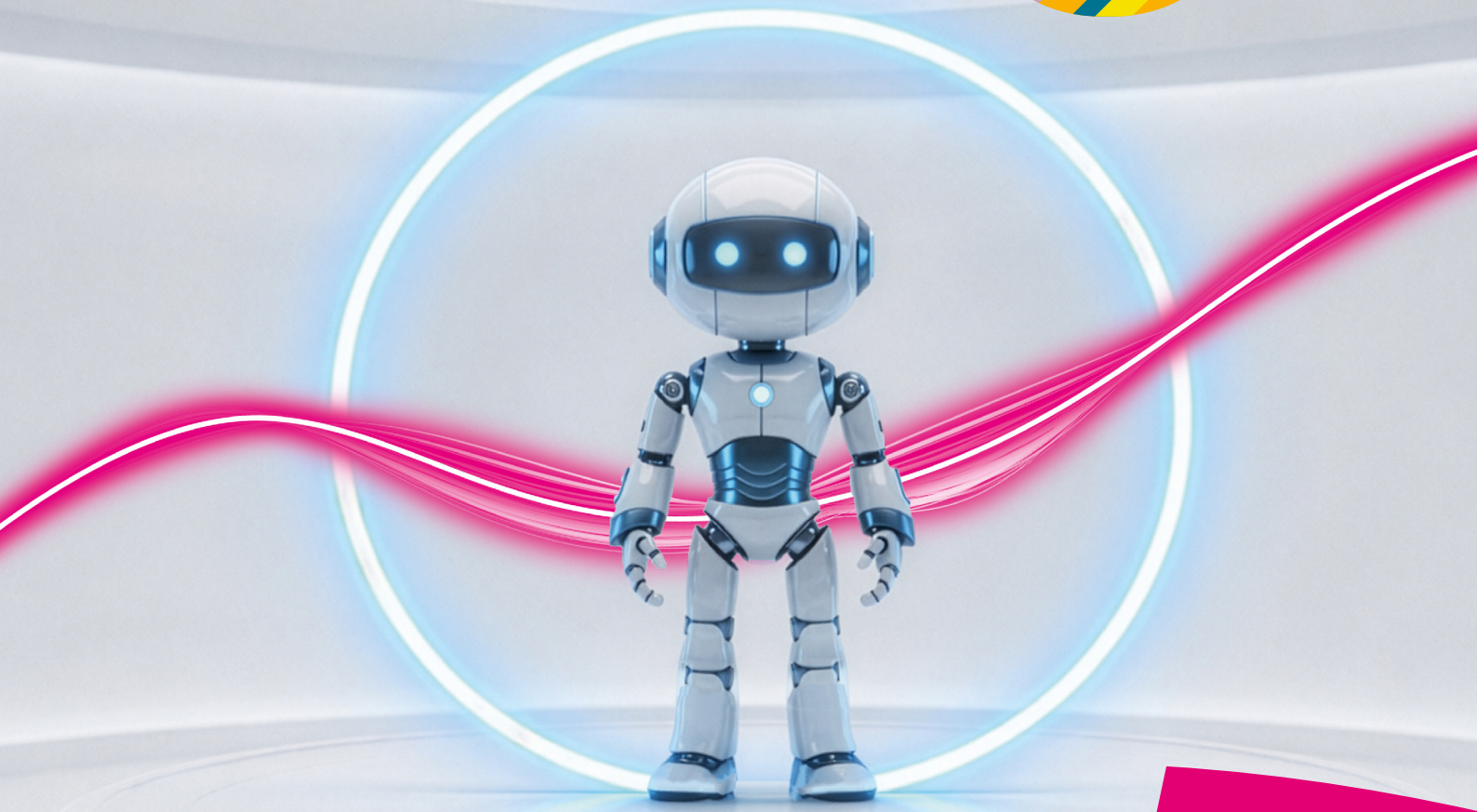
T SECURITY

In Kooperation mit PAC

Sicherer Rahmen für künstliche Intelligenz

Zero Trust Identitäts-management für KI-Agenten



INHALTSVERZEICHNIS

MANAGEMENT SUMMARY

KI-Agenten übernehmen zunehmend Aufgaben wie Entscheidungsfindung, Transaktionsabwicklung und Prozesssteuerung. Ihre Vernetzung in Multi-Agenten-Systemen birgt neuartige Sicherheitsrisiken: Ein kompromittierter Supervisor-Agent kann Kaskadenfehler auslösen, die gesamte Prozessketten lahmlegen. Identitätsmanagement ist hier keine technische Randaufgabe, sondern die strategische Basis für Sicherheit und Compliance.

Kurzlebige Agenteninstanzen mit Lebensdauern unter fünf Minuten stellen traditionelle Identity-Management-Systeme vor Herausforderungen. Drei Probleme sind besonders relevant: Sicherheitslücken durch verzögerte Widerrufe kompromittierter Identitäten, exponentielle Komplexität durch hierarchische Agentenstrukturen sowie Medienbrüche bei der Integration in CI/CD-Pipelines. Herkömmliche manuelle Prozesse können mit der Dynamik und Skalierung nicht Schritt halten.

Als Lösungsrahmen etabliert sich eine Zero-Trust-Architektur. Diese implementiert das Prinzip "Never Trust, Always Verify" durch Mikrosegmentierung von Netzwerken in isolierte Sicherheitszonen. Innerhalb dieser Zonen operieren Agenten mit strikt minimalen Zugriffsrechten. Hardware Security Modules härten das System ab, indem sie Schlüsselmaterial physisch schützen. Für die Verwaltung flüchtiger Identitäten hat sich das SPIFFE/SPIRE-Framework bewährt, das automatisierte Identitätslebenszyklen mit Just-in-Time-Ausstellung und Widerruf ermöglicht. Die Integration in Bereitstellungspipelines ist entscheidend, wobei Policy-as-Code-Validierung Konfigurationsfehler reduziert.

Die Kommunikation zwischen Agenten erfordert spezifische Protokolle. Mutual TLS mit Post-Quantum-Optionen dient als Fundament für gegenseitige Authentifizierung. Zur Absicherung von Agent-zu-Agent-Interaktionen kommen ressourcenoptimierte Protokolle wie A2A zum Einsatz, die Ende-zu-Ende-Verschlüsselung gewährleisten. Nonce-basierte Mechanismen verhindern Replay-Angriffe. Handoff-Prozesse zwischen Agenten werden durch eine dreistufige Sicherheitsarchitektur gesichert: kryptografische Verifizierung des Zielagenten, sessionspezifische Verschlüsselung und Integritätssicherung via Hashketten.

