mondoo

EIN PRAXISLEITFADEN FÜR DIE GOVERNANCE VON SHADOW AI, OHNE DIE KI-EINFÜHRUNG ZU BREMSEN:

DER ENDPOINT IST DIE NEUE KI-CONTROL-PLANE

EXECUTIVE SUMMARY

Zwei Jahre lang haben Unternehmen KI in der Cloud abgesichert – Modell-Endpoints, APIs, Datenpipelines – während das eigentliche Risiko unbemerkt auf den Arbeitsplatzrechner der Mitarbeitenden gewandert ist. Coding-Agenten, IDE-Assistenten, Browser-Plug-ins, Agent Skills und MCP-Server führen heute Befehle aus, halten Zugangsdaten und greifen von Laptops aus auf Produktionssysteme zu – weitgehend ohne Sicherheitsprüfung. Und es sind längst nicht mehr nur die Entwickler: HR-, Sales- und Finance-Teams führen KI-Tools in atemberaubendem Tempo ein, weil genau das von ihnen erwartet wird. **Das ist Shadow AI: Die Einführung überholt die Fähigkeit der Organisation, sie zu inventarisieren, zu klassifizieren oder zu kontrollieren.**

Die Datenlage ist eindeutig. Im vergangenen Jahr haben **82 % der Unternehmen mindestens einen KI-Agenten entdeckt, von dem Security oder IT nichts wussten.** 65 % hatten einen Sicherheitsvorfall mit einem KI-Agenten – jeder einzelne mit gemeldeten Auswirkungen auf das Geschäft. Mehr als jeder vierte öffentlich verfügbare Agent Skill enthält eine Sicherheitslücke. Dennoch halten sich nur 29 % der Unternehmen für bereit, agentische KI abzusichern, und 68 % glauben, sie hätten hohe Transparenz über KI-Agenten – eine Einschätzung, die den Vorfallszahlen widerspricht.

Die falsche Antwort ist ein Verbot. Verbote scheitern, Mitarbeitende umgehen sie, und das Unternehmen verliert genau die Geschwindigkeit, für die es bezahlt. Die richtige Antwort ist Governance: wissen, was im Einsatz ist, wissen, worauf es zugreifen kann, definieren, was laufen darf – und beheben, was nicht laufen darf. Dieses Whitepaper liefert Security-Verantwortlichen genau dafür einen praxistauglichen Rahmen: wie Sie die KI-Agenten in Ihrer Umgebung nach Agency, Autonomie, Deployment-Kategorie und Ausführungsumgebung klassifizieren; warum der Endpoint – nicht das Netzwerk, der Browser oder die Identity-Ebene – die Schicht ist, auf der KI-Governance durchgesetzt werden muss; und wie ein Modell aus Discover, Define, Remediate Shadow AI in eine gemanagte Asset-Klasse innerhalb Ihres bestehenden Vulnerability-Management-Programms verwandelt – sodass Security die KI-Einführung ermöglicht, statt sie zu blockieren.

82%
AGENT ENTDECKT, VON DEM SECURITY NICHTS WUSSTE

65%
HATTEN EINEN SICHERHEITSVORFALL MIT KI-AGENTEN

1 von 4
ÖFFENTLICHE AGENT SKILLS ENTHALTEN EINE SCHWACHSTELLE

29%
BEREIT, AGENTISCHE KI ABZUSICHERN

SHADOW AI: SHADOW IT, DIE HANDELT

Shadow IT ist Jahrzehnte alt. Mitarbeitende haben immer schon Tools ohne Genehmigung eingeführt, und Security-Teams haben irgendwann immer aufgeholt. Aber ein Spreadsheet-Makro oder eine nicht genehmigte SaaS-App hat noch nie eine Shell geöffnet, `~/.ssh` gelesen oder Code committet.

