherheitsüberwachung und Netzwerkprotokollüberprüfungen in Industrieumgebungen zur Erkennung von Small Language Models (SLMs) und autonomen Agenten in Fertigungsanlagen, SPSs und eingebetteter Hardware, wo herkömmliche Endpoint-Software nicht bereitgestellt werden kann.
- **E-Mail- und Zusammenarbeitsflow:** Analyse des Nachrichten-, Chat- und Kalenderdatenverkehrs zur Erkennung nicht genehmigter Agenten, die im Namen von Benutzer:innen Nachrichten senden, automatisch antworten oder Kalender verwalten, bevor die Aktivität in Endpoint- oder Netzwerk-Egress-Logs sichtbar wird.
- **Gesteuerter Ausführungspfad als Erkennungsfläche:** Architekturgrenze, an der alle genehmigten Agenten über eine zentral verwaltete Kontrollebene ausgeführt werden. Jeder agentenbasierte Prozess außerhalb dieser Grenze wird automatisch als Schatten-KI-Signal markiert, ohne dass separate umgebungsspezifische Erkennungsregeln erforderlich sind.

3.1.2 AI Agent Security Posture Management (AIASPM)

Nach ihrer Erkennung müssen Agenten kontinuierlich bewertet werden. AIASPM stellt die Diagnose-Ebene bereit, die Risikoprofile, Konfigurationshygiene und bekannte Softwareschwachstellen im gesamten Agentenbestand bewertet.

- **Risiko-Inventar:** Zentrale, dynamische Datenbank, die die Sicherheitslage, Schwachstellen, Softwareabhängigkeiten und Konfigurationsfehler über alle erkannten Agenten hinweg unabhängig vom Lebenszyklusstatus trackt und das IAM-fokussierte Agenten-Directory um eine Sicherheitsrisiko-Perspektive ergänzt.
- **Schwachstellen:** Standardverwaltung von Software-Schwachstellen, die auf agentenbasierte Abhängigkeiten und Modell-Endpoints ausgedehnt wurde. Sie deckt CVEs in der Codebasis des Agenten, der Lieferkette und den Basismodellen ab, die für eine konsistente Risikobewertung standardisierten Frameworks zugeordnet werden.
- **Code- und Image-Herkunft:** Kryptografische Überprüfung der Herkunft, Build-Integrität und Signaturen von Agenten-Code, Container-Images, Abhängigkeiten und Modellgewichten (z. B. Attestierungen des SLSA-Frameworks) vor einer Ausführung in der Produktion. Minimale, vorab attestierte Basis-Images, die von unnötigen Paketen bereinigt und anhand aktueller CVE-Daten neu erstellt wurden, reduzieren die Angriffsfläche der Lieferkette und vereinfachen die Herkunftsüberprüfung zum Erstellungszeitpunkt.
- **Scans der Tool- und Skill-Lieferkette:** Automatisierte statische und dynamische Überprüfung von Anweisungssätzen lokaler Agenten, Ausführungs-Skills und Remote-Tool-Schemas des Model Context Protocol (MCP), einschließlich Einträgen, die aus MCP- und Skill-Registries abgerufen wurden, um vor der Registrierung eingebetteten schädlichen Code, verdeckte Operationen und nicht autorisierte ausgehende Pfade zu erkennen.
- **Konfigurationsfehler:** Kontinuierliche Scans zur Erkennung von schwachen Parametereinstellungen, unsicheren System-Prompts, festcodierten, langlebigen Anmeldedaten, Identitätsdiebstahl-Vektoren und Konfigurationsabweichungen, die nach einer erfolgreichen anfänglichen Überprüfung eines Agenten auftreten.
- **Angriffsvalidierung:** Kontinuierliche, automatisierte Tests von Agenten und deren Leitplanken in Bezug auf bekannte Pfade für Prompt Injection-Angriffe, Jailbreaks und Datenexfiltrationen innerhalb temporärer, isolierter Ausführungsumgebungen, um die reale Anfälligkeit sicher zu bewerten, ohne die Produktion zu beeinträchtigen.

3.1.3 Agenten-Directory (Registrierung und Verwaltung der Agentenidentität)

Während beim Import von Agenten (Abschnitt 3.1.1.2) im Rahmen der Erkennung die Herkunft und die Metadaten erfasst werden, ist das Agenten-Directory das maßgebliche Identity-Repository, in dem für alle validierten, im Unternehmen gehosteten Agenten die kryptografischen Anmeldedaten, verantwortlichen Personen und Richtlinienberechtigungen formell registriert und verwaltet werden. Die Registrierung unterscheidet einen erkannten von einem verwalteten Agenten und ist die Voraussetzung für jede Zugriffsentscheidung in Säule 2.

Jeder gehostete Agent und jeder Unteragent mit spezifischen Berechtigungen muss hier mit einem verifizierten Profil, einer kryptografischen Identität, einer verantwortlichen Person und klaren strukturellen Beziehungen registriert werden, bevor er Zugriff auf Unternehmenssysteme anfordern kann. Gestalten Sie die Registrierung so reibungslos wie möglich, damit Teams nicht dazu verleitet werden, sie mithilfe gemeinsam genutzter Anmeldedaten zu umgehen. Lokale Agenten und kurzlebige Unteragenten, die den Berechtigungsbereich eines übergeordneten Elements erben, werden durch die unten beschriebenen Isolierungs- und Delegierungskontrollen gesteuert (und nicht durch einzelne Directory-Einträge). Sobald ein Agent importiert wurde, beantwortet das Directory zwei weitere Fragen zu seinen möglichen Handlungen in Ihrem Unternehmen: Welche Rolle spielt er und wie weist er seine Identität nach?

Agentenklassen: Welche Rolle spielt der Agent?

- **Autonome Agenten:** Vollständig unabhängige Agenten, die asynchron agieren, um ohne unmittelbare menschliche Aufsicht langfristige Geschäftsziele zu erreichen.
- **OBO (On-Behalf-Of)-Agenten für Menschen:** Agenten, die Aufgaben ausführen, die direkt an die Session bestimmter menschlicher Benutzer:innen gebunden sind, als Proxy fungieren und durch die zugrunde liegenden Berechtigungen dieser Benutzer:innen eingeschränkt sind.
- **Orchestrator-Agenten:** Übergeordnete Agenten, die ein Ziel in Teilaufgaben zerlegen und nachgelagerte Unteragenten erzeugen, koordinieren und überwachen. Das Directory muss Delegierungsketten und Verbreitungsbeziehungen als vollwertige Attribute tracken.
- **Unteragenten:** Spezialisierte untergeordnete Agenten, die von einem übergeordneten Orchestrator dynamisch erstellt oder aufgerufen werden, um diskrete Teilaufgaben auszuführen, und die durch vererbte Session-bezogene Berechtigungsbereiche statt individuelle Directory-Registrierungen gesteuert werden. Da innerhalb eines LLMs durchgesetzte modellinterne Aufgabengrenzen fehlschlagen können, müssen für Unteragenten isolierte Ausführungsgrenzen festgelegt werden, die den vererbten Berechtigungsumfang strukturell durchsetzen und damit verhindern, dass Logikfehler zu einer Rechteauserweiterung führen.

- **Physische/verkörperte Agenten:** Physische Agenten, die Roboter, Antriebe oder andere physische Systeme steuern. Diese Architektur kontrolliert Identity-Management und Zugriffe und soll vorhandene Kontrollen für Betriebssicherheit, OT-Sicherheit und Änderungsverwaltung nicht ersetzen, sondern ergänzen.
- **Guardian- und Security-Agenten:** Spezialisierte administrative KI-Agenten, die auf der Sicherheitskontrollebene agieren (z. B. SOC-Triage-Agenten, dynamische Richtlinien-Engines, kontinuierliche AIASPM-Scanner). Aufgrund ihrer erweiterten Berechtigungen, Telemetriedaten zu überprüfen und Isolierungsaktionen auszuführen, erfordern Guardian-Agentenidentitäten zwingend die kryptografische Attestierung der Workload, eine strikte Aufgabentrennung und kontinuierliche Freigaben durch die verantwortlichen Personen.

