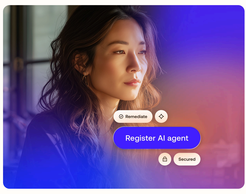
Ein Konzept für den sicheren Einsatz von Agenten in Unternehmen

Drei Fragen, die alle Sicherheits- und IT-Verantwortlichen beantworten sollten, bevor der Einsatz von Agenten überhand nimmt.



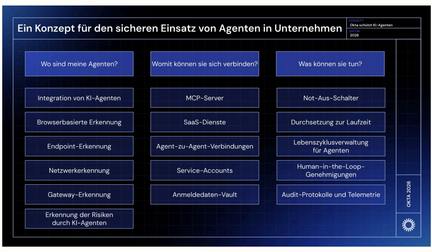
Die Identity-Management-Lücke bei der Gewährleistung von KI-Sicherheit

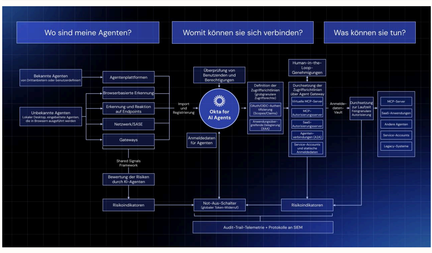
Früher hat die Software genau das getan, was man von ihr wollte. Nun entscheidet, handelt und vernetzt sie sich selbstständig. KI-Agenten unterstützen bereits Mitarbeitende, bedienen Kundinnen und sind entlang von Lieferketten im Einsatz – doch sie skalieren schneller als die entsprechenden Sicherheitsmaßnahmen. Im Laufe der letzten zehn Jahre haben Unternehmen die Identity-Sicherheit für Menschen durch Vaulting, das Least-Privilege-Prinzip und kontinuierliche Authentifizierung gestärkt. Der rasante Aufstieg von KI-Agenten schafft jedoch eine neue Identity-Management-Lücke. Jeder kann einen Agenten starten, ein Agent kann weitere Agenten starten, und jeder von ihnen verbindet sich mit Anwendungen, APIs, SaaS-Tools und Datensystemen. Das Ergebnis sind Tausende neue Entitäten mit privilegiertem Zugriff, die mit Maschinengeschwindigkeit arbeiten – oft außerhalb bestehender Sicherheitskontrollen.

Deshalb müssen Agenten als vollwertige Identitäten behandelt werden. Die Sicherheit eines agentenbasierten Unternehmens erfordert von Anfang an klare Zuständigkeiten und volle Transparenz, um eine unkontrollierte Skalierung der Agenten zu verhindern. Unternehmen sollten hierzu drei Fragen beantworten können:

1. Wo sind meine Agenten?
2. Womit können sie sich verbinden?
3. Was können sie tun?

Diese Fragen definieren das operative Konzept für die Absicherung von KI-Agenten. Bei der Beantwortung dieser Fragen geht es nicht um eine Bestandsaufnahme. Um ein sicheres agentenbasiertes Unternehmen zu betreiben, sind die richtigen Systeme, Identity-Kontrollen und ein geeignetes Governance-Modell notwendig.





Frage 1: Wo sind meine Agenten?

Nach Meinung unserer Kundinnen und Kunden ist Transparenz der zentrale Faktor.

Fragen Sie Ihr Team, wie viele Agenten in Ihrer Umgebung eingesetzt werden. Die meisten können Ihnen keine genaue Zahl nennen. Die Agenten, die von Mitarbeitenden in einem Browser gestartet werden oder die unbemerkt auf Desktop-Computern laufen, sind oft unbekannt und unterliegen keinerlei Kontrolle.

Daher benötigen Sie Möglichkeiten, um Agenten überall zu erkennen – unabhängig davon, wo sie erstellt oder eingesetzt wurden.