Für die Autorisierung transformieren intelligente Modelle statische Zugriffskontrollen. Policy Engines wie der Open Policy Agent implementieren kontextbewusste Entscheidungen, die Faktoren wie Systemlast oder Risikoscores integrieren. Adaptive Sicherheitsschleifen koppeln Runtime-Monitoring mit automatisierten Rechteanpassungen. Jede Entscheidung generiert unveränderliche Audit Trails mit vier Kernelementen: Identitätsnachweis des Agenten, angefragte Ressource/Aktion, angewendete Policy-Regeln und Umgebungszustand. Diese erfüllen Compliance-Anforderungen des EU AI Act oder der DORA-Verordnung.

Für die Praxis ergeben sich drei Handlungsfelder: Erstens die Klärung organisatorischer Zuständigkeiten für KI-Governance. Zweitens die konsequente Umsetzung von Zero-Trust-Architekturen mit Mikrosegmentierung und Minimalrechten. Drittens die vollständige Automatisierung dynamischer Identitätslebenszyklen durch Integration von Frameworks wie SPIFFE/SPIRE in CI/CD-Pipelines. Wie das Whitepaper betont, ist die Frage, ob Agenten zum Vorteil oder zur Schwachstelle werden, eine Führungsentscheidung. Klare Richtlinien und definierte Verantwortlichkeiten schaffen den Rahmen, in dem Teams Automatisierung vorantreiben können, ohne die Kontrolle zu verlieren.

VORWORT

KI-Agenten sind keine Zukunftsmusik mehr, sondern ein relevantes und aktuelles Thema. Sie sind zuständig für die Recherche, Entscheidungsfindung, Bestellung, Buchung und Genehmigung von Projekten – oft schneller als ganze Teams. Genau darin liegt ihre Stärke, aber auch das Risiko. Sobald Agenten eigenständig mit Systemen und Daten interagieren, ergeben sich Managementfragen. Es geht um die Zuständigkeit, die Grenzen der Interaktion und die Nachvollziehbarkeit von Entscheidungen für Mitarbeiter, Vorgesetzte, den Vorstand und Kunden.

Das vorliegende Whitepaper basiert auf einer grundlegenden Beobachtung: Eine Vielzahl von Unternehmen startet mit großem Enthusiasmus Pilotprojekte mit Agenten, jedoch kommt es bei der Umsetzung in den regulären Betrieb häufig zu Schwierigkeiten. Dies ist nicht auf das Fehlen von Modellen zurückzuführen, sondern darauf, dass Identität, Berechtigungen, Nachweise und Abläufe nicht im gleichen Tempo mit der technologischen Entwicklung gewachsen sind. Ohne diese Grundlagen besteht jedoch die Gefahr von Datenlecks, Fehlentscheidungen, Projektstopps aufgrund von Prüfungsfeststellungen sowie unnötigen Kosten. Diese Grundlagen befähigen Agenten, dauerhaft zur Steigerung von Produktivität und Qualität beizutragen.

Vertrauen durch Identität

Die wesentlichen Punkte lassen sich auf drei Kernthemen zurückführen. Zunächst ist das Vertrauen in automatisierte Entscheidungen von entscheidender Bedeutung. Jede Aktion eines Agenten muss einer identifizierbaren Einheit zugeordnet werden können, rechtlich begründbar und technisch nachweisbar sein.

Sicherheit als integraler Bestandteil

Zweitens ist eine beherrschbare Geschwindigkeit erforderlich. Skalierung darf nicht dahingehend missverstanden werden, dass zunächst gebaut und anschließend gesichert wird. Sicherheit ist ein integraler Bestandteil des Agenten-Lieferprozesses und muss konsequent gewährleistet werden. Dies umfasst die Implementierung testbarer Richtlinien, die Attestierung als Eintrittstor sowie die lückenlose Nachvollziehbarkeit der Spuren. Regulatorische Klarheit ist von entscheidender Bedeutung. Compliance ist kein nachträglicher Zusatz, sondern ein integraler Bestandteil der Lösung. Die Anforderungen aus Gesetzen und Normen werden in klare Kontrollen, belastbare Nachweise und Kennzahlen umgesetzt, damit Audits planbar und beherrschbar sind.

Für Entscheidungsträger ergibt sich daraus eine klare Prioritätenfolge. Die Identität der Agenten steht dabei an erster Stelle. Statt gemeinsamer Konten oder dauerhafter Rechte sollte jedem Agenten eine eindeutige, zeitlich begrenzte Identität zugewiesen werden. Auf dieser Grundlage sollten Rechte nur im Bedarfsfall vergeben werden. Least Privilege und Just-in-Time werden zum Standard, ergänzt um klar definierte Notfallpfade auf Basis des Vier-Augen-Prinzips. Ebenso wichtig sind überprüfbare Entscheidungen. Durch die Signierung von Entscheidungsketten und die Verwendung eines WORM-Speichers, der Fälschungen verhindert, wird eine lückenlose Nachvollziehbarkeit sowohl intern als auch im Rahmen von Audits sichergestellt. Die Attestierung etabliert sich somit als zentrale Form der Zugriffskontrolle: Vor dem Erhalt von Schlüsseln oder Daten erfolgt eine kryptografische Verifizierung der Laufzeitumgebung.

Integrierte Ausfallsicherheit

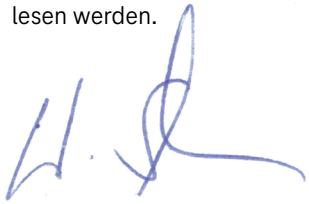
Schließlich ist integrierte Ausfallsicherheit essenziell. Widerrufe von Rechten und Schlüsseln müssen innerhalb von Minuten wirksam werden. Es ist ein definierter Degradationsmodus erforderlich anstelle unkontrollierter Freigaben. Zudem sind klare Ziele für die Wiederherstellungszeit nach einem Ausfall sowie für den maximal tolerierten Datenverlust, festzulegen.

Der Business Case lässt sich wie folgt zusammenfassen: Sichere Agenten beschleunigen Prozesse, reduzieren Fehlerquoten und schaffen Vertrauen, weil jede Aktion von einer identifizierten, autorisierten und attestierten Instanz stammt, deren Spur nachweisbar bleibt.

Für die kommenden zwölf Monate könnte ein realistisches Zielbild wie folgt aussehen: Mehr als 99 Prozent aller Agentensitzungen sind beglaubigt und vollständig protokolliert. Die Sperrung kompromittierter Zugriffe erfolgt innerhalb von Minuten in allen relevanten Zonen. Die Richtlinien decken mindestens 95 Prozent der kritischen Abläufe ab und werden wie Software getestet, mit Canary als kontrollierter Teilausführung und Dry Run als Testlauf, bei dem Entscheidungen bereits sichtbar, aber noch nicht durchgesetzt werden. Die Audits zur Nutzung von KI und zur IT-Sicherheit ergaben keine wesentlichen Beanstandungen, da die erforderlichen Nachweise strukturiert und in standardisierter Form vorliegen.