Identity-Substrat: Was beweist, dass er wirklich dieser Agent ist?

- **Metadatenprofil zur Agentenidentität:** Umfangreiche Metadatenprofile (CIMD), die Absicht, Herkunft, Abstammung und den operativen Kontext des Agenten definieren, um risikobasierte Zugriffsentscheidungen zu unterstützen.
- **Workload-Identität:** Kryptografische, plattformunabhängige, standardbasierte Identity-Produktionsframeworks, die für Agenten in dynamischen Cloud-Umgebungen sichere, verifizierbare IDs ausstellen.

- **Agenten-Governance-Metadaten:** Strukturierte Metadaten, die an den Directory-Eintrag des Agenten gebunden sind, einschließlich der verantwortlichen menschlichen Person, des zugewiesenen operativen Teams oder der Rufbereitschaft, des Lebenszyklus-Status (z. B. aktiv, gesperrt, stillgelegt), der Risikostufe und des operativen Berechtigungsbereichs. Diese Daten werden von Richtlinien-Engines für die Erstellung der Zugriffskontrolle und die Laufzeitauswertung kontinuierlich verwendet.
- **A2A-Agentenprotokolle:** KI-Agenten, die standardisierte Frameworks für die Interoperabilität nutzen, um Fähigkeiten dynamisch zu erkennen, Aufgaben zu delegieren und über Unternehmensgrenzen hinweg peer-to-peer zusammenzuarbeiten. Wenn Agenten über Unternehmensgrenzen hinweg ohne gemeinsames Directory oder gemeinsame Vertrauensbasis interagieren, nutzen sie kryptografische Standards für die Attestierung und den Token-Austausch (z. B. WIMSE/AIMS, OAuth Token Exchange, HTTP Message Signatures), um Delegierungsketten aufrechtzuerhalten und die Hauptidentität über Zwischensysteme hinweg zu verifizieren.

3.2 Säule 2:

Was können meine Agenten tun? (Zugriff und Berechtigungen)

Autonome Agenten wurden für Aktionen konzipiert und benötigen dafür Verbindungen zu Daten, APIs und Systemen. Aufbauend auf den in Säule 1 etablierten verifizierten Identitäten definiert diese Säule die Grenzen und Berechtigungen, die regeln, wie Agenten mit Unternehmensressourcen interagieren – sowohl in direkten, von Menschen delegierten Sessions als auch in komplexen Multi-Agenten-Workflows. Dabei werden menschenzentrierte Legacy-Modelle um Machine-to-Machine-KI-Governance ergänzt.

3.2.1 Zugriffsrichtlinien (Definition der Zugriffsrichtlinien)

Zugriffsrichtlinien legen die operativen Grenzen fest, was ein Agent mit verifizierter Identität lesen, schreiben oder ausführen darf. Aufgrund des unvorhersehbaren Verhaltens von generativer KI müssen diese Richtlinien stark kontextabhängig und adaptiv sein.

- **Grobkörnig:** Umfassende, rollenbasierte Kontrollen, die definieren, mit welchen übergeordneten Umgebungen oder Makroservices sich ein Agent verbinden kann.
- **Feinkörnig:** Attributbasierte und objektbezogene Berechtigungen, die einen Agenten auf bestimmte Datenbankzeilen, Dateien oder spezifische API-Methoden beschränken. Dabei werden Klassifizierungslabells für Unternehmensdaten (z. B. Vertraulich, personenbezogene Daten) explizit berücksichtigt, um richtlinienkonforme Datengrenzen durchzusetzen. Richtlinienkonstrukte können Datenklassifizierung, Agentenidentität, Risikoumfang, Umgebungskontext und Aktionstyp umfassen.

- **Beziehungsbasiert:** Zugriff, der sich aus den Graph wer ihn erstellt hat, in wessen Auftrag er handelt und wem er untergeordnet ist.
- **Absichtsbasiert:** Fortschrittliche Richtlinien-Engines, die die semantische Absicht des Prompts oder Plans des Agenten analysieren, bevor sie Zugriff auf Ressourcen gewähren. Pläne können mit den aktuellen Tools inspiziert werden. Zu glauben, dass ein Agent nur das tut, was Benutzer:innen eins zu eins im Prompt eingeben, ist illusorisch, weil Agenten nicht nur auf Prompts reagieren. Kurzlebige, zeitgebundene Berechtigungen werden ausschließlich für die Dauer einer bestimmten Aufgabe erstellt und nach Abschluss sofort widerrufen.
- **Netzwerkbasier:** Makrosegmentierung, die den Datenverkehr von Agenten über Unternehmens-VPCs hinweg einschränkt, kombiniert mit L3/L4-Mikrosegmentierungskontrollen, die die Kommunikation zwischen Agent und Ziel auf autorisierte IP-Adressbereiche oder Service Meshes beschränken.

- **Zeitgebunden:** Zeitgebundene Berechtigungen sind auf eine einzelne Aufgabe oder Session beschränkt und werden nach Abschluss oder Zeitüberschreitung automatisch entzogen. Da automatisch genehmigte, automatisch ablaufende Berechtigungen für sich genommen nur einen begrenzten Kontrollwert bieten, kombinieren Sie das zeitgebundene Scoping mit der Anomalieerkennung aus Abschnitt 3.3.1.2, statt den Ablaufzeitpunkt als einzige Schutzmaßnahme zu betrachten.
- **Anwendungsübergreifender Zugriff:** Richtlinien-Frameworks und Begrenzungskontrollen, die die Fähigkeit eines Agenten steuern, verschiedene Anwendungsökosysteme zu durchlaufen, und so eine sichere Datenübersetzung, den Token-Austausch und die Ausführung von Funktionen ermöglichen, während er Workflows über verschiedene Unternehmensplattformen hinweg orchestriert.
- **Delegierung und Vererbung:** Die effektive Berechtigungsgrenze für einen Unteragenten wird als Schnittmenge seiner eigenen Berechtigungen und der Berechtigungen des delegierenden Agenten oder Menschen berechnet, um eine Confused-Deputy-Eskalation (verwirrter Stellvertreter) zu verhindern. Die Richtlinie muss die Delegierung zudem in ihrer Breite und Tiefe begrenzen, damit sie sich nicht ungehindert ausbreiten kann.
- **Leitplanken:** Harte, deterministische Sicherheitseinschränkungen und Verhaltensgrenzen, die autonome Entscheidungsschleifen außer Kraft setzen, um nicht konforme Aktionen, toxische Ausgaben oder hochriskante operative Schritte global und unabhängig von der Absicht oder den feinkörnigen Berechtigungen des Agenten zu verhindern.
- **Bedingte Überprüfung:** Lösen Sie automatisierte oder Human-in-the-Loop-Workflow-Genehmigungen aus, sobald die Aktion eines Agenten festgelegte finanzielle oder operative Schwellenwerte überschreitet (z. B. ein Kauf oder eine Buchung über 1.000 USD), unabhängig davon, ob die Aktion im Rahmen der gewährten technischen Berechtigungen zulässig ist.
- **Zugriffsüberprüfungen:** Automatisierte, kontinuierliche Berechtigungs-scans, die Systemverantwortliche bei der Validierung nicht konformer Berechtigungen unterstützen.
- **Zugriffsanforderungen:** Formelle, auditierte Workflows zur Provisionierung neuer Funktionen, Modelle oder Systemzugriffe für einen bestehenden Agenten. Während Anfragen von menschlichen verantwortlichen Personen oder von Agenten dynamisch zur Laufzeit gesendet werden können, müssen von Agenten initiierte Anfragen streng anhand statischer Berechtigungsobergrenzen evaluiert werden. Bei einer Erweiterung des Berechtigungsbereichs erfordern sie eine obligatorische Verifizierung durch die verantwortliche Person, um zu verhindern, dass Agenten automatisierte Workflows manipulieren, um überhöhten Zugriff zu erhalten.
- **Aufgabentrennung (SoD):** Algorithmische Regeln, die verhindern, dass ein einzelner Agent über widersprüchliche Berechtigungen verfügt, z. B. Erstellen UND Genehmigen von Finanztransaktionen.
- **Joiner-Mover-Leaver (JML):** Wenden Sie bei Veränderungen des Berechtigungsbereichs, der verantwortlichen Personen oder der Modellversion eines Agenten dieselben strikten Prinzipien für Onboarding, Change-Management und Deprovisionierung an, die für menschliche Identity-Lebenszyklen gelten.

