

LEITFADEN · TOMORIS TOOLBOX

KI-Agenten sicher einsetzen: Rechte, Freigaben, MCP-Server und Protokollierung

Risiken nach OWASP, Grundsätze für Rechte und Freigaben, MCP-Server prüfen und isolieren, Einführung in drei Stufen und Prüfliste vor dem Produktivbetrieb

STAND

25.10.2026

UMFANG

PDF, 8 Seiten

ONLINE

tomoris.com/downloads/ki-agenten-sicher-einsetzen

AUF EINEN BLICK

3

Ursachen übermäßiger Handlungsmacht laut OWASP LLM06

ASI01–10

OWASP Top 10 für agentische Anwendungen 2026

MUST NOT

Token-Weitergabe in der MCP-Spezifikation

3

Stufen für die Einführung, dazu laufende Prüfung

FÜR WEN

- IT-Leitung und IT-Sicherheit, die KI-Agenten freigeben und betreiben
- Entwicklung, die Agenten, Werkzeuge oder MCP-Server baut oder einbindet
- Fachbereiche, die Agenten mit Zugriff auf E-Mail, Dateien oder Fachanwendungen planen
- Geschäftsführung, die über Einsatz und Grenzen entscheidet

WURUM ES GEHT

KI-Agenten planen selbstständig mehrere Schritte und rufen dabei Werkzeuge auf: Sie lesen Dateien, schreiben E-Mails, ändern Code oder fragen Datenbanken ab. Damit wird aus einer falschen oder manipulierten Antwort eine falsche Handlung. Dieser Leitfaden ordnet die wichtigsten Risiken nach OWASP ein und zeigt, wie Sie Rechte begrenzen, folgenreiche Aktionen freigeben lassen, Inhalte Dritter als nicht vertrauenswürdig behandeln, MCP-Server prüfen und Agenten isoliert und nachvollziehbar betreiben. Am Ende steht eine Prüfliste für den Schritt in den Produktivbetrieb.

Dieser Leitfaden ist eine Empfehlung von TOMORIS auf Grundlage der genannten Quellen. Funktionen von Produkten sind nach Herstellerdokumentation zum Stand 25.10.2026 beschrieben und können sich ändern. Er ersetzt keinen Sicherheitstest und keine Rechtsberatung.

INHALT

01	Was Agenten riskanter macht als Chat	2
02	Sechs Grundsätze für Rechte und Freigaben	3
03	Identitäten und Zugangsdaten	4
04	MCP-Server prüfen und betreiben	5
05	Beispiel: Wie ein Werkzeug die Grundsätze umsetzt	6
06	Einführung in drei Stufen	6
07	Prüfliste vor dem Produktivbetrieb	7
08	Quellen	8

01

Was Agenten riskanter macht als Chat

Agenten handeln: Die wichtigsten Risiken nach OWASP Top 10 for LLM Applications 2025 und den Top 10 für agentische Anwendungen 2026.

Ein Chat-Assistent antwortet, ein Agent handelt: Er erhält Werkzeuge, etwa Dateizugriff, E-Mail, Shell oder Schnittstellen, und entscheidet selbst, wann er sie aufruft. OWASP nennt die Folge „Excessive Agency“ (LLM06:2025): schädliche Aktionen als Reaktion auf unerwartete, mehrdeutige oder manipulierte Ausgaben des Modells, gleich warum es sich fehlverhält. Die Ursachen sind zu viel Funktionalität, zu viele Berechtigungen und zu viel Autonomie. Die OWASP Top 10 für agentische Anwendungen 2026 (veröffentlicht am 9.12.2025) erweitern das um den Grundsatz der geringsten Handlungsmacht (Least Agency): Agentisches Verhalten dort einsetzen, wo es nötig ist, weil es sonst die Angriffsfläche ohne Nutzen vergrößert.