- **Agentenbasierte Plattformintegrationen:** Registrieren Sie Agenten von großen Drittanbieterplattformen und Ihre eigenen benutzerdefiniert erstellten Agenten bei Ihrem Identity-Anbieter. Wenn Sie diese Agenten bei ihrer Entstehung nicht sehen können, lassen sie sich auch nicht kontrollieren.
- **Browserbasierte Erkennung:** Entdecken Sie Schatten-Agenten, die über Browser und Erweiterungen agieren und sich außerhalb der Kontrolle Ihres Identity-Anbieters befinden. Dies sind Agenten, die von Mitarbeitenden ohne vorherige Berechtigung gestartet werden.
- **Endpoint-Erkennung:** Identifizieren Sie Agenten, die auf verwalteten Geräten ausgeführt werden. Dies sollte sich nahtlos in Ihre bestehende Lösung für Mobilgerätemanagement oder Endpoint-Sicherheit integrieren lassen.
- **Netzwerkerkennung:** Erkennen Sie nicht autorisierten Datenverkehr von Agent-zu-Agent und Agent-zu-Ressource auf der Netzwerkebene. Die Agenten kommunizieren untereinander und mit den verschiedenen Diensten, und Sie müssen diese Verbindungen erkennen.
- **Gateway-Erkennung:** Identifizieren und kontrollieren Sie nicht registrierte KI-Agenten und OAuth-Clients, die mit Ihrer API, Ihrem MCP und Agenten-Gateways interagieren. Wenn Agenten Ihre APIs aufrufen, sollten sie authentifiziert und ihre Aktionen protokolliert werden.
- **Risiko-Bewertung für KI-Agenten:** Überwachen Sie fortlaufend Konfigurationsfehler, durch die KI-Agenten angreifbar werden, und decken Sie diese auf. Durch die Analyse der Sicherheitsdaten jeder Agentenidentität werden Risiken proaktiv aufgedeckt.

Unabhängig vom verwendeten Tech-Stack müssen Sie Signale aus all diesen Quellen verarbeiten. Ihre Erkennungsebene muss Plattform-, Tool- und Team-übergreifend funktionieren. Fragmentierte Transparenz ist keine echte Transparenz.

Frage 2: Womit können sie sich verbinden?

Sobald Sie einen Agenten sehen können, müssen Sie alle für ihn verfügbaren Ressourcen zuordnen und die Zugriffsrichtlinie durchsetzen. Ohne zentrale Kontrolle über die Verbindungen kann ein einziger kompromittierter Agent mit Maschinengeschwindigkeit Zugriffe in Ihrer gesamten Umgebung verknüpfen.

Der Wirkungsradius eines kompromittierten Agenten wird durch seine Verbindungen bestimmt.

- **MCP-Server und Ressourcen:** MCP-Server (Model Context Protocol) ermöglichen Agenten den Zugriff auf Tools und Datenquellen – sowohl interne (Apps, APIs, Datenbanken, geistiges Eigentum) als auch externe (Drittanbieter-MCPs wie Slack, GitHub und Notion). Ihr Sicherheitsperimeter erstreckt sich nun auf alle Ressourcen, die Ihr Agent erreichen kann.
- **SaaS-Anwendungen:** Agenten verbinden sich mit denselben SaaS-Tools, die auch Ihre Mitarbeitenden täglich verwenden. Der Unterschied besteht darin, dass Agenten wesentlich schneller arbeiten und auf mehr Daten zugreifen. Ein kompromittierter Agent kann schneller Daten exfiltrieren oder Änderungen an jeder verbundenen SaaS-Anwendung vornehmen, als es ein Mensch je könnte.
- **Agent-zu-Agent-Verbindungen:** Sichern Sie den Handshake und die Autorisierung zwischen autonomen Entitäten ab. Die laterale Bewegung beginnt, wenn ein Agent einen anderen Agenten aufruft. Beide Seiten müssen die Identität überprüfen, bevor Daten ausgetauscht oder Aufgaben delegiert werden.
- **Service-Account:** Beseitigen Sie den Wildwuchs an statischen, langlebigen Anmeldedaten. Dies sind die Generalschlüssel, die Agenten von Legacy-Machine-to-Machine-Mustern erben. Alle statischen Anmeldedaten stellen eine dauerhafte Hintertür dar, die nur darauf wartet, ausgenutzt zu werden.
- **Im Vault gesicherte Anmeldedaten:** Schützen und rotieren Sie Secrets automatisch. Ein nicht rotiertes Token ist für Angreifende eine offene Tür. Anmeldedaten sollten sicher gespeichert, dynamisch herausgegeben und regelmäßig rotiert werden. Kein Agent sollte mit Anmeldedaten arbeiten, die ihren Zweck überdauert haben.