3.2.2 Governance

Governance bildet die Compliance-Grundlage und eine kontinuierliche Überwachungsschleife. Dabei werden automatisierte Tools anstelle von manuellen Pro-forma-Prüfungen eingesetzt, um die Berechtigungen von Agenten in Agentengeschwindigkeit mit dem Least-Privilege-Prinzip im Einklang zu halten.

3.3 Säule 3:

Was tun meine Agenten tatsächlich? (Laufzeitautorisierung und -kontrolle)

Transparenz und Identity-Management sind ohne integrierte Laufzeitüberwachung und -durchsetzung wirkungslos. Diese Säule konzentriert sich darauf, die Aktionen von Agenten bei der Ausführung von Aufgaben in Echtzeit abzufangen, zu überprüfen und zu autorisieren. Dadurch wird sichergestellt, dass sie kein schädliches Verhalten zeigen, nicht unsachgemäß mit sensiblen Daten umgehen oder kritische Zielsysteme des Unternehmens kompromittieren.

3.3.1 Laufzeitautorisierung und -überwachung

Die Laufzeitautorisierung und -überwachung fungiert als aktive Überprüfungsebene für Live-Sessions von Agenten. Sie befindet sich direkt im Ausführungspfad und analysiert sowohl die Eingaben, die in den Agenten einfließen, als auch die zurückgegebenen Aktionen oder Ausgaben.

3.3.1.1 Gateways (Durchsetzung der Zugriffsrichtlinien)

Gateways fungieren in einer abgestuften Verteidigungskette als Richtlinien-Durchsetzungspunkte (PEPs). Sie fangen den Datenverkehr zwischen der Agenten-Ebene und den restlichen Technologien des Unternehmens ab, parsen Protokolle und blockieren nicht autorisierte Aufrufe, bevor diese ihr Ziel erreichen. Nachgelagerte API-Gateways behalten letztendlich die Autorisierungshoheit und validieren jeden Aufruf ihrerseits, unabhängig davon, was vorgelagerte KI-/Agent-Gateways bereits geprüft haben. KI/LLM-, MCP-, Security-, Agenten- und API-Gateways werden oft zusammen bereitgestellt, aber jedes verfügt über einen eigenen Kontrollpunkt:

KI/LLM-Gateways steuern den Modell-API-Datenverkehr; MCP-Gateways steuern den Datenverkehr von Tool-/Kontextfreigabeprotokollen; Security-Gateways führen protokollunabhängige Bedrohungsprüfungen durch; Agenten-Gateways regeln agentenspezifische Identity- und Delegierungsbelange und API-Gateways setzen für die Tools und Dienste, die Agenten aufrufen, klassische Kontrollen auf API-Ebene durch.

- **KI/LLM-Gateway:** Fängt Low-Level-API-Aufrufe an grundlegende Modelle ab und verwaltet Ratenbegrenzung, Token-Anzahl und Außerkraftsetzungen von System-Prompts.
- **MCP-Gateway (Model Context Protocol):** Spezialisierter Proxy, der für die Verwaltung von Protokollen zur Kontextfreigabe und von Integrationen offener Standards zwischen Agenten und Anwendungen optimiert wurde. Der Schwerpunkt liegt dabei auf MCP-Servern.
- **Security-Gateway:** Eine direkt integrierte Ebene für die Bedrohungsabwehr, die dedizierte Content-Security und Protokollverifizierung für Agenteninteraktionen bietet. Es wird direkt im Datenverkehrsflow platziert, um Payloads nach dem Aufbrechen der TLS-Verschlüsselung in Echtzeit eingehend zu inspizieren. Dabei fängt es komplexe Prompt Injections, Jailbreaks, versteckte schädliche Textzeichenfolgen und Payloads mit schädlichem Code aktiv ab und filtert sie heraus, bevor sie den Agenten oder nachgelagerte Ressourcen erreichen. Darüber hinaus scannt es die Datenklassifizierung, um ausgehende Daten zu redigieren, die strukturelle Integrität von Modellausgaben zu kontrollieren und verschlüsselte Payloads von Agenten kryptografisch zu inspizieren.

- **Agenten-Gateway:** Dient als agentenspezifische Laufzeit-Kontrollebene: Es bindet die Agentenidentität an jede Anfrage, verwaltet die Tool-/MCP-Registrierung, auf die Agenten zugreifen, und steuert Agentenaktionen (nicht nur den Zugriff) basierend auf Identität, Kontext und spezifischer angeforderter Aktion. Es stoppt unsichere Interaktionen in Echtzeit, indem es riskanten Modelldatenverkehr blockiert oder bereinigt und sensible Tool-Aktionen mithilfe integrierter und benutzerdefinierter Richtlinien verweigert oder genehmigungspflichtig macht. Zudem injiziert es zur Ausführungszeit Secret-freie, kurzlebige nachgelagerte Anmeldedaten in API-Aufrufe, sodass Agenten niemals dauerhafte Enterprise-API-Keys oder Service-Account-Secrets handhaben, speichern oder offenlegen. Das Gateway kann erfasste Telemetriedaten für unterstützte Multi-Agenten-Interaktionen bereitstellen und vertrauliche Daten (z. B. personenbezogene Informationen) erkennen oder löschen. Dies erfolgt entsprechend konfigurierten Richtlinien und Funktionen und schützt vor unsicherem Verhalten und promptbasierten Angriffen.
- **API-Gateway:** Erzwingt klassische Kontrollen auf API-Ebene, Ratenbegrenzung, Kontingente, Schema-Validierung, mTLS, WAF, Traffic Shaping und Widerruf am Gateway für die Tools und Dienste, die Agenten und MCP-Server nutzen. Dabei wird bei jedem Aufruf eine eigene abschließende Autorisierungsprüfung durchgeführt, anstatt davon auszugehen, dass vorgelagerte KI-/Agenten-Gateways bereits jede Anomalie erkannt haben.

3.3.1.2 Überwachung (Laufzeitüberwachung)

Die Überwachung ermöglicht umfassende Beobachtbarkeit, Bedrohungs-erkennung und aktive Durchsetzung, indem Laufzeit-Telemetriedaten auf ungewöhnliches Verhalten, Datenlecks und externe Ausnutzungsversuche überprüft und in Echtzeit Kontrollen wie Sandboxing und Endpoint-Isolierung angewendet werden.

- **Sandboxing:** Isolierung der Agentenausführung (insbesondere von nicht vertrauenswürdigen, von Drittanbietern stammenden oder neu aktualisierten Agenten) in sicheren, eingeschränkten Laufzeitumgebungen, um laterale Bewegungen zu verhindern und Ausführungsrisiken durch strukturelle Grenzen einzudämmen.
- **Endpoint-Isolierung und Allowlisting:** Lokale Laufzeitkontrollen für auf Endpoints gehostete Agenten, einschließlich Sandboxing auf Betriebssystemebene, Allowlisting von Endpoint-Anwendungen und lokaler Überprüfungen des Gerätestatus, um die Befehlsausführung lokaler Agenten und den Dateizugriff einzuschränken.
- **Tool-/Datenzugriff:** Strukturelle Überprüfung und Auditierung in Echtzeit von Argumenten, Payloads, Schemas und Abfragen, die an nachgelagerte Tools und Repositories übergeben werden. Die Durchführung von Überprüfungen innerhalb oder direkt an der Ausführungsgrenze gewährleistet, dass dynamische Ausführungspläne nicht von autorisierten Parametern abweichen oder während der Session unbeabsichtigte Änderungen verursachen.
- **Bedrohungs-/Anomalieerkennung:** Machine-Learning-Engines, die Verhaltens-Baselines von Agenten überwachen, um ungewöhnliche Zunahmen bei Datenzugriffen, schnelle API-Schleifen oder auffällige Ausführungsmuster zu melden.