Führung ist von entscheidender Bedeutung. Ob Agenten zu einem Vorteil oder zu einer Schwachstelle werden, ist keine rein technische Frage, sondern eine Führungsentscheidung. Die Festlegung klarer Richtlinien und die Definition von Verantwortlichkeiten, Kennzahlen und Nachweisen schaffen einen Rahmen, in dem Teams die Automatisierung von Prozessen vorantreiben können, ohne dabei die Kontrolle zu verlieren. Dieses Vorwort dient somit weniger als Einführung, sondern vielmehr als Einladung: Stellen Sie sicher, dass Sicherheit, Governance und Nachweisbarkeit zu den Gestaltungsprinzipien werden. Agenten tragen nicht nur zur Effizienz, sondern auch zur Reputation, zur Resilienz und zur Einhaltung gesetzlicher Vorschriften bei.

Ich hoffe, dass Sie dieses Whitepaper mit großem Interesse lesen werden.



Wolfgang Schwab, Head of Cybersecurity bei PAC

1. EINLEITUNG

In der dynamischen Welt autonomer KI-Systeme wird effektives Identitätsmanagement zur entscheidenden Stellschraube für Sicherheit und Vertrauen. KI-Agenten übernehmen zunehmend komplexe Entscheidungen und führen eigenständig Handlungen aus. Ihre Vernetzung in Multi-Agenten-Systemen birgt jedoch neuartige Sicherheitsrisiken: Bereits ein einziger kompromittierter Supervisor-Agent kann hier Kaskadenfehler auslösen, die ganze Prozessketten lahmlegen und erheblichen Schaden verursachen.

Moderne Identitätslösungen transformieren dieses fundamentale Risiko in eine strategische Chance für Ihr Unternehmen. Sie ermöglichen nicht nur maschinenlesbare, überprüfbare Identitäten für flüchtige Agenteninstanzen, sondern gewährleisten auch eine automatisierte Durchsetzung von Sicherheitsrichtlinien und schaffen die unverzichtbare Grundlage für Compliance.

Dieses Whitepaper entschlüsselt konkrete Strategien, wie Sie skalierbare Identitätsframeworks etablieren, um auch kurzlebige Agenteninstanzen sicher zu verwalten. Es zeigt auf, wie Sie Rechteeskalationen innerhalb hierarchischer KI-Systemarchitekturen proaktiv verhindern können. Zudem erfahren Sie, wie Sie die Kommunikationssicherheit zwischen Agenten mit ressourcenoptimierten Protokollen gewährleisten – entscheidend für den effizienten Betrieb.

Die Dringlichkeit dieser Maßnahmen zeigt sich in branchenspezifischen Bedrohungsszenarien: Im Finanzsektor können beispielsweise gespoofte Agenten unautorisierte Handelsaktionen auslösen. Das Gesundheitswesen sieht sich Risiken wie manipulierten Diagnoseempfehlungen durch Prompt-Injection-Angriffe ausgesetzt. In Industrie 4.0-Umgebungen drohen kostspielige Produktionsausfälle durch die genannten Kaskadenfehler in vernetzten KI-Systemen.

Die in diesem Whitepaper vorgestellten Sicherheitskonzepte und Architekturen sind keine theoretischen Modelle. Sie erfahren, wie Identitätsmanagement vom operativen Hindernis zum strategischen Enabler wird, der autonome Systeme nicht nur sicher, sondern erst wirklich vertrauenswürdig macht.

Tauchen Sie ein und erfahren Sie, wie Sie die Sicherheit Ihrer KI-Agenten-Ökosysteme auf das nächste Level heben.

1.1 Warum Agentensicherheit jetzt strategische Priorität braucht

Die Sicherheitsrisiken, die mit der Vernetzung von KI-Agenten einhergehen, erfordern eine strategische Priorisierung der Agentensicherheit. Unter KI-Agenten sind autonome Systeme, die Aufgaben basierend auf künstlicher Intelligenz ausführen und oft nur kurz aktiv sind, zu verstehen. Diese Agenten agieren in einer Art und Weise, die die Grenzen zwischen menschlich gesteuerten Aktionen und Automatisierung verschwinden lassen. Die zunehmende Integration von KI-Agenten in kritische Infrastrukturen und Geschäftsprozesse macht es unerlässlich, proaktive Sicherheitsmaßnahmen zu implementieren. Hierbei geht es besonders um die Vermeidung operativer Störungen, finanzieller Verluste sowie Reputationsschäden.

Risiko: Prompt-Injection-Manipulationen

Ein zentraler Aspekt ist die Anfälligkeit von KI-Agenten für Prompt-Injection-Manipulationen. Diese Angriffe können Sicherheitsbeschränkungen umgehen und zu unerwünschten Handlungen führen.

Durch die Implementierung von Eingabevalidierung, semantische Firewalls und Ausgabesanitisierung, die gemeinsam Angriffe in Echtzeit erkennen und neutralisieren können, kann diese Bedrohung effektiv abgewehrt werden.

Risiko: Identitätsfälschung

Ein weiterer kritischer Faktor ist die Identitätsfälschung, auch Spoofing genannt. Gefälschte Agenten können unautorisierte Aktionen in automatisierten Workflows ausführen, insbesondere bei der Integration von Drittsystemen. Hier wird der Bedarf an robusten Authentifizierungs- und Autorisierungsmechanismen deutlich.

Risiko: Systemische Risiken

Die größte Gefahr jedoch bilden systemische Risiken. Ein fehlgesteuerter Agent kann in Netzwerken unkontrollierbare Kettenreaktionen auslösen. Diese systemischen Risiken erfordern zum einen eine umfassende Risikobewertung und die Entwicklung von Notfallplänen und zum anderen vorbeugend Netzwerksegmentierung und eine granulare Berechtigungssteuerung.

Die strategische Priorisierung der Agentensicherheit erfordert eine ganzheitliche Herangehensweise. Unternehmen müssen nicht nur technische Sicherheitsmaßnahmen implementieren, wie sie in diesem Whitepaper betrachtet werden, sondern auch die organisatorischen Rahmenbedingungen schaffen sowie Schulungen für Mitarbeiter durchführen und regelmäßige Sicherheitsaudits durchführen. Um eine Grundlagen für eine sichere und zuverlässige Nutzung von KI-Agenten zu schaffen.

Ob Agenten zu einem Wettbewerbsvorteil oder zu einer Schwachstelle werden, ist keine rein technische Frage, sondern eine Führungsentscheidung. Die Festlegung klarer Richtlinien und die Definition von Verantwortlichkeiten, Kennzahlen und Nachweisen schaffen einen Rahmen, in dem Teams die Automatisierung von Prozessen vorantreiben können, ohne dabei die Kontrolle zu verlieren.