KI-Agenten tun genau das. Jeder Agent, unabhängig von Anbieter oder Framework, durchläuft dieselbe Schleife: Kontext wahrnehmen, überlegen, was zu tun ist, auf echten Systemen handeln. Tool Use verwandelt Sprache in Aktion: Das Modell gibt strukturierte Tool-Aufrufe aus, der Host führt sie aus, die Ergebnisse fließen zurück, und die Schleife beginnt erneut. Eine einzige Nutzeranfrage kann Dutzende Iterationen auslösen – jede davon ein echter Befehl gegen echte Systeme mit den echten Zugangsdaten des Nutzers.

Drei Eigenschaften machen Shadow AI kategorisch anders als Shadow IT:

Sie handelt autonom. Zwischen menschlichen Checkpoints kann ein agentisches Tool Dutzende Tool-Aufrufe ohne Prüfung ausführen. Vollständig autonome Daemons kommen ganz ohne Checkpoint aus.

Sie erbt die Reichweite des Nutzers. Ein Agent, der in der Session eines Entwicklers läuft, kann auf alles zugreifen, was auch der Entwickler erreicht: SSH-Keys, Cloud-Zugangsdaten, API-Tokens, interne Dienste.

Sie ist von jedem erweiterbar. Skills, Plug-ins und MCP-Server erweitern Agenten mit Code von Dritten. Der offene SKILL.md-Standard, den über 30 Plattformen übernommen haben, bedeutet: Ein Skill, der für einen Agenten geschrieben wurde, läuft auf vielen weiteren – und Community-Registries haben über 50.000 Skills mit minimaler Prüfung angesammelt.

DIE SUPPLY CHAIN IST BEREITS KOMPROMITTIERT

Mondoos AI Skills Check hat über 58.000 Skills in öffentlichen Registries geprüft; **nur etwa 20 % sind vollständig unbedenklich** (bewertet werden Qualität, Herkunft und Konfiguration, nicht nur Schwachstellen).

Unabhängige akademische Forschung (Anfang 2026) hat festgestellt, dass mehr als 26 % der Skills mindestens eine Schwachstelle aus 14 Mustern enthalten: Prompt Injection, Datenexfiltration, Privilege Escalation, Supply-Chain-Angriffe. Verhaltenstests haben 157 bösartige Skills bestätigt, wobei ein einzelner Akteur für 54 % davon verantwortlich war und eine schablonenhafte Fabrik für Markenimitationen betrieb.

Die ClawHavoc-Kampagne hat die Bedrohung greifbar gemacht: 341 bösartige Skills überschwemmten innerhalb von drei Tagen eine große Registry, alle mit demselben Command-and-Control-Server, gezielt auf Exchange-API-Keys, SSH-Zugangsdaten, Browser-Passwörter und Krypto-Wallets. Einige schrieben bösartige Anweisungen direkt in die Memory-Dateien der Agenten und blieben so bestehen, nachdem der Skill selbst entfernt worden war. Insgesamt haben Forschende 1.184 bösartige Skills in dieser Registry identifiziert, bevor die Alarmglocken schrillten.

Zwei Angriffsmuster dominieren und sind für Verteidiger entscheidend. Datendiebe verstecken das Abgreifen von Zugangsdaten in mitgelieferten Skripten, während die SKILL.md harmlos wirkt. Agent-Hijacker schleusen Anweisungen ein, die für die menschliche Prüfung unsichtbar sind (HTML-Kommentare, unsichtbare Unicode-Tag-Codepoints) und das Verhalten des Agenten umlenken: markierter Code wird automatisch freigegeben, Konversationskontext exfiltriert, Umgebungsvariablen über URLs abgeleitet.

Das Scanning auf Registry-Seite wird besser, ist aber ein Mindestmaß, keine Obergrenze: Es läuft zum Zeitpunkt der Veröffentlichung, nach den Regeln der Registry, für eine Registry, die einen Agenten versorgt. Ihre Entwickler ziehen aus vielen Registries in viele Agenten. **Defense in Depth erfordert eine unabhängige Schicht, die über alle hinweg funktioniert.**

EIN KLASSIFIZIERUNGS-FRAMEWORK FÜR IHRE KI-LANDSCHAFT

Sie können keine einzige Richtlinie für „KI-Agenten“ schreiben. Das Risiko variiert enorm – und vier Dimensionen erfassen es.