- **Prompt Injection:** Inline-Scanner, die darauf ausgelegt sind, direkte und indirekte schädliche Textzeichenfolgen abzufangen, die die Kernlogik des Agenten kapern oder Systemanweisungen außer Kraft setzen sollen.
- **DLP/Ausgabefilterung:** Tools zur Datenverlustprävention, die die Ausgaben von Agenten scannen, um das Risiko eines Abflusses von proprietärem Quellcode, personenbezogenen Daten oder Geschäftsgeheimnissen an externe Anbieter oder unbefugte Benutzer:innen zu mindern.
- **Tracing:** Umfassende Transaktionszuordnung, die den genauen Ausführungspfad, die Call-Stacks und die Aufrufe von Unteragenten darstellt, die durch einen primären Benutzer-Prompt ausgelöst werden.
- **Kosten und Leistung:** Überwachung von Token-Nutzung, Betriebskosten, Latenz und Ressourcenverbrauch, um außer Kontrolle geratene Schleifen von Agenten oder Angriffe zur Ressourcenerschöpfung zu verhindern.
- **Workflow-in-the-Loop:** Programmgesteuerte Auslöser, die die Ausführung des Agenten bei risikoreichen Aktionen pausieren und an automatisierte Richtlinien-Gegenprüfungen oder eine explizite menschliche Freigabe weiterleiten, bevor sie fortgesetzt werden.
- **Kontinuierliche Überwachung:** Kontinuierliche Überwachung und Exportierbarkeit der Aktivitäten des Agenten, einschließlich granularer Nachverfolgung, Audit-Trail-Protokollierung und Nachweisbarkeit der Agentenaktionen.

3.3.1.3 Kontinuierliche Durchsetzung von Absicht und Ausrichtung

Die kontinuierliche Durchsetzung von Absicht und Ausrichtung bestimmt, ob ein Agent während der gesamten Ausführung innerhalb der Grenzen seiner delegierten Aufgabe bleibt oder in eine unbefugte Zweckausweitung abgewichen ist. Dies ergänzt die Richtlinien für den Erstzugriff: Die absichtsbasierte Richtlinie bestimmt, worauf ein Agent zunächst zugreifen darf, während die kontinuierliche Durchsetzung der Ausrichtung bewertet, ob sich entwickelnde, mehrstufige Verhaltensweisen im Laufe der Zeit angemessen bleiben.

Zu den wichtigsten Funktionen gehören:

- **Zweckbindung:** Durchsetzung einer strikten Verknüpfung zwischen delegierenden Benutzer:innen, der Agentenidentität, dem genehmigten Workflow und der angeforderten Aufgabe.
- **Vergleich von Plan und Aktion:** Laufzeitvergleich geplanter Tool-Aufrufe mit beobachteten Aktionen, um Aufgabenabweichung, Injection-bedingte Abweichungen und Confused-Deputy-Verhalten zu erkennen.
- **Kontextbezogene Durchsetzung:** Ausführung von Echtzeitaktionen wie Zulassen, Verweigern, Löschen, Step-up-Genehmigung, Einschränkung des Berechtigungsbereichs, Sperrung oder Eindämmung.
- **Nachvollziehbare Entscheidungsnachweise:** Führen unwiderlegbarer Audit-Protokolle, die für jede wesentliche Aktion erklären, warum diese zugelassen, geändert oder blockiert wurde.

3.3.2 Ressourcenzugriff

Der Ressourcenzugriff definiert den tatsächlichen Wirkungsradius eines Agenten. Er kategorisiert die diversen nachgelagerten Ziele, mit denen ein Agent interagieren kann, und stellt sicher, dass jede Schnittstelle von geeigneten Sicherheits-Wrappern umgeben ist.

3.3.2.1 Zielressourcen (Komponenten, auf die zugegriffen wird)

Zielressourcen stellen die Zielorte der Aktionen eines Agenten dar. Unabhängig davon, ob Kontextinformationen abgerufen oder Änderungen übertragen werden, müssen diese Endpoints kontinuierlich katalogisiert und geschützt werden.

- **MCP-Server:** Daten-Repositorys und Dienstprogramm-Server, die explizit dafür konfiguriert sind, über das Model Context Protocol Kontextinformationen (strukturierte Daten, Dokumente, Einbettungen, Tools und Status) an Agenten weiterzugeben.
- **SaaS-Apps:** Produktivitäts- und zentrale Geschäftsanwendungen für Unternehmen, die große Mengen sensibler Unternehmensdaten enthalten.
- **Datenspeicher und Wissensdatenbanken:** Strukturierte und unstrukturierte Speicher, einschließlich Vektordatenbanken, Data Lakes, Warehouses und transaktionaler Speichersysteme, die von Agenten genutzt werden, um geschäftlichen Kontext zu extrahieren oder operative Ergebnisse festzuschreiben. Diese Repositorys unterliegen Praktiken für das Data Security Posture Management (DSPM), die Herkunft und Klassifizierung nachverfolgen, während Agenten auf die einzelnen Speicher zugreifen.

- **CLI-Tools:** Befehlszeilenschnittstellen, die es Agenten ermöglichen, Skripte auf Systemebene, Konfigurationen oder operative Befehle auszuführen.
- **Weitere Agenten:** Nachgelagerte Unteragenten oder spezialisierte Agenten, die Teilkomponenten eines umfassenderen Workloads ausführen sollen.
- **Privilegierte Ressourcen:** Hochsensible Zielumgebungen, einschließlich Stammverzeichnissen, Identity Stores und Master-Konfigurationskonsolen.
- **Standard- und Legacy-Endpoints:** Enterprise-REST/GraphQL-APIs, Datenbanken und Legacy-Anwendungen, die agentenbasierte Kontextprotokolle wie MCP nicht nativ unterstützen.
- **Zahlungssysteme:** Finanznetzwerke, Transaktions-Gateways und Buchhaltungsjournale erfordern ein Höchstmaß an kryptografischer Validierung und menschlicher Aufsicht.
- **Browser:** Automatisierte Web-Sessions, die von Agenten genutzt werden, um öffentliche Informationen zu extrahieren oder mit webbasierten externen Konsolen zu interagieren.
- **Skill-Laufzeitumgebungen und ausführbare Code-Dateien:** Code-Interpreter, Skriptausführungs-Sandboxen und prozedurale Module, die von Agenten dynamisch aufgerufen werden, um generierten Code auszuführen, und die isolierte Computing-Grenzen erfordern, um Runtime-Escape-Risiken zu minimieren.
- **Grundlegende Modelle:** Die Anbieter von Rohdaten, bei denen die Gewichte gehostet werden, was sichere, validierte Modell-API-Keys erfordert.
- **Kanäle für Kommunikation und Zusammenarbeit:** E-Mail-, Chat- und Ticketing-Systeme, über die Agenten interagieren und die nicht vertrauenswürdige Prompt-Eingangspfade darstellen, sodass sie tiefgehenden Payload-Überprüfungen und -Scans unterliegen.

3.3.2.2 Zweckgebundene Datensicherheit

Zweckgebundene Datensicherheit bewertet, ob ein bestimmter Agent auf bestimmte Daten zugreifen, diese verarbeiten, transformieren, zusammenfassen, offenlegen oder übertragen sollte, basierend auf dem delegierten Zweck, der Berechtigung des bzw. der Benutzer:in, der Agentenidentität, dem Workflow-Status, der Datenklassifizierung, dem bzw. der Empfänger:in, dem Ziel und der nachgelagerten Aktion. Dabei werden Datenkontrollen zum Zeitpunkt der Nutzung und nicht nur zum Zeitpunkt des Zugriffs angewendet. Dies ist wichtig, da Agenten Daten systemübergreifend kombinieren, vertrauliche Informationen aus nicht vertraulichen Eingaben ableiten und Ergebnisse über Ausgaben, APIs, Nachrichten, Dateien und Unteragenten-Aufrufe offenlegen. Klassifizierung und DLP bestimmen, um welche Daten es sich handelt und ob sie sensibel sind; zweckgebundene Datensicherheit bestimmt, ob diese Nutzung im jeweiligen Kontext angemessen ist. Zu den wichtigsten Funktionen gehören die Zweckbindung zwischen der Benutzeranfrage, der Aufgabe des Agenten, der Datenquelle und der zulässigen Nutzung; die Laufzeitauswertung, ob Zugriff und Offenlegung zum aktuellen Workflow-Schritt passen; Kontrollen für Zusammenfassung, Transformation, Export, anwendungsübergreifende Übertragung sowie Offenlegung gegenüber Drittanbieter-Agenten; die Erkennung von Datenüberschreitungen, übermäßigem Abruf, unbefugten Rückschlüssen und unangemessener Weitergabe; sowie Durchsetzungsmaßnahmen, die Löschung, Maskierung, teilweise Antworten, Genehmigungs-Workflows, Protokollierung, Blockierung oder Eindämmung umfassen.