Aufsichtsbehörden, interne Revision und Datenschutz richten dabei ihren Blick im Kern auf drei Fragen: Wer entscheidet über die Befugnisse eines Agenten? Ist jede Aktion nachvollziehbar und belegbar? Und erfüllt das Zusammenspiel von Technik und Organisation die Vorgaben aus Gesetzen, Verordnungen und Normen? Viele Anforderungen aus dem EU AI Act, DORA, NIS2, der DSGVO und ISO-Normen sind bewusst abstrakt formuliert. Für Technikteams bleibt dadurch oft unklar, welche Architektur- und Kontrollmaßnahmen konkret erforderlich sind. Für das Management ist es schwer zu erkennen, ob aktuelle Projekte tatsächlich auditfest sind.

Die Deutsche Telekom Security bietet Lösungen und Dienstleistungen, die Unternehmen bei der Entwicklung und Implementierung einer robusten Agentensicherheitsstrategie unterstützen. Dabei greifen wir auf unser jahrelanges Know-How zurück.

1.2 Zero-Trust Identity Management als Lösungsansatz

Eine Zero-Trust-Architektur (ZTA) etabliert sich als grundlegendes Sicherheitsparadigma für KI-Agenten-Systeme, indem sie das Prinzip "Never Trust, Always Verify" konsequent umsetzt. Dieser Ansatz stellt einen grundlegenden Wandel dar: Statt sich auf traditionelle Perimeter-Absicherung zu verlassen, wird jede Zugriffsanfrage – unabhängig von ihrer Herkunft – als potenziell gefährlich behandelt. Für autonome Agentensysteme ist dies besonders relevant, da ihre dynamischen Interaktionen und dezentralen Entscheidungsprozesse statische Sicherheitsgrenzen obsolet machen.

Kern des Zero-Trust-Modells ist die Mikrosegmentierung, die Netzwerke in isolierte Sicherheitszonen unterteilt. Innerhalb dieser Zonen operieren Agenten mit strikt minimalen Zugriffsrechten (Least Privilege), was Rechteeskalationen und unerlaubte Werkzeugnutzung (Tool Misuse) effektiv eindämmt. Diese strukturelle Absicherung wird durch kontinuierliche Verhaltensüberwachung ergänzt, bei der Algorithmen Abweichungen von etablierten Betriebsmustern erkennen – etwa ungewöhnliche Prompt-Interaktionen oder anomale Ressourcennutzung, die auf Manipulationsversuche hindeuten können.

Die operative Umsetzung erfolgt durch Policy-Engines wie Open Policy Agent (OPA), die attribut-/ bzw. rollenbasierte Zugriffskontrolle (ABAC/RBAC) in Echtzeit umsetzen. Dabei fließen dynamische Kontextfaktoren wie Agenten-Vertrauensscore, Systemrisikostatus oder Handlungskritikalität in die Entscheidungsfindung ein. Ein Beispiel: Ein Agent mit niedrigem Vertrauensscore erhält automatisch reduzierte Zugriffe, bis eine manuelle Überprüfung durch einen Administrator erfolgt ist.

Für Compliance-Anforderungen, wie sie zum Beispiel der EU AI Act stellt, werden unveränderliche Audit Trails zur unverzichtbaren Komponente. Diese kryptografisch signierten Protokolle dokumentieren jede Agentenaktion mit Identitätsnachweis und schaffen so die geforderte Transparenz über automatisierter Entscheidungen. Hardware Security Modules (HSM) härten das System zusätzlich ab, indem sie Schlüsselmaterial in physisch geschützten Chips speichern und so vor Software-basierten Angriffen schützen.

In der Praxis integrieren Unternehmen diese Komponenten in Security-Orchestration-Plattformen (SOAR), die automatisierte Incident-Response ermöglichen: Bei Anomalieerkennung werden kompromittierte Agenten automatisch isoliert (Sandboxing) oder deren Berechtigungen widerrufen. Dieser mehrschichtige Ansatz adressiert die spezifischen Schwachstellen agentenbasierter Systeme, indem er Sicherheit als durchgängigen Prozess – nicht als statischen Perimeter – implementiert.

2. HERAUSFORDERUNGEN VON AGENTEN IDENTITY MANAGEMENT

Die Verwaltung von Identitäten für kurzlebige KI-Agenten stellt traditionelle Identity-Management-Systeme vor große Herausforderungen. Diese kurzlebigen Agenteninstanzen mit Lebensdauern von weniger als fünf Minuten erfordern neue Ansätze für die Identitätserstellung, -verteilung und den -widerruf. Herkömmliche manuelle Prozesse können der Geschwindigkeit und Skalierung dieser dynamischen Lebenszyklen oft nicht gerecht werden.

Drei Herausforderungen sind besonders relevant:

Sicherheitslücken können durch verzögerte Widerrufe kompromittierter Identitäten entstehen. Oft liegen Stunden oder sogar Tage zwischen der Kompromittierung und der Deaktivierung einer Identität, was Angreifern ein großes Zeitfenster für Missbrauch bieten kann.

Durch hierarchische Agentenstrukturen entsteht exponentielle Komplexität. Wenn ein Supervisor-Agent Dutzende Subagenten steuert, müssen Identitäten und Berechtigungen dynamisch delegiert werden. Wenn Tausende von Agenten pro Stunde gestartet werden, führt dies in der Praxis zu nicht mehr manuell beherschbaren Aufwänden.

Oft fehlt die nahtlose Integration in moderne CI/CD-Bereitstellungspipelines (Continuous Integration/Delivery), was zu Medienbrüchen führen kann. Wenn IAM-Lösungen keine passende Schnittstellen bieten oder hohe Latenzen aufweisen, entstehen Blockaden. Wenn Tausende Agenten gleichzeitig Identitäten anfordern, werden Zertifizierungsstellen, Hardware-Security-Module und Policy-Engines zu Flaschenhalsen. Praktisch bedeutet das, dass KI-Agenten auf ihre Identitäten warten, während teure Infrastrukturressourcen ungenutzt bleiben.

Lösung: Das SPIFFE/SPIRE-Framework

Als Lösungsrahmen etabliert sich das SPIFFE/SPIRE-Framework. Dieses Framework ermöglicht die automatisierte Identitätsausstellung mit integriertem Widerruf. Es generiert kurzlebige, workload-spezifische Identitäten, die ohne manuelle Interaktion ausgestellt und nach Aufgabenende automatisch entzogen werden können.

Dabei greift das Framework auf bekannte Technologien zurück: X.509-Zertifikate und JSON Web Token (JWT). Ein X.509-Zertifikat bestätigt die Identität eines Agenten und kann zur Authentifizierung verwendet werden. Ein JWT enthält Informationen über die Identität und Berechtigungen eines Agenten und kann zur Autorisierung verwendet werden. Beide Methoden können in Kombination oder separat eingesetzt werden, um die Sicherheit und Zuverlässigkeit der Identitätsverwaltung zu gewährleisten.