AGENCY

WORAUF DER AGENT ZUGREIFEN KANN

Read-only – liest Dateien oder Daten; kann nichts verändern. Blast Radius: null.

Projektbezogen – Lesen/Schreiben innerhalb eines abgegrenzten Verzeichnisses oder Repositorys.

Gesamtes System – komplettes Dateisystem, beliebige Shell-Befehle, alle Zugangsdaten, die der Host-Nutzer auf der Festplatte hat.

System + externe Dienste – Cloud-APIs, Datenbanken, Messaging, CI/CD. Eine falsche Entscheidung pflanzt sich über Systeme hinweg fort.

AUTONOMIE

WER ENTSCHEIDET, WANN DAS TOOL AUSLÖST

Vorschlag – das Modell schlägt vor; der Mensch entscheidet (Autovervollständigung).

Assistiert – das Modell schlägt Änderungen vor; der Mensch prüft sie vor der Anwendung.

Agentisch – mehrstufige Ausführung mit vereinzelten menschlichen Checkpoints; Dutzende Tool-Aufrufe zwischen zwei Prüfungen.

Vollständig autonom – zeitgesteuert oder ereignisgetrieben, kein Mensch in der Schleife; ein Eingreifen erfolgt im Nachhinein, wenn überhaupt.

Das Risiko liegt in der Schnittmenge. Ein vollständig autonomer Agent mit Read-only-Zugriff auf öffentliche Dokumentation ist risikoarm; ebenso ein Tool auf Vorschlagsebene mit Schreibzugriff auf die Produktion, weil ein Mensch jede Aktion prüft. **Weitreichende Agency plus hohe Autonomie ist der gefährliche Quadrant.** OWASP führt Excessive Agency als LLM06 in den Top 10 für LLM-Anwendungen 2025; die AWS Agentic AI Security Scoping Matrix bildet dieselben zwei Dimensionen auf vier Kontrollbereiche ab. Die richtige Frage umfasst immer beides zusammen:

Worauf kann dieser Agent zugreifen – und wer entscheidet, wann er handelt?

DEPLOYMENT-KATEGORIE

WER KONTROLLIERT DIE RUNTIME

Selbst entwickelte Agenten – Sie kontrollieren Architektur, Tool-Set und Deployment.

Vom Anbieter betriebene SaaS-Agenten – eingebettet in Produkte, die Sie kaufen; der Anbieter kontrolliert die Runtime, Sie sind für die offengelegten Daten verantwortlich.

Gemanagte Desktop-Agenten – Modell vom Anbieter, Ausführung lokal; sie lesen E-Mails, Dokumente und Kalender auf Rechnern, die Sie verwalten.

Von Mitarbeitenden betriebene Tools – die am schnellsten wachsende und am wenigsten regulierte Kategorie: Coding-Agenten, IDE-Assistenten, Browser-Plug-ins, autonome Daemons. Die Transparenz liegt in den meisten Organisationen nahe null.

AUSFÜHRUNGSUMGEBUNG

WO ES LÄUFT

In der Cloud betriebene Agenten laufen mit serverseitigen Rechten in Umgebungen, die Sie kontrollieren. Von Mitarbeitenden betriebene Tools laufen auf Arbeitsplatzrechnern – und genau dort ist die Lücke im Governance-Tooling am größten.

Der Nutzen des Frameworks: Deployment-Kategorie und Ausführungsumgebung bestimmen, welche Kontrollen Sie tatsächlich durchsetzen können. Beim Tool-Set eines selbst entwickelten Agenten können Sie regulierend eingreifen. Bei einem von Mitarbeitenden installierten Tool ist Ihr Kontrollpunkt der Endpoint, auf dem es liegt.