3.4 Säule 4:

Wie reagiere ich? (aktive Eindämmung)

Die letzte Säule des Konzepts wandelt Telemetriedaten in sofortige Eindämmung um. Wenn Überwachungssysteme ungewöhnliches Verhalten, kompromittierte Anmeldedaten oder einen Prompt-Injection-Angriff melden, muss das Unternehmen über hochgradig automatisierte, präzise Abwehrmechanismen verfügen, die verhältnismäßige, minimal störende Reaktionen (Ratenbegrenzung, Quarantäne) gegenüber destruktiven Maßnahmen bevorzugen, welche einen selbst verursachten Ausfall nach sich ziehen könnten.

3.4.1 Reaktion und Durchsetzung (Drosselung und Eindämmung)

Die automatisierte operative Kontrollschleife ist darauf ausgelegt, in Echtzeit auf Risikoindikatoren zu reagieren. Sie fungiert als koordinierte Kontrollebene (nicht als einzelner zentraler Engpass) und führt im gesamten technischen Stack gezielte, verhältnismäßige Sicherheitsreaktionen aus.

Die Durchsetzung ist am effektivsten, wenn sie mit einem standardisierten Benachrichtigungsprotokoll kombiniert wird: das System, das ein Risiko erkennt, veröffentlicht ein Not-Aus-Ereignis, die abonnierenden Systeme reagieren, und das Ergebnis wird als Bestätigungssignal zurückgemeldet.

3.4.1.1 Durchsetzungsmaßnahmen (Eindämmungsmechanismen)

Durchsetzungsmaßnahmen stellen die konkreten, mehrschichtigen Hebel dar, mit denen Security-Teams und automatisierte Orchestratoren einen problematischen oder kompromittierten Agenten isolieren, eindämmen oder vollständig neutralisieren können. Diese Maßnahmen bilden eine vom SOC verwaltete Bibliothek von Reaktionsoptionen, keine isolierten Einzelschalter: Sicherheitsverantwortliche wählen – gemessen an den aggregierten Risikoindikatoren – die Maßnahme mit den geringsten Auswirkungen.

Von Guardian-Agenten ausgelöste Eindämmungsmaßnahmen (Sperren, Isolieren, Beenden) setzen voraus, dass der Ziel-Agent innerhalb einer Ausführungsgrenze operiert, die granulare Kontrolle auf Session-Ebene unterstützt, unabhängig davon, ob die Ausführung in einem Managed Cloud Service, einer SaaS-Anwendung oder einer Container-Laufzeitumgebung erfolgt. Ohne zugrunde liegende Plattform- oder Laufzeitunterstützung, die einen Agenten ohne negative Auswirkungen auf angrenzende Workflows pausieren kann, vermittelt eine API-gesteuerte Eindämmung ein trügerisches Sicherheitsgefühl. Wenn die Eindämmung von Agenten mithilfe von physischer Technik oder OT erfolgt, müssen automatisierte Auslöser kalibriert werden, damit die abrupte Isolierung in der Anlage keine physischen Sicherheits- oder etriebsprobleme verursacht.

- **Token-Widerruf:** Sofortiges Ungültigmachen aktiver OAuth- oder API-Access-Token, wodurch die Fähigkeit des Agenten, verbundene Anwendungen aufzurufen, umgehend blockiert wird.
- **Session-Token-Widerruf:** Ungültigmachung aktiver OAuth-Token und Widerruf nachgelagerter Session-Kontexte über verbundene Unternehmensdienste hinweg und/oder Trennung der aktiven Laufzeitverbindung von einem einzelnen, spezifischen Interaktionsstrang, ohne notwendigerweise die allgemeine Identität des Agenten zu löschen.
- **Kontinuierliche Autorisierung:** Dynamische Echtzeit-Neubewertung von Sicherheitskontexten, die Berechtigungen sofort herabstuft, sobald sich die Risikolage ändert.

- **Prozessbeendigung:** Erzwungenes Beenden von Computing-Laufzeitumgebungen als letztes Mittel, wenn die Quarantäne fehlschlägt. Beachten Sie, dass durch die Beendigung forensische In-Memory-Telemetriedaten verloren gehen. Nutzen Sie daher nach Möglichkeit Netzwerkquarantäne, die Sperrung der Session oder die Stilllegung der Ausführungsumgebung.
- **Netzwerkquarantäne:** Verwendung von Mikrosegmentierungsregeln zur Isolierung der Hosting-Umgebung des Agenten, wodurch dessen Kommunikation sowohl mit dem internen Netzwerk als auch mit dem offenen Internet unterbrochen wird.
- **Beendigung von Cloud-Diensten:** Sofortige Stilllegung der zugrunde liegenden Cloud-Instanzen, serverlosen Funktionen oder Virtual-Machine-Cluster, die den schädlichen Agenten hosten.
- **Beendigung der Agenten-Plattform:** Ein administrativer Befehl auf oberster Ebene, der die zentralisierte Agenten-Orchestrierungsplattform anweist, den operativen Account des Agenten zu sperren und alle geplanten Aufgaben anzuhalten.
- **Drosselung und Ratenbegrenzung:** Schrittweises Einschränken der Anfragerate, des Token-Budgets oder der Parallelität eines Agenten als angemessene erste Reaktion, wodurch die Leistungsfähigkeit verringert wird, ohne die Session vollständig zu trennen.

3.4.2 Zugriffswiederherstellung

Ein Not-Aus-Schalter ist nur die Hälfte der Kontrollschleife: Jede Eindämmungsmaßnahme erfordert einen entsprechenden Weg zurück zu einem verifizierten, letzten bekannten fehlerfreien Zustand. Die Zugriffswiederherstellung regelt, wie ein eingedämmter, unter Quarantäne gestellter oder beendeter Agent wieder in Betrieb genommen wird. Dabei wird sichergestellt, dass die Wiedereinsetzung bewusst, dokumentiert und nachvollziehbar erfolgt und keine automatische Aufhebung der Durchsetzungsmaßnahme darstellt. Vollständig automatisierte, standardisierte Wiederherstellungs-Workflows bleiben ein Zielzustand. Die meisten Unternehmen benötigen bei einem oder mehreren der unten stehenden Schritte heute weiterhin eine manuelle Freigabe.

- **Erneute Attestierung:** Erneute Überprüfung von Identität, Code und Konfiguration des Agenten anhand einer als sicher bekannten Baseline, bevor der Zugriff in einer neuen, isolierten Ausführungsinstanz wiederhergestellt wird.
- **Abgestufte erneute Registrierung:** Schrittweise Wiederherstellung des Zugriffs (z. B. Lese- vor Schreibzugriff, Sandbox vor Produktion), anstatt sofort alle Berechtigungen wiederherzustellen.
- **Bestätigung der Ursache:** Erfordert, dass die menschliche verantwortliche Person oder das Security-Team die Ursache und Behebung dokumentiert, bevor der Agent wieder aktiviert wird.
- **Abschluss des Audit-Trails:** Verknüpfung des ursprünglichen Ereignisses, das den Not-Aus-Schalter ausgelöst hat, der Behebungsmaßnahmen und der Entscheidung zur erneuten Registrierung zu einem geschlossenen, auditierbaren Vorfalldatensatz.