Durch die kontinuierliche Weiterentwicklung und Anpassung der Sicherheitsstrategien können Unternehmen diese Herausforderungen bewältigen und die Sicherheit ihrer KI-Agenten gewährleisten.

Entscheidend ist der mentale Shift: Agenten-Identitäten sollten als temporäre Ressourcen behandelt werden und nicht als dauerhafte Benutzerkonten.

2.1 Exkurs: Sicherheit der Agent2Agent-Kommunikation

Die Kommunikation zwischen KI-Agenten ist ein kritischer Punkt in verteilten Systemen. Wenn diese Kommunikation nicht sicher ist, kann dies zu einer Vielzahl von Problemen führen, wie z. B. Datenlecks oder unberechtigten Zugriffen.

Ein wichtiger Aspekt ist die Verwendung von eindeutigen Zahlen, die nur einmal verwendet werden. Diese Zahlen, auch als Nonces bezeichnet, verhindern, dass abgefangene Nachrichten später wiedereingespeist werden (Replay-Angriffe), indem sie jeder Nachricht eine eindeutige, einmalig gültige Kennung zuweisen.

Die Verschlüsselung der Kommunikation ist ein weiterer wichtiger Punkt. Hierbei kommt eine spezielle Verschlüsselungstechnik (AES-GCM-SIV) zum Einsatz, die sowohl Vertraulichkeit als auch Integrität der übertragenen Daten gewährleistet. Diese Technik ist besonders sicher, da sie – im Gegensatz zu klassischem AES-GCM selbst bei Wiederverwendung von Initialisierungsvektoren sicher bleibt.

Neben den etablierten Sicherheitsmaßnahmen spielen spezifische Kommunikationsprotokolle eine zentrale Rolle. Zwei relevante Ansätze sind das Agent-to-Agent-Protokoll (A2A) und das Model-Context-Protokoll (MCP), die jeweils unterschiedliche Sicherheitsanforderungen adressieren.

Model-Context-Protokoll (MCP)

MCP überträgt nicht nur Daten, sondern auch den Kontext eines KI-Modells – etwa Trainingsdatenherkunft oder Entscheidungslogik. Es nutzt verschlüsselte Kontextcontainer, die nur von berechtigten Agenten geöffnet und modifiziert werden können. Integrität wird durch Hash-basierte Prüfsummen gesichert, die Manipulationen sofort erkennbar machen. Es ermöglicht nachvollziehbare Entscheidungspfade und beugt falschen Schlussfolgerungen vor, indem der Modellkontext transparent bleibt. Gleichzeitig erhöht die Übertragung von Kontextmetadaten die Angriffsfläche. Über einen ungesicherten Kontextcontainer könnten etwa bösartige Code-Snippets eingeschleust werden. Inkompatible MCP-Versionen zwischen Agenten können zu Interpretationsfehlern oder abgebrochenen Kommunikationsketten führen.

MCP stellt allein keinen vollständigen Sicherheitslayer bereit, da es auf externe Mechanismen angewiesen ist. Seine Container erfordern daher zwingend Kombination mit Protokollen wie A2A oder TLS und entsprechendes Schlüsselmanagement, um Vertraulichkeit zu gewährleisten. Ohne diese ist MCP lediglich ein strukturiertes, aber ungeschütztes Datenformat.

Beide Protokolle ergänzen die bestehenden Maßnahmen wie AES-GCM-SIV-Verschlüsselung und dreistufige Agent-Handoffs. Während A2A die Basis für vertrauliche 1:1-Kommunikation legt, adressiert MCP komplexe Anforderungen an Transparenz und Nachvollziehbarkeit.

Die größte verbleibende Herausforderung liegt in der **protokollübergreifenden Sicherheitskonsistenz,** insbesondere wenn Nachrichten zwischen unterschiedlichen Systemwelten transferiert werden. Adaptive Sicherheitsgateways lösen dies durch Echtzeit-Protokolltranslation und synchronisieren die Sicherheitssysteme. Diese Gateways überbrücken insbesondere Konfigurationslücken zwischen A2A- und MCP-Implementierungen. Durch die kontinuierliche Weiterentwicklung und Anpassung der Sicherheitsstrategien können Unternehmen diese Herausforderungen bewältigen und die Sicherheit ihrer KI-Agenten gewährleisten.

Durch diese Protokolle entsteht ein mehrschichtiges Sicherheitsnetz – allerdings nur, wenn ihre Implementierung konsistent über alle Systemgrenzen hinweg erfolgt. Kontinuierliche Protokollaudits und automatisiertes Patch-Management sind hier unverzichtbar.

Agent-to-Agent-Protokoll (A2A)

Das Agent-to-Agent-Protokoll operiert wie ein abhörsicherer Direktkanal zwischen zwei Parteien. Es etabliert pro Sitzung einen temporären Verschlüsselungstunnel, der durch gegenseitige Authentifizierung mittels digitaler Zertifikate initialisiert wird. Jede Nachricht enthält neben den verschlüsselten Nutzdaten einen kryptografischen Fingerabdruck (HMAC), der Manipulationen während der Übertragung erkennbar macht. Ein entscheidender Sicherheitsvorteil ist die Sitzungsisolation: Selbst wenn ein Angreifer einen Schlüssel kompromittiert, bleiben andere Kommunikationsstränge unberührt. Allerdings zeigt A2A Schwächen in Szenarien mit häufigen Verbindungswechseln, da der fortlaufende Neuaufbau von Sitzungen Rechenlast verursacht. Zudem bietet es keine integrierte Schutzfunktion gegen Denial-of-Service-Angriffe, da die Authentifizierung vor Nachrichtenverarbeitung erfolgt. Die strikte Punkt-zu-Punkt-Architektur limitiert auch die Skalierbarkeit bei Gruppenkommunikation.

Besondere Aufmerksamkeit erfordern **Agent-Handoffs,** also strukturierte Aufgabenübergaben zwischen Agenten, bei denen Kontextinformationen wie Bearbeitungsstatus oder Zwischenergebnisse transferiert werden müssen. Spezifizierte Kontexttransferprotokolle sichern diese kritischen Übergabepunkte durch eine dreistufige Sicherheitsarchitektur: Zunächst erfolgt eine kryptografische Verifizierung des Zielagenten, um dessen Identität und Berechtigung zu bestätigen. Anschließend wird die Nutzlast mit session-spezifischen Schlüsseln verschlüsselt, die ausschließlich für diese Übergabe generiert werden. Abschließend erfolgt eine Integritätsprüfung via kryptografischer Hash-Verkettung (Hash Chains), die jede Modifikation der Daten während des Transfers erkennbar macht.