Prompt Injection (LLM01)	Eingaben verändern das Verhalten des Modells ungewollt. Indirekt geschieht das über Inhalte Dritter wie Websites, Dateien oder E-Mails; die Anweisungen müssen für Menschen nicht sichtbar sein. OWASP hält einen sicheren Ausschluss für unklar und empfiehlt, die Folgen zu begrenzen.
Agent Goal Hijack (ASI01)	Angreifer verändern Ziele, Aufgabenwahl oder Entscheidungswege des Agenten, etwa über Prompts, täuschende Werkzeugausgaben, manipulierte Dateien oder vergiftete externe Daten. Agenten können Anweisungen nicht zuverlässig von Inhalten unterscheiden.
Tool Misuse (ASI02)	Der Agent missbraucht legitime Werkzeuge innerhalb seiner Pläne und Delegationsketten, etwa indem er mit vorhandenen Rechten mehr tut als vorgesehen. OWASP baut dabei auf den Maßnahmen von LLM06 auf.
Identity & Privilege Abuse (ASI03)	Delegationsketten, geerbte Rollen oder zwischengespeicherte Zugangsdaten werden ausgenutzt, um Rechte auszuweiten. Ohne eigene, verwaltete Identität des Agenten lässt sich geringste Berechtigung laut OWASP nicht wirklich durchsetzen.
Agentic Supply Chain (ASI04)	Agenten, Werkzeuge, Modelle, MCP-Server oder Verzeichnisse von Dritten sind bösartig, kompromittiert oder auf dem Weg manipuliert. Agenten laden solche Fähigkeiten oft erst zur Laufzeit.

02

Sechs Grundsätze für Rechte und Freigaben

Was OWASP gegen übermäßige Handlungsmacht und Prompt Injection empfiehlt, als Regeln für Ihre Agenten.

1

Werkzeuge auf das Nötige beschränken

Nur die Erweiterungen anbieten, die der Agent für seine Aufgabe braucht, und darin nur die nötigen Funktionen: Wer E-Mails zusammenfasst, braucht Lesen, nicht Senden oder Löschen. Offene Werkzeuge wie beliebige Shell-Befehle oder URL-Abrufe vermeiden (OWASP LLM06, Maßnahmen 1 bis 3).

2

Berechtigungen minimieren

Jedem Werkzeug nur die Rechte geben, die es braucht, etwa Lesezugriff auf eine Tabelle statt auf die ganze Datenbank, und es im Kontext des jeweiligen Benutzers ausführen (LLM06, Maßnahmen 4 und 5).

3

Folgenreiche Aktionen freigeben lassen

Ein Mensch bestätigt Aktionen mit großer Wirkung, bevor sie ausgeführt werden: senden, löschen, bezahlen, Berechtigungen ändern, veröffentlichen (LLM06, Maßnahme 6; LLM01, Maßnahme 5).

4

Autorisierung nicht dem Modell überlassen

Ob eine Aktion erlaubt ist, entscheidet das Zielsystem nach seinen Regeln, nicht das Sprachmodell. Jede Anfrage über ein Werkzeug wird dort gegen die Sicherheitsrichtlinie geprüft (LLM06, Maßnahme 7: complete mediation).

5

Fremde Inhalte als nicht vertrauenswürdig behandeln

Websites, Dokumente, E-Mails und Werkzeugbeschreibungen von MCP-Servern können Anweisungen enthalten. Sie werden getrennt und gekennzeichnet verarbeitet und dürfen keine Aktionen ohne Prüfung auslösen (LLM01, Maßnahme 6; MCP-Spezifikation).

6

Protokollieren und begrenzen

Aufrufe von Werkzeugen und Aktionen in Zielsystemen protokollieren und überwachen, Raten und Budgets begrenzen, damit Fehlverhalten auffällt, bevor großer Schaden entsteht (LLM06; OWASP ASI02, Maßnahme 5).