3.5

Übergreifende Basiskomponenten

Diese Architekturebenen existieren nicht isoliert. Sie bilden die Grundlage und erstrecken sich über alle vier Säulen. Dabei dienen sie als Bindeglied, das Transparenz, Identität, Überwachung und Reaktion zu einer einheitlichen Sicherheitsstruktur verbindet.

3.5.1 Ausführungskontext und Risikoindikatoren

Ausführungskontext und Risikoindikatoren fungieren als das analytische Bindeglied des Konzepts. Diese Orchestrings-Engine erfasst kontinuierlich Echtzeit-Laufzeitlemetrie, Bewertungen der Sicherheitslage und delegierende Benutzeridentitätssignale (z. B. Indikatoren für kompromittierte Accounts oder ein erhöhtes Principal-Risiko) und bewertet Bedrohungen dynamisch, indem sie individuell risikoarme Bedingungen zu toxischen Kombinationen korreliert. Die Engine veröffentlicht zudem positiven Ausführungskontext, eine strukturelle Bestätigung dessen, was bereits als gültig bekannt ist, und nicht nur Warnmeldungen darüber, was schiefgelaufen ist. Wenn ein Schwellenwert überschritten wird (z. B. eine plötzliche Prompt Injection in Kombination mit dem Versuch, Finanzdaten zu exfiltrieren), lösen Risikoindikatoren die entsprechenden Reaktions-/Durchsetzungsmechanismen direkt und programmgesteuert aus. Um Echtzeit-Interoperabilität und koordinierte Reaktionen über heterogene Anbieterumgebungen hinweg zu gewährleisten, kann dieses System offene Sicherheitsstandards wie das **Shared Signals Framework (SSF)** und das **Continuous Access Evaluation Profile (CAEP)** nutzen, um Zero-Trust-Statusänderungen asynchron zu veröffentlichen und zu verarbeiten.

3.5.2 Telemetrie, Protokollierung und Beobachtbarkeit von KI-Agenten

Dies ist die grundlegende Basisebene, die alle Säulen der Referenzarchitektur unterstützt. Sie dient als unveränderliche Datenebene, die Ausführungs-Metadaten im gesamten Enterprise-Stack erfasst, und ist durch eine obligatorische Inline-Löschung von Payloads abgesichert, damit keine personenbezogenen Daten, Secrets und Anmeldedaten in den Logging Lake gelangen. Ohne diesen strukturierten, hochpräzisen Logging Lake ist eine automatisierte Bedrohungserkennung unmöglich. Für Sicherheitsbewertungen fehlen die Daten und forensische Incident-Responder:innen tapen völlig im Dunkeln. Um Konsistenz zu gewährleisten und Anbietersilos in komplexen Multi-Agenten-Interaktionen zu reduzieren, kann diese Ebene offene Standards wie das **Open Cybersecurity Schema Framework (OCSF)** nutzen, um verschiedene Log-Streams in einer einzigen, einheitlichen Sicherheitstaxonomie zu normalisieren.

- **Manipulationssicherer Telemetrie-Ursprung:** Die Erfassung von Telemetriedaten an der Ausführungsgrenze, anstatt sich auf vom Agenten selbst gemeldete Protokolle zu verlassen, trägt zur Gewährleistung der Nichtabstreitbarkeit bei und verhindert, dass ein kompromittierter Agent sein Ausführungsprotokoll unterdrückt, ändert oder fälscht.

4

Operationalisierung mit globalen Systemintegratoren (GSIs)

Ein Architekturkonzept erfordert ein operatives Bereitstellungsmodell, um einen dauerhaften Unternehmensvorteil zu bieten. Die folgenden Abschnitte beschreiben, wie Security-Teams und Beratungspartner diese vier Säulen innerhalb bestehender Unternehmens-Workflows umsetzen. Unternehmen können die Architektur je nach betrieblichen Möglichkeiten und Anforderungen mithilfe von internen Teams, GSIs, Managed Service Providern oder anderen qualifizierten Partnern umsetzen. Eine Referenzarchitektur, die ohne ein Betriebsmodell veröffentlicht wird, wird innerhalb eines Jahres zu Lagerware. Die folgenden Abschnitte definieren, wie diese Architektur in einem Unternehmen tatsächlich aufgebaut, personell besetzt und aufrechterhalten wird.

4.1 Bereitstellungsberatung und Architekturunterstützung

GSIs setzt die vier Säulen in einen schrittweisen, organisationsspezifischen Rollout um, anstatt eine gleichzeitige Komplettbereitstellung durchzuführen. Dies umfasst:

- **Bewertung des Ist-Zustands:** Abgleich des bestehenden IAM-, Netzwerk- und Monitoring-Stacks eines Unternehmens mit der Referenzarchitektur der Blueprint Alliance, um echte Lücken sowie solche Funktionen zu identifizieren, die bereits unter einem anderen Namen existieren.

- **Design des Zielbetriebsmodells:** Definition klarer operativer Verantwortlichkeiten für alle Säulen, um Verantwortlichkeitslücken zwischen den Teams zu vermeiden. Dieses Modell weist typischerweise Gateways und Überwachung dem Bereich Security Engineering; Agenten-Directory, Zugriffsrichtlinien und Governance dem Bereich IAM; Reaktion/Durchsetzung dem SOC sowie Ausführungsgrenzen (Sandboxing, Isolationslaufzeiten, Lebenszyklus-Management) dem Bereich Compute und Platform Engineering zu.
- **Planung des phasenweisen Rollouts:** Schrittweise Bereitstellung, angepasst an die bestehenden Tool-Investitionen des Kundenunternehmens und den regulatorischen Zeitplan.
- **Integration der Anbieter-Interoperabilität:** Konfiguration der spezifischen KI/LLM-, MCP-, Security-, Agenten- und API-Gateways, die ein Unternehmen bereits lizenziert hat, sodass die Interoperation über die gemeinsame Telemetrie- und Risikoindikator-Ebene erfolgt, anstatt dass ein kompletter Austausch der bestehenden Infrastruktur erforderlich ist.

4.2 Standardisierte Risikobewertungsmatrizen

Nicht jeder Agent erfordert die gleiche Kontrolltiefe, und GSIs bieten kalibrierte Bewertungs-Frameworks, die verhindern, dass Unternehmen risikoarme Bereitstellungen überdimensionieren oder risikoreiche unzureichend schützen. Dies umfasst:

- **Risikoeinstufungsmodelle**, die Agenten nach operativem Umfang (schreibgeschützt bzw. transaktional), Datenklassifizierungen, Autonomiegrad (OBO bzw. vollautonom) und dem Wirkungsradius der Zielressourcen bewerten.
- **Branchenspezifisch kalibrierte Schwellenwerte**: Ein Finanzdienstleistungs-Agent, der Zahlungs-Workflows ausführt, und ein Fertigungs-Agent, der OT-Sensordaten ausliest, erfordern vollkommen unterschiedliche Risiko-Baselines. GSIs bringen die branchenübergreifende Musterbibliothek mit, mit denen diese Schwellenwerte leicht kalibriert werden können. So müssen Unternehmen nicht bei null anfangen.
- **Bewertung der geschäftlichen Auswirkungen für semantische Governance**: Festlegung der finanziellen, regulatorischen oder reputationsbezogenen Schwellenwerte, die eine obligatorische Human-in-the-Loop-Genehmigung auslösen, abgestimmt auf die tatsächliche Risikobereitschaft des Kundenunternehmens anstelle eines generischen Standardwerts.
- **An die Risikostufe angepasste Isolationstiefe**: Die Risikoeinstufung muss die Stärke der strukturellen Ausführungsgrenze (z. B. Sandboxing auf Kernel-Ebene, kurzlebige Laufzeit-Lebenszyklen) parallel zu den Review-Gates anpassen. Dadurch wird sichergestellt, dass ein Hochrisiko-Agent im Falle einer Kompromittierung nicht denselben physischen Wirkungsradius hat wie ein risikoarmes Tool.