Für ressourcenbeschränkte Umgebungen wie IoT-Edge-Netzwerke gibt es speziell optimierte Sicherheitslösungen. Diese Lösungen reduzieren den Protokoll-Overhead und ermöglichen so interoperable Sicherheit in heterogenen Agentenplattformen. Insbesondere die Schlüsselverwaltung für MCP wird hier durch vereinheitlichte Policy-Templates optimiert. Dies reduziert manuelle Konfiguration und stellt konsistente Verschlüsselungsregeln für Kontextcontainer in heterogenen Umgebungen sicher.

Durch die Kombination aus kryptografischer Absicherung und protokolltechnischen Kontrollen entsteht eine durchgängige Sicherheitskette von der Nachrichtengenerierung bis zur finalen Verarbeitung.

3. KERNKOMPONENTEN UND PRINZIPIEN VON AGENT IDENTITY MANAGEMENT

3.1 Identitätsverwaltungssysteme für KI-Agenten

Die Gestaltung von Identitätsverwaltungssystemen für KI-Agenten erfordert eine strategische Abwägung zwischen zentralisierten und dezentralen Ansätzen. Zentralisierte Systeme eignen sich für homogene Umgebungen, wo einheitliche Richtlinien und einfache Verwaltungsprozesse Priorität haben. Dezentrale Ansätze hingegen reduzieren das Risiko von Single Points of Failure und sind besonders wertvoll in föderierten oder hierarchischen Multi-Agenten-Systemen.

Drei Prinzipien bilden den Rahmen für die Implementierung solcher Systeme: Eindeutigkeit, Widerrufbarkeit und Schutz kryptografischer Schlüssel. Die Eindeutigkeit gewährleistet, dass jeder KI-Agent über eine unverwechselbare Identität verfügt. Dies kann zum Beispiel durch die Kombination von zeitgestempelten Identifikatoren mit hardwarebasierten Fingerabdrücken erreicht werden. Die Widerrufbarkeit stellt sicher, dass kompromittierte Identitäten umgehend entzogen werden können. Dies kann durch spezielle Mechanismen wie OCSP-Stapling unterstützt werden. Der Schutz kryptografischer Schlüssel erfolgt durch Hardware Security Module, die physische Sicherheitsbarrieren gegen Extraktion bieten und automatisierte Schlüsselrotation durchführen.

Für die technische Darstellung von Identitäten stehen zwei Optionen zur Verfügung: X.509-Zertifikate und dezentrale Identifikatoren (DIDs). X.509-Zertifikate sind ein bewährter Standard für PKI-basierte Unternehmensumgebungen, während DIDs für selbstverwaltete Identitäten in Szenarien geeignet sind, in denen über Grenzen hinweg interagiert werden soll. Es ist wichtig, dass diese Systeme post-quanten-resilient sind, um langfristige Sicherheit zu gewährleisten.

Moderne Systeme erfordern auch kontextbewusste Identitätsbindung, bei der Agentenrechte dynamisch an operative Faktoren wie geografische Position oder Systemlast angepasst werden. Beispielsweise kann ein Agent im Produktionsnetzwerk automatisch eingeschränktere Rechte erhalten als in isolierten Testumgebungen.

Für kurzlebige Agenten in Serverless-Umgebungen kommen Just-in-Time-Identitäten zum Einsatz, bei denen kurzlebige Credentials bei Agentenstart automatisch ausgestellt und nach Aufgabenende widerrufen werden. Dieser Ansatz kombiniert dezentrale Flexibilität mit zentraler Auditierbarkeit und reduziert Angriffsflächen durch Gültigkeitsfenster von typischerweise unter fünf Minuten.

Die nahtlose Integration in Multi-Agenten-Architekturen wird durch standardisierte Handoff-Protokolle gewährleistet, die den Identitätskontext bei Aufgabenübergaben durch eine dreistufige Sicherheitsarchitektur erhalten. Diese Technik verkürzt Übergabelatenzen für Echtzeitsteuerungsketten in industriellen IoT-Umgebungen.

Durch die Kombination dieser Prinzipien und Technologien können Unternehmen eine sichere und effiziente Identitätsverwaltung für ihre KI-Agenten implementieren.

3.2 Authentifizierungsprotokolle

Die Absicherung der Agent-zu-Agent-Interaktion erfordert spezialisierte Protokollarchitekturen, die sich mit Ende-zu-Ende-Verschlüsselung für durchgängige Vertraulichkeit, delegierte Autorisierung für präzise Zugriffskontrolle und ressourcenoptimierten Sicherheitsmechanismen für hardwarebeschränkte Umgebungen beschäftigen. Dieser mehrschichtige Ansatz bildet die Grundlage vertrauenswürdiger Kooperation in verteilten KI-Systemen.

Als Fundament etabliert sich Mutual TLS (mTLS) mit Post-Quantum-Kryptografie-Optionen, das eine gegenseitige Authentifizierung durchführt. Jeder Agent weist dabei seine Identität kryptografisch nach und verifiziert gleichzeitig die seines Kommunikationspartners, was Man-in-the-Middle-Angriffe wirksam unterbindet. Für dynamische Umgebungen löst SPIFFE/SPIRE die Identitätsverwaltung durch standardisierte Workload-APIs, die SPIFFE-IDs als X.509-SVIDs oder JWT-SVIDs ausstellen und so serviceübergreifende Vertrauensstellungen ohne manuelle PKI-Interaktion ermöglichen.

Die kontrollierte Zugriffserteilung realisiert OAuth 2.0 mit mTLS-bound Access Tokens (RFC 8705), wo ein Autorisierungsserver berechtigungsgebundene Token ausstellt, die physisch an kryptografische Schlüssel gekoppelt sind. Dieser Mechanismus verhindert effektiv den Missbrauch gestohlener Credentials und übertrifft ältere DPoP-Ansätze in Sachen Sicherheit.

Ergänzend implementieren Service-Mesh-Architekturen automatisiertes Policy-Enforcement durch Sidecar-Proxys. Diese Proxys verschlüsseln nicht nur den Datenverkehr, sondern ermöglichen auch Laufzeit-Attestation über das RATS Framework (Remote Attestation Procedures). Dieses standardisierte Protokoll definiert Verfahren zur kryptografischen Verifikation der Systemintegrität. Dabei erfasst der Sidecar-Proxy maschinenlesbare Beweise über den Zustand der Ausführungsumgebung, wie Kernel-Version oder geladene Bibliotheken. Ein vertrauenswürdiger Verifier prüft diese Evidenz gegen referenzierte Integritätsmesswerte und validiert die Signaturkette bis zu einer Hardware-Root-of-Trust, beispielsweise einem TPM oder HSM. Das resultierende Attestation Result bestätigt oder widerlegt die Unversehrtheit der Umgebung und erkennt damit manipulierten Agentencode in Echtzeit. Token-Sicherheitsmechanismen nutzen moderne TLS Exporter Keys gemäß RFC 5705 zur Sitzungskopplung, während ephemere Credentials mit Gültigkeitsfenstern unter fünf Minuten das Angriffsfenster für Replay-Attacken minimieren

Hardware-basierte Roots-of-Trust in TPM 2.0/HSM schützen kryptografisches Material durch physische Isolierung, wobei Service-Mesh-Integrationen wie Istio Authorization Policies die konsistente Durchsetzung von Zugriffsregeln über Systemgrenzen hinweg garantieren. Für Delegationsszenarien kommen W3C Verifiable Credentials zum Einsatz, die Berechtigungen kryptografisch nachweisbar zwischen Agenten übertragen.