Ein Beispiel aus OWASP LLM06: Ein Assistent, der E-Mails zusammenfasst, wird mit einer Erweiterung gebaut, die nur lesen kann, meldet sich mit einem OAuth-Zugang mit reinem Lese-Umfang an, und jede von ihm entworfene E-Mail sendet erst der Mensch nach Durchsicht ab.

03

Identitäten und Zugangsdaten

Eigene, eng begrenzte und kurzlebige Zugänge statt geteilter Administratorkonten.

- Eigene Identität je Agent und Aufgabe statt des persönlichen Kontos einer Administratorin oder eines Entwicklers; so bleibt nachvollziehbar, wer gehandelt hat (OWASP ASI03).
- Kurzlebige, eng begrenzte Token je Aufgabe statt dauerhafter Schlüssel; Rechte über Grenzen deckeln (ASI03, Maßnahme 1; ASI02: just-in-time und kurzlebige Zugänge).
- Zugangsdaten je MCP-Server getrennt und eng begrenzt, nichts aus der eigenen Shell geerbt, nichts wiederverwendet (MCP Local Server Security, Checkliste).
- MCP-Server dürfen keine Token annehmen, die nicht ausdrücklich für sie ausgestellt wurden, und Token nicht ungeprüft an nachgelagerte Schnittstellen weiterreichen (MCP Security Best Practices: Token Passthrough, „MUST NOT“).
- Geheimnisse gehören in einen Tresor oder die Anmeldedatenverwaltung des Systems, nicht in Prompts, Chatverläufe oder Konfigurationsdateien im Repository.
- Zugänge eines Agenten müssen sich zentral sperren und erneuern lassen, wenn er sich auffällig verhält (ASI04 und ASI10: Notabschaltung und Widerruf von Zugangsdaten).

04 —

MCP-Server prüfen und betreiben

Das Model Context Protocol verbindet Agenten mit Werkzeugen; jeder Server ist Code mit Zugriff. So prüfen Sie ihn.

MCP-Server stellen einem Agenten Werkzeuge, Daten und Vorlagen bereit. Die Spezifikation (Fassung 2026-07-28) sagt dazu: Werkzeuge bedeuten beliebige Codeausführung und sind entsprechend vorsichtig zu behandeln; Beschreibungen des Werkzeugverhaltens gelten als nicht vertrauenswürdig, außer sie stammen von einem vertrauenswürdigen Server; Hosts müssen vor dem Aufruf eines Werkzeugs die ausdrückliche Zustimmung des Benutzers einholen. Durchsetzen kann das Protokoll diese Grundsätze nicht selbst, das müssen die Anwendungen tun.

- ☐ Braucht der Server lokale Dateien, Werkzeuge oder das lokale Netz? Wenn nicht, die entfernte Fassung über Streamable HTTP mit Autorisierung statt eines lokalen Prozesses nutzen
- ☐ Herausgeber identifiziert, Quellcode einsehbar, exakter Paketname geprüft (Tippfehler-Pakete), Version fest angeheftet
- ☐ Bei Starthilfen wie npx oder uvx bedenken, dass die Versionsangabe nur das oberste Paket festlegt; ein Container-Image mit Digest friert alle Abhängigkeiten ein
- ☐ Werkzeugbeschreibungen vor dem Einsatz gelesen; Server mit fremden Anweisungen in Beschreibungen entfernt; Client fragt erneut, wenn sich Werkzeuge ändern
- ☐ Fremde Server isoliert in Container, Sandbox oder virtueller Maschine, ohne unnötige Laufwerkseinbindungen
- ☐ Dateizugriff auf bestimmte Verzeichnisse beschränkt, wo möglich nur lesend
- ☐ Ausgehender Netzverkehr gesperrt oder auf eine bekannte Liste von Zielen beschränkt
- ☐ Nur so viele Server wie nötig, weil alle konfigurierten Server denselben Kontext des Modells teilen
- ☐ Im Unternehmen: Liste freigegebener Server mit Kriterien, laufende Inventur der Client-Konfigurationen, Protokolle der MCP-Aktivität, Server zentral sperrbar

Die Prüfpunkte folgen dem Leitfaden „Local Server Security“ der MCP-Dokumentation (Fassung 2026-07-28). Er nennt die Isolation des Prozesses die wirksamste einzelne Maßnahme, weil sie auch dann schützt, wenn die Prüfung der Herkunft versagt.