4.3 Kontinuierliche Lebenszyklusverwaltung

Governance für Agenten ist keine einmalige Zertifizierung; GSIs steuern die fortlaufenden Lebenszyklusprozesse, die das Agenten-Directory und die Zugriffsrichtlinien bei Änderungen der Umgebung korrekt halten:

- **Automatisierte Zyklen für die Zugriffszertifizierung**: Abrufe des Status des Agenten-Directorys und der Zugriffsrichtlinien durch Managed Services, wobei Agenten markiert werden, deren menschliche verantwortliche Personen das Unternehmen verlassen haben oder deren Berechtigungen vom Least-Privilege-Prinzip abgewichen sind.
- **Joiner-Mover-Leaver-Parität (JML) für Agenten**: Anwendung derselben Disziplin bei der Deprovisionierung, die für das Offboarding von Personen verwendet wird, auf Änderungen der für den Agenten verantwortlichen Person, Upgrades von Modellversionen und die Stilllegung von Unteragenten, damit sich verwaiste Anmeldedaten nicht als Hauptursache für Identity-Lücken ansammeln.
- **Weiterleitung von Ausnahmen bei der Aufgabentrennung (SoD)**: Betrieb der Workflows, die Ausnahmen bei Berechtigungskonflikten (z. B. ein Agent, der eine Finanztransaktion sowohl erstellen als auch genehmigen kann) zur Lösung an die zuständigen Systemverantwortlichen weiterleiten, und zwar im gleichen Rhythmus wie die Identity Governance für menschliche Identitäten.
- **Drift-Management für Ausführungsumgebungen**: Planmäßiges Patchen, Aktualisieren von Abhängigkeiten und erneute Attestierung von Agenten-Laufzeitumgebungen und zugrunde liegenden Plattformkonfigurationen, um Schwachstellen-Drift in gehosteten, Cloud- und Container-Bereitstellungen zu beseitigen.

4.4 Vorschriften- und Compliance-Einhaltung

GSIs verbinden die Architektur mit den spezifischen regulatorischen Rahmenbedingungen, denen ein Unternehmen unterliegt, und übersetzen technische Kontrollen in Nachweise, die Vorschriften, Compliance-Vorgaben und Audit-Anforderungen erfüllen:

- **Zuordnung von Kontrollen zu regulatorischen Frameworks:** Abstimmung von Telemetrie, Audit-Trails und Ergebnissen von Zugriffsüberprüfungen auf branchenspezifische Anforderungen (z. B. Modellrisikomanagement bei Finanzdienstleistungen, Datenverarbeitung im Gesundheitswesen, Schutz kritischer Infrastrukturen), während sich diese parallel zu spezifischen Richtlinien für agentenbasierte KI weiterentwickeln.
- **Audit-bereite Nachweispakete:** Strukturierung der Ausgabe der Telemetrie- und Protokollierungsebenen, sodass Zugriffsüberprüfungen, Eindämmungs- und Drosselungsereignisse sowie Datensätze zur Zugriffswiederherstellung (erneute Bestätigung, Bestätigung der Fehlerursache, Abschluss des Audit-Trails) die Dokumentationsstandards von Gutachter:innen und Prüfer:innen und nicht nur interne SOC-Anforderungen erfüllen.
- **Beratung zu grenzüberschreitenden Daten und Datenresidenz:** Beratung dazu, wie Agentenimport-Kanäle (SaaS-Agenten, First-Party-Agenten, die Unternehmensgrenzen überschreiten) mit den spezifischen Anforderungen an Datenresidenz und grenzüberschreitende Übertragungen in jeder für das Unternehmen relevanten Rechtsordnung zusammenhängen.
- **Manipulationssichere Audit-Herkunft:** Strukturierung der Telemetrieerfassung, sodass Aufsichtsbehörden und Prüfer:innen unwiderlegbare Audit-Protokolle erhalten, die direkt an der Ausführungsgrenze generiert werden. Dies verringert die Abhängigkeit von einer sekundären Protokollrekonstruktion aus der umgebenden Netzwerkinfrastruktur.

4.5 Branchenspezifische Playbooks

Die Kontrollanforderungen variieren je nach Branche, weshalb GSIs Playbooks pflegen müssen, die auf unterschiedliche Betriebsumgebungen zugeschnitten sind

- **Finanzdienstleister:** Priorisieren strikte Ausführungsisolierung, hochsichere Transaktionsschwellenwerte und regulatorische Datenkontrollen.
- **Fertigung und Industrie:** Legen den Schwerpunkt auf die Erkennung von OT-, IoT- und Edge-Agenten auf Legacy-Hardware.
- **Einzelhandel und Handel:** Konzentrieren sich auf die Abwehr von B2C-Prompt-Injections, den Markenschutz und die Verhinderung des Abflusses personenbezogener Kundendaten.

Die Gründungsmitglieder aktualisieren diese Playbooks kontinuierlich, sobald gemeinsame Interoperabilitätsergebnisse veröffentlicht werden, um das Betriebsmodell kontinuierlich an neue Agentenfähigkeiten anzupassen.

Fazit: Gestaltung der Zukunft der KI-Sicherheit

Die Blueprint Alliance regelt das Verhalten von Agenten zur Produktionslaufzeit. Diese Architektur ist direkt in den Enterprise-Stack integriert, anstatt nachträglich als externe Ebene hinzugefügt zu werden. Daher bietet sie die notwendigen strukturellen Schutzmaßnahmen, mit denen agentenbasierte Automatisierung sicher und im großen Maßstab bereitgestellt werden kann.

Um echte herstellerübergreifende Interoperabilität zu gewährleisten, validieren die Gründungsmitglieder ihre Plattformen aktiv anhand zentraler offener Standards, darunter das Model Context Protocol (MCP) für die Tool-Interaktion, das Open Cybersecurity Schema Framework (OCSF) für einheitliche Protokollierung, Shared Signals and Events (SSF/CAEP) für den Risikodaten-Austausch in Echtzeit sowie HTTP Message Signatures (RFC 9421) für die Authentifizierung von Anfragen. Diese gemeinsame Telemetrie-Basis gewährleistet, dass ein in einer Laufzeitumgebung erkannter Bedrohungsindikator oder eine Grenzverletzung nativ erfasst und über die gesamte Kontrollebene hinweg darauf reagiert werden kann. Die Möglichkeiten unterscheiden sich je nach Anbieter und Implementierung. Sobald die Tests abgeschlossen sind, werden Referenzintegrationen und validierte Interoperabilitätsergebnisse veröffentlicht.

Durch die regelmäßige Veröffentlichung gemeinsamer Interoperabilitätsergebnisse und Referenzintegrationen schafft die Blueprint Alliance ein sich entwickelndes, herstellerübergreifendes Ökosystem, das mit dem Tempo der Innovation bei Agenten Schritt hält.

Konzept in der Praxis

Die fünf nachfolgenden Szenarien zeigen, wie die Säulen in Abschnitt 3 in der Praxis zusammenwirken. Jedes ist eine Zusammensetzung, d. h. sie basieren auf Mustern aus frühen Alliance-Interaktionen und nicht auf einzelnen, namentlich genannten Kundenunternehmen. Diese Darstellungen veranschaulichen das beabsichtigte Verhalten der Architektur und sind kein Beleg für heute implementierte Ergebnisse in der Größenordnung von Alliance-Mitgliedern. Betrachten Sie sie als Architektur in Aktion, nicht als Fallstudie einer einzelnen Bereitstellung.

Vom Schatten-Agenten zur registrierten Identität. Eine Finanzanalystin verbindet einen browserbasierten KI-Assistenten mit der unternehmensweiten SSO-Session, um Lieferantenverträge zusammenzufassen. Die Schatten-KI-Erkennung kennzeichnet den nicht erkannten ausgehenden Datenverkehr bei der ersten Verbindung. AI Agent Security Posture Management prüft das Tool auf bekannte Schwachstellen und leitet es an das Agenten-Directory weiter. Noch bevor es zum ersten Zugriff auf vertrauliche Daten kommt, ist der Assistent entweder als verwaltete Identität mit einem verifizierten Metadatenprofil und eingeschränktem Zugriff registriert oder auf Netzwerkebene blockiert. Was früher monatelang unsichtbar blieb, ist nun sichtbar, wird geprüft und entschieden, bevor es jemals mit sensiblen Daten in Kontakt kommt.