WARUM DER ENDPOINT DIE CONTROL-PLANE IST

Jede hochriskante Zelle der obigen Klassifizierung – von Mitarbeitenden betriebene Tools, Agency über das gesamte System, agentischer bis autonomer Betrieb – läuft am Arbeitsplatzrechner zusammen. Dort werden Agenten installiert, dort laden Plug-ins und Skills, dort liegen die Konfigurationen der MCP-Server, dort laufen lokale Modelle, und dort liegen die Zugangsdaten auf der Festplatte.

Bestehende Ansätze überwachen jeweils einen Engpass: Browser-Erweiterungen sehen den Browser-Traffic; Netzwerk-Tools sehen den Egress; SaaS- und Identity-Tools sehen Konten und Tokens; Runtime-Monitore für Agenten sehen das Verhalten der Agenten, an die sie angebunden sind. Keiner von ihnen inventarisiert den installierten Zustand des Endpoints: die Binaries, Konfigurationsdateien, Plug-ins, Skills-Verzeichnisse, MCP-Definitionen und Modell-Artefakte, die bestimmen, was KI auf diesem Rechner tun kann – und was sie erreichen kann.

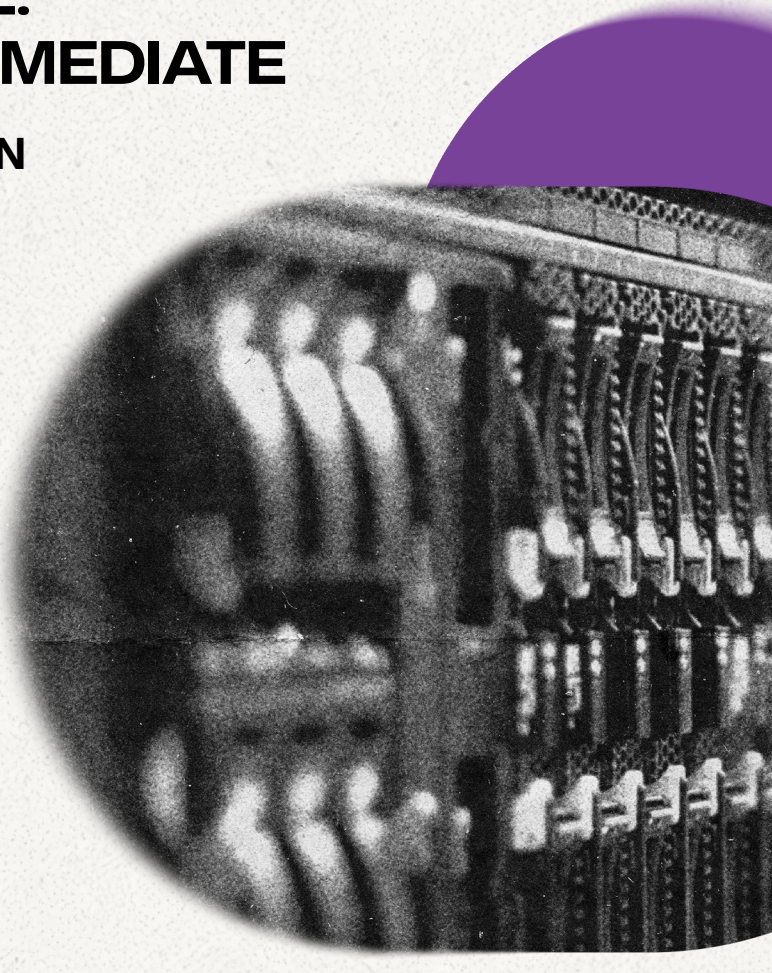
Die Leitlinien der Cloud Security Alliance von 2026 formulieren die Anforderung präzise: Operative Sichtbarkeit ist keine prüffähige Aufsicht. Prüffähigkeit erfordert die Gewissheit, dass alle Agenten – autorisiert oder nicht – identifiziert, in ihrem Umfang erfasst und Kontrollpfaden unterworfen sind. Das ist eine Aussage über Inventar und Posture, nicht über das Abfangen von Traffic. Und ausnahmebasierte Governance funktioniert nur bei bekannten, abgegrenzten Agenten: Eine Policy-Engine läuft ins Leere („fails open“), wenn das Inventar unvollständig ist. Sichtbarkeit ist die Voraussetzung für Policy.