05 —

Beispiel: Wie ein Werkzeug die Grundsätze umsetzt

Funktionen von Claude Code laut Anthropic-Dokumentation, als Beispiel für die Prüfung beliebiger Agenten-Werkzeuge.

Grundsatz und Umsetzung in Claude Code laut Sicherheitsdokumentation von Anthropic (Stand 25.10.2026)

GRUNDSATZ	UMSETZUNG LAUT DOKUMENTATION
Freigaben	Berechtigungsmodi legen fest, was ohne Rückfrage geschieht; im manuellen Modus fragt Claude Code vor Dateiänderungen, Tests und Befehlen, und Befehle ohne passende Regel brauchen eine Freigabe
Isolation	Sandbox für Shell-Befehle mit Trennung von Dateisystem und Netzwerk; für Skripte und Werkzeugaufrufe mit externen Diensten empfiehlt Anthropic virtuelle Maschinen
Fremde Inhalte	Befehle, die Inhalte aus dem Web laden, werden standardmäßig nicht automatisch freigegeben; Anthropic rät, nicht vertrauenswürdige Inhalte nicht direkt einzuspeisen
MCP-Server	Server aus der Projektkonfiguration brauchen eine eigene Zustimmung; Anthropic empfiehlt eigene Server oder solche von vertrauenswürdigen Anbietern
Vorgaben und Überwachung	Verwaltete Einstellungen setzen Vorgaben der Organisation durch, Berechtigungen lassen sich regelmäßig prüfen, die Nutzung über OpenTelemetry überwachen

Das Beispiel beschreibt Funktionen laut Herstellerdokumentation, keine Bewertung oder Empfehlung eines Produkts. Anthropic weist selbst darauf hin, dass kein System gegen alle Angriffe immun ist und Vorschläge vor der Freigabe zu prüfen sind.

06 —

Einführung in drei Stufen

Autonomie schrittweise erweitern und laufend prüfen, jede Stufe mit Protokollen und klarer Verantwortung; Empfehlung von TOMORIS.



Stufe 1

Nur lesen

Der Agent recherchiert, fasst zusammen und schlägt vor; er hat nur Lesezugriff auf freigegebene Quellen.



Stufe 2

Vorschlag mit Freigabe

Der Agent bereitet Änderungen vor; jede Aktion mit Wirkung bestätigt ein Mensch einzeln.



Stufe 3

Begrenzt automatisch

Eng umrissene, umkehrbare Aktionen laufen ohne Rückfrage, in einer Sandbox und mit Budget; alles andere bleibt freigabepflichtig.



Laufend

Überprüfen

Protokolle auswerten, Rechte und Server-Liste regelmäßig prüfen, nach Vorfällen oder neuen Funktionen eine Stufe zurück.

Unabhängig von der Stufe verlangt Art. 4 der KI-Verordnung in der Fassung der VO (EU) 2026/1744, dass Anbieter und Betreiber Maßnahmen ergreifen, um die KI-Kompetenz der Personen zu unterstützen, die in ihrem Auftrag mit KI-Systemen arbeiten. Wer Agenten freigibt oder überwacht, muss ihre Grenzen kennen.

07 —

Prüfliste vor dem Produktivbetrieb

Zum Abhaken, bevor ein Agent mit echten Daten und Systemen arbeitet.