In der Praxis kristallisieren sich zwei Muster heraus: Cloud-native Agenten nutzen typischerweise mTLS-Varianten mit SPIFFE-Identitäten und OAuth-Token-Delegation, während Edge-Agenten auf IETF ACE mit EDHOC-Handshakes setzen. Kritische Operationen erfordern hardwaregestützte Schlüsselgenerierung mit TPM-Bindung und RATS-Attestation. Dieser adaptive Ansatz gewährleistet, dass Sicherheitsanforderungen die Agilität autonomer Systeme nicht beeinträchtigen – eine essentielle Voraussetzung für Echtzeitinteraktionen in industriellen Steuerungstopologien oder finanziellen Transaktionsnetzwerken.

So entsteht eine zukunftssichere Architektur, die sowohl aktuellen Bedrohungen als auch kommenden regulatorischen Anforderungen gerecht wird. Die Kombination aus automatisiertem Identity-Lifecycle-Management, hardwaregestützter Sicherheit und kontextbewusster Autorisierung transformiert dabei Sicherheit vom Hindernis zum Ermöglicher skalierbarer Autonomie.

3.3 Intelligente Autorisierungsmodelle

Moderne Autorisierungssysteme für KI-Agenten transformieren statische Zugriffskontrollen in dynamische Entscheidungsprozesse, die Kontext, Risiko und Absicht synthetisieren. Attributbasierte Zugriffskontrolle (ABAC) bildet hier das Fundament, indem sie Umgebungsvariablen, Agenteneigenschaften und Handlungskonsequenzen in Echtzeit analysiert – ein Paradigmenwechsel, der traditionelle Rollenmodelle (RBAC) durch kontextsensitive Bewertungen ersetzt. Diese evaluieren nicht nur "Wer darf was", sondern integrieren fluktuierende Faktoren wie Systemlast, geografische Position oder dynamische Risikoscores, wodurch autonome Systeme flexibel auf variable Bedrohungslagen reagieren.

Die operative Umsetzung erfolgt durch Policy Engines wie dem Open Policy Agent (OPA), in welchem die Autorisierungslogik in deklarativen REGO-Richtlinien implementiert wird. Diese regelbasierten Policies werden als Code versioniert, automatisiert getestet und konsistent durchgesetzt. Für latenzkritische Szenarien werden ABAC-Entscheidungen vorberechnet und in Edge-Caches vorgehalten, während Hardware-Integrationen wie Keylime oder OpenTitan TPM-Attestationen direkt in die Policy-Logik einbinden, um die physische Integrität von Ausführungsumgebungen zu verifizieren.

Adaptive Sicherheitsschleifen koppeln Runtime-Monitoring mit dynamischem Enforcement: LSTM-basierte Anomaliedetektoren identifizieren Abweichungen wie unerwartete Tool-Invokationen oder Ressourcenzugriffe in unter drei Millisekunden und lösen automatische Rechteanpassungen aus. Dieser risikoadaptive Ansatz reduziert Angriffsflächen proaktiv, ohne menschliche Intervention zu benötigen. Bei Hochrisikooperationen – etwa Finanztransaktionen oder Steuerung kritischer Infrastrukturen – aktiviert Step-up-Authentication zusätzliche Sicherheitsebenen: Menschliche Freigaben via FIDO2 oder QR-Code-basierte MFA wirken als letzte Verteidigungslinie gegen autonome Eskalationen. Für KI-Agenten kommt Model-Attestation hinzu: Vor Entscheidungsausführung verifiziert eine Trusted Computing Base (TCB) die Hash-Summe des Agentenmodells gegen zertifizierte Referenzen. Abweichungen, wie sie etwa Adversarial-Poisoning erzeugt, lösen automatische Isolation aus.

Delegationssicherheit wird durch OAuth Token Exchange (RFC 8693) gewährleistet, das Berechtigungskaskaden zwischen Agenten mit kryptografischen Nachweisen absichert. Jede Entscheidung generiert zudem unveränderliche Audit Trails mit vier Kernelementen: Einem kryptografischen Hash des ausführenden Agenten, die angefragte Ressource und Aktion, die angewandten Policy-Regeln sowie den Umgebungskontext zum Entscheidungszeitpunkt. Diese Struktur ermöglicht nicht nur forensische Analysen, sondern erfüllt auch Compliance-Anforderungen.

In der Praxis realisieren solche Architekturen damit eine neue Sicherheitsparadoxie: Je autonomer Agenten handeln, desto dynamischer muss ihre Kontrolle sein. Durch die Integration von Hardware-Vertrauensankern, kryptografischer Delegation und selbstoptimierenden Modellen entsteht so ein Schutzrahmen, der nicht restriktiv wirkt, sondern erst die volle Handlungsfähigkeit ermöglicht – besonders in Szenarien wie vernetzter Produktion oder adaptiver Logistik, wo Millisekunden über Erfolg oder Systemversagen entscheiden.

4. TECHNOLOGIELANDSCHAFT ZUR ABSICHERUNG VON KI-AGENTEN

In diesem Kapitel werden bewusst keine KI-Modelle, Agenten-Frameworks oder Formen der Prompt-Orchestrierung behandelt. Im Fokus stehen ausschließlich die Sicherheits- und Governance-Bausteine, die für den Betrieb von KI-Agenten erforderlich sind. Die eigentliche KI-Engine lässt sich vergleichsweise schnell austauschen, das Sicherheitsfundament hingegen ist eine langfristige Investition: Es muss Fragen zu Identität, Rechten, Protokollierung, Kryptografie und Resilienz zuverlässig beantworten.

Die Technologielandschaft zur Absicherung von KI-Agenten lässt sich in wenige wiederkehrende Kategorien gliedern. Jede Kategorie erfüllt einen klaren Sicherheitszweck: Wer oder was handelt hier, wer entscheidet über erlaubte Aktionen, wie werden Aktivitäten protokolliert und nachweisbar gemacht, wie werden Daten und Modelle geschützt und wie bleibt die eingesetzte Kryptografie langfristig tragfähig.