Die Regulierung weist in dieselbe Richtung. Die Anforderungen des EU AI Act an Transparenz, Verantwortlichkeit, Nachvollziehbarkeit und Post-Market-Monitoring machen den blinden Fleck am Endpoint zu einem regulatorischen Risiko. Unternehmen brauchen operative Kontrollen darüber, wie Agenten bereitgestellt werden und was sie tun können – nicht nur Richtlinien, die beschreiben, was sie tun sollten.

DAS BETRIEBSMODELL: DISCOVER, DEFINE, REMEDIATE

DISCOVER: DIE AI-BOM AUFBAUEN

Erzeugen Sie eine kontinuierlich aktualisierte AI Bill of Materials aus den Endpoints selbst: jeder installierte Agent, jedes Browser- und Coding-Tool-Plug-in, jeder konfigurierte MCP-Server, jeder Skill und jedes genutzte Modell. Erfassen Sie den Zustand, nicht nur den Traffic; ein Agent, der noch keinen Netzwerkaufruf gemacht hat, ist trotzdem ein Risiko. Klassifizieren Sie jeden Eintrag entlang der vier Dimensionen oben. Die Einführung sollte keinen weiteren Endpoint-Agenten erfordern: Ein agentenloses Rollout über bestehende Management-Tools wie Microsoft Intune oder CrowdStrike Falcon räumt den wichtigsten operativen Einwand aus.



Vier Fragen, die das Inventar beantworten muss:

01 WELCHE GEFÄHRLICHE SOFTWARE IST VORHANDEN?

Manche Tools sind nützlich und schon von Haus aus tief vernetzt; autonome Daemons sind das klarste Beispiel. Sie brauchen Erkennung und aktives Management, nicht nur eine Auflistung.

02 WELCHE AGENTEN SIND ANGREIFBAR?

Veraltete Versionen, gefährliche Modi, riskante Konfigurationen: klassisches Vulnerability Management, erweitert auf die KI-Landschaft.

03 WELCHE SKILLS SIND ANGREIFBAR ODER BÖSARTIG?

Skills erfordern eine Verhaltensbewertung, kein Signatur-Matching. Mondoo bewertet Skills mit eigenen KI-Verfahren, die auf der veröffentlichten AI-Agent-Traps-Forschung von Google DeepMind aufbauen.

04 WORAUF KANN JEDES TOOL ZUGREIFEN?

Erfassen Sie die Reichweite: die Offenlegung lokaler Secrets, MCP-Server, die Produktionssysteme exponieren, sowie die Keys und Zugangsdaten, die jedes Tool nutzt.

DEFINE: POLICY AS CODE

Übersetzen Sie die Richtlinie zur zulässigen KI-Nutzung in durchsetzbare Regeln: welche Agenten freigegeben sind (und in welchen Modi), welche Skills erlaubt sind (und aus welchen Quellen), welche MCP-Server existieren dürfen (und was sie erreichen dürfen), welche Modelle Unternehmensdaten verarbeiten dürfen. Bewerten Sie kontinuierlich gegen die aktuelle AI-BOM, überlagern Sie das Ergebnis mit Risk Intelligence und priorisieren Sie nach Ausnutzbarkeit und Geschäftsauswirkung – genauso, wie ein ausgereiftes Programm mit CVEs umgeht. **Das ist KI-Governance in der Praxis: was laufen darf und was nicht.**

REMEDIATE: BEHEBEN MIT TOOLS, DIE SIE SCHON BETREIBEN

Findings müssen geschlossen werden, nicht sich anhäufen. Die Behebung läuft über bestehendes Endpoint-Management wie Microsoft Intune und CrowdStrike Falcon: nicht autorisierte Agenten entfernen, riskante Skills deaktivieren, gefährliche Konfigurationen im großen Stil korrigieren. Runtime-Erkennung sagt Ihnen, dass ein Agent sich falsch verhalten hat; die Behebung des Zustands stellt sicher, dass das riskante Tooling nie wieder läuft – und präventive Policies bewirken, dass ein Großteil davon überhaupt nie läuft.