- ☐ Aufgabe, Grenzen und verantwortliche Person des Agenten schriftlich festgelegt
- ☐ Agentisches Verhalten ist für die Aufgabe nötig; einfachere, feste Abläufe wurden geprüft
- ☐ Nur nötige Werkzeuge und Funktionen freigeschaltet, keine offenen Shell- oder Abrufwerkzeuge ohne Grund
- ☐ Eigene Identität mit geringsten Rechten, kurzlebige Zugänge, keine geteilten Administratorkonten
- ☐ Folgenreiche Aktionen brauchen eine menschliche Freigabe; die Liste dieser Aktionen ist festgelegt
- ☐ Zielsysteme prüfen jede Anfrage selbst gegen ihre Berechtigungen
- ☐ Inhalte Dritter werden als nicht vertrauenswürdig behandelt; Tests mit eingeschleusten Anweisungen durchgeführt
- ☐ MCP-Server nach Herkunft geprüft, Version angeheftet, isoliert, mit eigenem eng begrenztem Zugang
- ☐ Netzzugang des Agenten und seiner Werkzeuge auf bekannte Ziele beschränkt
- ☐ Alle Werkzeugaufrufe und Aktionen protokolliert, Auswertung und Aufbewahrung geregelt
- ☐ Budgets und Ratenbegrenzung gesetzt, Notabschaltung und Widerruf der Zugänge getestet
- ☐ Datenschutz geprüft: Rechtsgrundlage, Vertrag mit dem Anbieter, gegebenenfalls Datenschutz-Folgenabschätzung
- ☐ Nutzende geschult; Meldeweg für auffälliges Verhalten bekannt

Die Prüfliste ist eine Empfehlung von TOMORIS auf Grundlage der genannten OWASP- und MCP-Veröffentlichungen; sie ersetzt keinen Sicherheitstest des konkreten Agenten.

NÄCHSTER SCHRITT

Selbst prüfen oder gemeinsam umsetzen

UMSETZUNG BESPRECHEN

Kontakt aufnehmen

tomoris.com/contact

contact@tomoris.com

08 —

Quellen

Alle Aussagen sind belegt. Gesetzestexte gelten in der jeweils geltenden Fassung.

- OWASP Top 10 for LLM Applications 2025: LLM06 Excessive Agency**
Ursachen, Maßnahmen 1 bis 7, Begrenzung durch Protokolle und Ratenlimits, Beispiel E-Mail-Assistent
genai.owasp.org/llmrisk/llm062025-excessive-agency
- OWASP Top 10 for LLM Applications 2025: LLM01 Prompt Injection**
direkte und indirekte Prompt Injection, Maßnahmen 4 bis 6
genai.owasp.org/llmrisk/llm01-prompt-injection
- OWASP Top 10 for Agentic Applications for 2026 (9.12.2025)**
Least Agency, ASI01 bis ASI04 und ASI10 mit Maßnahmen
genai.owasp.org/resource/owasp-top-10-for-agentic-applications-...
- Model Context Protocol, Spezifikation 2026-07-28**
Abschnitt Security and Trust & Safety
modelcontextprotocol.io/specification/2026-07-28
- Model Context Protocol: Security Best Practices (2026-07-28)**
Token Passthrough, Local MCP Server Compromise
modelcontextprotocol.io/docs/2026-07-28/tutorials/security/secu...
- Model Context Protocol: Local Server Security (2026-07-28)**
Herkunft, Werkzeugbeschreibungen, Isolation, Zugangsdaten, Netz, Checkliste
modelcontextprotocol.io/docs/2026-07-28/tutorials/security/loca...
- Anthropic: Claude Code Security**
Berechtigungen, Sandbox, Prompt Injection, MCP, Team-Sicherheit
code.claude.com/docs/en/security
- Verordnung (EU) 2026/1744 (Digital-Omnibus-Verordnung zur KI)**
neuer Art. 4 der VO (EU) 2024/1689, in Kraft seit 27.7.2026
eur-lex.europa.eu/eli/reg/2026/1744/oj/deu