Alle diese Verbindungen müssen in Ihrem SIEM-System protokolliert werden. Was man nicht sieht, kann man nicht sichern. Jede Agentenverbindung und Angaben dazu, wer die Inhalte zu welchem Zeitpunkt und mit welchen Anmeldedaten aufgerufen hat, sollte zur Überwachung und Untersuchung an Ihr Security Operations Center weitergeleitet werden.

Frage 3: Was können sie tun?

Es genügt nicht zu wissen, wo sich die Agenten befinden und womit sie sich verbinden können, wenn Sie nicht kontrollieren und unterbinden können, was sie tatsächlich tun. Wenn ein Agent mit der Exfiltration von Daten beginnt oder unautorisierte Prozesse startet, müssen Sie schnell reagieren.

- **Not-Aus-Schalter:** Wenn ein Agent von seiner beabsichtigten Aufgabe abweicht, unerwartet auf sensible Daten zugreift oder eine Bedrohung erkannt wird, müssen Sie den Zugriff auf allen Systemen sofort widerrufen können, um das Risiko einzudämmen.
- **Durchsetzung zur Laufzeit:** Autorisieren Sie Agenten basierend auf dem, was sie in Echtzeit erreichen wollen. Bewerten Sie Kontext, Abfolge und Volumen. Eine Anfrage nach 10 Kundendatensätzen sieht anders aus als eine Abfrage nach 10.000. Sie müssen Prompt-Injection-Angriffe erkennen und Richtlinien auf Tool-Ebene durchsetzen, bevor Aktionen ausgeführt werden.
- **Lebenszyklusverwaltung für Agenten:** Agentenberechtigungen, die am ersten Tag sinnvoll sind, haben nach 90 Tagen kaum noch ihre Gültigkeit. Überprüfen Sie die Zugriffsrechte kontinuierlich, um das Least-Privilege-Prinzip durchzusetzen, automatisieren Sie Zertifizierungen und widerrufen Sie den Zugriff unverzüglich, wenn Agenten außer Betrieb genommen werden oder Mitarbeitende das Unternehmen verlassen.
- **Human-in-the-Loop-Genehmigungen:** Für sensible oder potenziell riskante Agentenaktionen sollte eine Genehmigung durch einen Menschen erforderlich sein. Verhindern Sie destruktive Operationen, massenhaften Datenzugriff oder Rechteausweitung für Agenten.
- **Audit-Protokolle und Telemetrie:** Jede Agentenaktion muss protokolliert und an Ihr SIEM-System gesendet werden. Das gilt für jeden Tool-Aufruf, jede Autorisierungsentscheidung und jeden Zugriffsversuch. Laufzeitüberwachung und Not-Aus-Schalter funktionieren nur, wenn Sie vollständige Transparenz haben.

Das Konzept als Grundlage

Die drei Fragen – wo sind meine Agenten, womit können sie sich verbinden, was können sie tun – sind kein fernes Ziel, sondern sie stellen den Mindeststandard dar, den Sie für den Betrieb von KI-Agenten in der Produktion benötigen.

Unternehmen, die diese Fragen nicht beantworten können, tappen im Dunkeln. Und wenn nach einer Sicherheitsverletzung der Vorstand, prüfende Personen und die Aufsichtsbehörde Fragen stellen, ist „Keine Ahnung" die falsche Antwort.

Die Vorreiter haben das bereits erkannt. Führende Unternehmen handeln proaktiv und warten nicht erst auf einen Sicherheitsvorfall. Sie behandeln Agenten schon jetzt wie vollwertige Identitäten und integrieren Erkennung, Durchsetzung und Governance von Anfang an in ihre Agenten-Implementierungen. Dadurch können sie die drei Fragen beantworten, bevor ihr Agenteneinsatz überhand nimmt.

Das Konzept ist nicht als einmalige Prüfung gedacht. Die Beantwortung dieser drei Fragen ist eine fortlaufende operative Disziplin, wenn Ihr Agentenbestand von Dutzenden auf Tausende anwächst. Sie haben ein Jahrzehnt mit der Implementierung von sicherem Identity-Management verbracht. Diese Arbeit darf nicht durch den unkontrollierten Einsatz von Agenten gefährdet werden.

Okta hat dieses Konzept durch kontinuierliche Zusammenarbeit mit führenden Unternehmen mit dem Ziel entwickelt, KI-Agenten im großen Umfang abzusichern. Unter okta.com/ai-agents erfahren Sie, wie die Okta Platform dies implementiert.