Kategorien der Technologielandschaft zur Absicherung

Identity und Trust

Identity und Trust beziehen sich auf die technische Identität und die Vertrauenswürdigkeit von Agenten und Diensten. Workload-Identität wird über Zertifikate oder vergleichbare Mechanismen bereitgestellt, idealerweise mithilfe kurzlebiger Tokens, klarer Delegationsmodelle und sauberer Governance für Service-Konten. Remote Attestation ergänzt die reine Identität um den Nachweis des Systemzustands, etwa durch Messungen über Trusted Platform Modules oder Enklaven sowie attestierte Nachweise, die in Richtlinien einfließen. Identity-Provider und Föderation sorgen dafür, dass Agenten in bestehende Identitätslandschaften eingebettet werden, etwa über offene Protokolle (OpenID Connect, SAML), Schnittstellen für die automatisierte Verwaltung (SCIM) und fein granulare Berechtigungen.

Policy und Authorization

In dieser Kategorie wird die Entscheidungslogik für Zugriffe und Aktionen verankert. Eine Richtlinien-Engine entscheidet, ob eine Aktion erlaubt ist; sie gehört zum Sicherheitskonzept, nicht zur KI-Logik. Wichtig sind die Unterstützung rollen- und attributbasierter Modelle, die Trennung von Entscheidungs- und Durchsetzungspunkten sowie die Verwaltung von Policies als Code mit Versionierung, Tests, Dry-Run und schrittweisem Rollout. Ergänzend stellen Secrets-Management, Key-Management und Hardware-Sicherheitsmodule sicher, dass Agenten nur mit kontrolliertem Schlüsselmaterial arbeiten, etwa durch Umschlagverschlüsselung, regelmäßige Schlüsselrotation und Modelle wie Bring-Your-Own-Key oder Post-Quantum-fähige Verfahren.

Connectivity und Runtime

Connectivity und Runtime sorgen dafür, dass Agenten sicher mit Systemen kommunizieren und in einer kontrollierten Umgebung laufen. Service Mesh und API-Gateways bilden dabei einen Steuerungs- und Schutzlayer rund um Agenten und Dienste, mit durchgängig verschlüsselten Verbindungen, Zugriffskontrollen auf Anwendungsebene, Weitergabe von Identität und Kontext sowie Rate-Limits. Die Laufzeitumgebung wird durch Sicherheitsleitplanken, Prozessisolation, zentrale Token-Vergabe und Quotenmechanismen so gestaltet, dass Agenten agieren können, ohne die Umgebung zu gefährden.

Daten und Modellschutz

Daten- und Modellschutz konzentrieren sich auf die Absicherung der zentralen Assets im Agenten-Umfeld. Auf Datenebene geht es um die Klassifikation von Daten, die Steuerung und Überwachung von Datenflüssen, den Schutz vor Datenverlust sowie Verfahren wie Pseudonymisierung oder Anonymisierung. Modelle werden als schützenswerte Ressourcen verstanden: Zweck, Datenquellen und Grenzen werden dokumentiert, Zugriffe auf Modelle und Konfigurationen gesteuert, Ausgaben kontrolliert und regelmäßige Sicherheitstests (z. B. Red-Team-Übungen) verankert.

Observability und Risk

Observability und Risk adressieren Sichtbarkeit, Nachvollziehbarkeit und Nachweismöglichkeiten. Benötigt werden Telemetrie- und Sicherheitsinformationssysteme, die verteilte Traces, konsistente Identitäten sowie Integrationsmöglichkeiten mit SIEM-Plattformen bereitstellen. Ergänzend sorgen evidenzorientierte Speicher, Schutz vor Manipulation und aussagekräftige Berichte dafür, dass die Wirksamkeit von Kontrollen überprüfbar ist und die Audit-Anforderungen erfüllt werden.

Kryptografie und Zukunftsfähigkeit

Schließlich geht es um die langfristige Tragfähigkeit der Kryptografie. Dazu gehören ein Inventar der eingesetzten kryptografischen Verfahren, erste Schritte hin zu hybriden Schlüsselaustauschverfahren sowie ein Migrationspfad für besonders schützenswerte Verbindungen und Daten. Ergänzend werden Schlüssel-Escrow und Notfallpfade klar geregelt, mit definierten Schwellen, Vier-Augen-Kontrolle und vollständiger Protokollierung, um im Ausnahmefall handlungsfähig zu bleiben.

FAZIT: JETZT DEN SICHEREN RAHMEN SETZEN!

Dieses Whitepaper verdeutlicht: Identitätsmanagement ist keine technische Randaufgabe, sondern die strategische Basis für den sicheren und skalierbaren Einsatz von KI-Agenten. Eine klare und verifizierbare Identität bildet die Grundlage für den Zugriff auf Informationen sowie die ordnungsgemäße Ausführung von Aktionen.

Learnings

1. Autonomie erfordert Kontrolle:
Je eigenständiger Agenten entscheiden und handeln, desto entscheidender wird eine lückenlose Identitätssicherung – besonders bei Aufgabenübergaben in hierarchischen Systemen.

2. Dynamik verlangt Automatisierung:
Kurzlebige Agenteninstanzen machen manuelle Prozesse obsolet; nur vollautomatisierte Lebenszyklen gewährleisten Sicherheit bei Massenskalierung.

3. Heterogenität braucht Standardisierung:
Unterschiedliche Plattformen erfordern interoperable Protokolle, um agentenübergreifende Kooperation ohne Sicherheitskompromisse zu ermöglichen.

Unsere Handlungsempfehlungen für die Praxis

- Klären Sie organisatorische Zuständigkeiten um die Regulatorik für KI-Agenten steuerbar und auditfest zu machen.
- Setzen Sie Zero-Trust-Architekturen konsequent um, indem Sie Agenten durch Mikrosegmentierung in isolierten Sicherheitszonen mit strikten Minimalrechten operieren lassen. Ergänzen Sie dies durch Hardware-basierte Roots-of-Trust wie TPM/HSM-Module, die kryptografische Schlüssel physisch absichern.
- Automatisieren Sie dynamische Identitätslebenszyklen vollständig, indem Sie Frameworks wie SPIFFE/SPIRE in Ihre CI/CD-Pipelines integrieren. Entscheidend ist hier die Policy-as-Code-Validierung, die Sicherheitsrichtlinien algorithmisch vor jedem Deployment prüft und so Konfigurationsfehler reduziert.
- Sichern Sie Agentenkommunikation durch mehrschichtige Schutzmechanismen ab. Verhindern Sie Replay-Angriffe mit Nonce-basierten Einmalkennungen und geeigneter Verschlüsselung. Besondere Priorität sollte der Absicherung von Handoff-Prozessen zukommen.

Sie wünschen Unterstützung?

Unsere Expert*innen stehen Ihnen jederzeit zur Verfügung.

Kontakt
security.dialog@telekom.de
security.telekom.de

Herausgeber
Deutsche Telekom Security GmbH
Consulting
Bonner Talweg 100
53113 Bonn

Deutsche Telekom Security GmbH | 03/2026