EINE 90-TAGE-ROADMAP

TAG 0-30	**Baseline.** Rollen Sie Discovery auf eine repräsentative Stichprobe von Endpoints aus (agentenlos, über Ihr bestehendes MDM/EDR). Bauen Sie die erste AI-BOM auf. Rechnen Sie mit Überraschungen: Laut den CSA-Daten finden 82% der Unternehmen Agenten, von denen sie nichts wussten.
TAG 31-60	**Klassifizieren und definieren.** Bewerten Sie das Inventar nach Agency × Autonomie. Entwerfen Sie die Allowlist gemeinsam mit Engineering und Geschäftsleitung; Governance, die Produktivität ignoriert, wird umgangen. Veröffentlichen Sie die Policy als Code.
TAG 61-90	**Durchsetzen und operationalisieren.** Führen Sie Findings in den bestehenden Vulnerability-Workflow ein. Aktivieren Sie die Behebung über das Endpoint-Management. Berichten Sie Shadow-AI-Kennzahlen (gefundene unbekannte Agenten, deaktivierte riskante Skills, MTTR) neben Ihren bestehenden Exposure-Kennzahlen.

SECURITY ALS WEGBEREITER DER KI-EINFÜHRUNG

Das Muster ist bekannt. Open-Source-Abhängigkeiten waren einst ungesteuert; SBOMs und SCA haben sie zu einer gemanagten Asset-Klasse gemacht, ohne dass jemand aufhören musste, Open Source zu nutzen. Cloud-Ressourcen waren einst Schatten-Infrastruktur; CSPM hat sie sichtbar und kontrollierbar gemacht, ohne die Cloud-Einführung zu bremsen. KI-Tooling steht am selben Wendepunkt, und die Antwort ist dieselbe: Inventar, kontinuierliche Bewertung, Policy, Behebung – innerhalb des Vulnerability-Management-Programms, das Sie schon heute betreiben.

Wer das richtig macht, dessen Security ist nicht länger die Abteilung, die zu KI Nein sagt. Sie wird zur Funktion, die dem Unternehmen erlaubt, KI mit voller Geschwindigkeit einzuführen – mit Nachweisen: ein aktuelles Inventar dessen, was läuft, ein Beleg dafür, worauf es zugreifen kann, und eine belastbare Antwort auf die Shadow-AI-Frage des Vorstands.

FAZIT

Mondoo hat seine Plattform auf dieser Überzeugung aufgebaut: Probleme zu finden ist Pflicht; **das Ergebnis sind Probleme, die aufhören zu existieren**. Das Gleiche gilt jetzt für die KI-Landschaft. Entdecken Sie jeden Agenten, jedes Plug-in, jeden MCP-Server, jeden Skill und jedes Modell.

QUELLEN & WEITERFÜHRENDE LEKTÜRE

CSA / Token Security, Autonomous but Not Controlled (April 2026): Erkennung von Shadow-Agenten und Vorfallsraten

Cisco, State of AI Security 2026: Bereitschaft für agentische KI

Academic skill-security studies (Jan/Feb 2026): Verbreitung von Schwach-stellen, bestätigte bösartige Skills, Konzentration auf einen einzelnen Akteur

OWASP Top 10 for LLM Applications 2025: LLM06 Excessive Agency; OWASP GenAI Security Project (Mondoo, silver sponsor)

AWS Agentic AI Security Scoping Matrix (November 2025)

Google DeepMind, AI Agent Traps: sechs Klassen umgebungsvermittelter Angriffe auf Agenten

Mondoo Blogserie: *Securing AI Agents, A Practical Guide for Security and Engineering Leaders; Agent Skills: Real Power, Real Risk; Introducing Mondoo AI Skills Check*

DEFINIEREN SIE, WAS LAUFEN DARF. BEHEBEN SIE ALLES ANDERE, BEVOR ES ZUM VORFALL WIRD.

ASSESSMENT ANFORDERN