Beibehaltung der Delegierung in einem Multi-Agenten-Workflow. Eine Leiterin der Vertriebsabteilung weist einen Orchestrator-Agenten an, eine Reihe von CRM-Datensätzen zu aktualisieren und eine Verteilerliste zu benachrichtigen. Der Orchestrator erzeugt einen Unteragenten für den Schreibvorgang und einen weiteren zum Entwerfen der Benachrichtigung. Jeder Hop überträgt einen verschachtelten actor-Claim zurück zur ursprünglichen menschlichen Session, sodass die Engine für das Agenten-Directory und die Zugriffsrichtlinien bei jedem Schritt OBO-Einschränkungen durchsetzen kann. Wenn ein Unteragent versucht, auf ein System zuzugreifen, auf das der bzw. die ursprüngliche Benutzer:in nicht zugreifen konnte, wird dies durch zentrale IAM- und Richtliniengrenzen (einschließlich anwendungsübergreifender Zugriff) blockiert. Da jeder Unteragent in seinem eigenen isolierten Ausführungskontext ausgeführt wird, kann ein blockierter Angriff weder die Orchestrator-Session noch parallele Aufgaben beeinträchtigen, während der gesamte Workflow in einem einzigen Ausführungs-Trace vereint bleibt.

Bedarfsgerechte Anpassung des Zugriffs an das, was der Agent tatsächlich zu tun versucht. Für eine Support-Mitarbeiterin wird grobgranularer Lesezugriff auf eine Kundendatenplattform bereitgestellt. Wenn der Agent versucht, eine Rückerstattung vorzunehmen, erkennt die intentionsbasierte Richtlinien-Engine den Wechsel von einer Abfrage zu einer Transaktion und verlangt eine Human-in-the-Loop-Genehmigung, bevor sie eine feingranulare, kurzlebige Berechtigung gewährt, die auf diese einzelne Bestellung beschränkt ist. Durch die Ausführung der Rückerstattung innerhalb einer flüchtigen Laufzeitumgebung, die auf diese einzelne genehmigte Aufgabe beschränkt ist, werden die erweiterten Berechtigungen und die Laufzeitumgebung nach Abschluss der Aufgabe stillgelegt, wodurch Standing-Berechtigungen und persistente Angriffsflächen reduziert werden. Die erweiterte Berechtigung erlischt in dem Moment, in dem die Aufgabe abgeschlossen ist.

Von der Anomalie zur automatisierten Eindämmung. Die Laufzeit-überwachung, bei der es sich in der Regel eher um ein herkömmliches ML-Anomalieerkennungsmodell als um ein LLM handelt, erkennt, dass ein Agent eine ungewöhnlich hohe Anzahl von Aufrufen einer internen API durchführt, auf die er sonst selten zugreift. Ein Inline-Guardian-Agent markiert separat eine Prompt-Injection-Signatur in seiner letzten Eingabe. Anhand von Risikoindikatoren werden beide Ereignisse über das Shared Signals Framework korreliert. Der definierte Risikoschwellenwert wird überschritten, sodass automatisch ein Token-Widerruf, die Netzwerkquarantäne und die Aussetzung der aktiven Ausführungsumgebung des Agenten ausgelöst werden. Dabei wird der In-Memory-Status für eine forensische Überprüfung beibehalten, anstatt ihn direkt zu beenden. Bei alldem muss nicht darauf gewartet werden, dass ein Mensch die Warnung bemerkt. Das Security-Team erhält den vollständigen Trace als Closed-Loop-Bestätigungsdatensatz, wodurch die Zeit bis zur Eindämmung von manuellen mehrstündigen Untersuchungen zu einer automatisierten, praktisch sofortigen Reaktion verkürzt wird.

Verwandlung von Zugriffsüberprüfungen in einen kontinuierlichen Prozess. Während jedes kontinuierlichen Zugriffszertifizierungszyklus ruft der Managed Service eines globalen Systemintegrators den vollständigen Status des Agenten-Directorys und der Zugriffsrichtlinien über die Open-Telemetry-Ebene der Konzeptarchitektur ab, markiert Agenten, deren verantwortliche Person das Unternehmen verlassen hat, und leitet Ausnahmen bei der Funktionstrennung an die zuständigen Systemverantwortlichen weiter. Eine manuelle, tabellenbasierte Überprüfung wird zu einem kontinuierlichen, revisionssicheren Workflow.

Best Practices für den Einstieg

Sie benötigen nicht bereits am ersten Tag jede Funktion in diesem Konzept. Hier ist die Reihenfolge, in der die Bereitstellungen der Alliance-Mitglieder tatsächlich erfolgt sind.

Richten Sie zuerst eine einheitliche Protokollierung und Telemetrie ein.

Sie können die Durchsetzung von Richtlinien oder das Baseline-Verhalten nicht ohne einen einzigen, normalisierten Logging Lake debuggen. Betrachten Sie dies als Voraussetzung für jeden der folgenden Schritte.

Beginnen Sie dann mit der Erkennung, nicht mit der Richtlinie. Was Sie nicht sehen, können Sie nicht steuern. Richten Sie die Schatten-KI-Erkennung und ein erstes Agenten-Inventar ein, bevor Sie eine einzige Zugriffsrichtlinie erstellen. Die meisten Teams sind überrascht, wie viele Agenten bereits ausgeführt werden.

Erst registrieren, dann einschränken. Überführen Sie alle entdeckten Agenten in ein zentrales Agenten-Directory mit verifizierten Identitäten und verantwortlichen Personen, bevor Sie feingranulare oder absichtsbasierte Richtlinien anwenden. Eine Richtlinie für einen nicht registrierten Agenten bietet keine Sicherheit.

Implementieren Sie gestaffelte Autorisierung. Feingranulare (FGA) und beziehungsbasierte (ReBAC) Kontrollen müssen vorhanden sein, bevor die absichtsbasierte Autorisierung die Berechtigungen von Agenten korrekt einschränken kann.

Stellen Sie die Laufzeit bereit, bevor Sie die Reaktion automatisieren.

Implementieren Sie Monitoring, Tracing und DLP/Ausgabefilterung umfassend genug, um eine echte Verhaltens-Baseline zu etablieren, bevor Sie die automatisierte Durchsetzung mit Risikoindikatoren verknüpfen. Wenn Sie die Eindämmung bei „verrauschten“ oder unvollständigen Telemetriedaten automatisieren, werden falsch-positive Ergebnisse das Vertrauen in das gesamte System untergraben.

Unterziehen Sie den Not-Aus-Schalter einem Belastungstest, bevor Sie ihn brauchen. Überprüfen Sie den Widerruf von Token, die Beendigung von Sessions und die Netzwerkquarantäne in regelmäßigen Abständen an einem Nicht-Produktions-Agenten. Ein Not-Aus-Schalter, den Sie noch nie getestet haben, ist eine Hypothese, keine Sicherheitsmaßnahme.

Betrachten Sie Governance als kontinuierlichen Prozess, nicht als Projekt. Zugriffsprüfungen, Zugriffsanforderungen und Prüfungen der Funktionstrennung sollten im selben Rhythmus ablaufen wie die Identity Governance für menschliche Identitäten und nicht als einmalige KI-Agenten-Initiative. Ein globaler Systemintegrator kann dabei helfen, diesen Übergang vom Projekt zum Betriebsmodell zu vollziehen.

Diese Materialien dienen ausschließlich allgemeinen Informationszwecken und stellen keine Rechts-, Datenschutz-, Sicherheits-, Compliance- oder Unternehmensberatung dar. Der Inhalt gibt womöglich nicht den aktuellen Stand der Sicherheits-, Rechts- oder regulatorischen Entwicklungen wieder. Es liegt in Ihrer Verantwortung, sich mit Blick auf die Rechtslage, den Datenschutz, die Sicherheit, die Compliance und das Business beraten zu lassen. Stützen Sie sich nicht allein auf diese Materialien. Die Autor:innen übernehmen keinerlei Zusicherungen oder Garantien in Bezug auf diesen Inhalt und haften nicht für Verluste oder Schäden, die aus Ihrer Implementierung dieser Empfehlungen resultieren.