CLAWGUARD ARBEITSDOKUMENT

MCP-Security-Checkliste 2026

Die 10 kritischen Prüfpunkte für KI-Agenten und MCP-Server — mit Test-Anleitung, Härtungs-Maßnahmen und EU-AI-Act-Zuordnung. Zum Durcharbeiten, Abhaken und als Nachweis für Ihre Dokumentation.

> **Basis dieses Dokuments:** Scan von **11.529 öffentlichen MCP-Servern** (ClawGuard Registry-Studie 2026). **7,4 %** hatten Zugriff auf Dateisysteme, Datenbanken, Zahlungssysteme oder Cloud-Infrastruktur — ohne wirksame Zugriffskontrolle.

Version 1.0 · August 2026 · Jörg Michno, ClawGuard · prompttools.co

Warum diese Checkliste jetzt?

Das Model Context Protocol (MCP) ist 2026 die Standard-Schicht zwischen LLMs und der Außenwelt — Dateisysteme, Datenbanken, APIs, Cloud. Claude, Cursor, VS Code und nahezu alle Agenten-Frameworks bringen MCP-Support mit. Dieselbe Konnektivität, die MCP mächtig macht, ist die neue Angriffsfläche.

Drei Gründe, warum „später" keine Option mehr ist:

1. **Regulierung läuft:** Seit 02.08.2026 gelten die Transparenzpflichten des EU AI Act (Art. 50). Die Hochrisiko-Pflichten folgen ab 02.12.2027 (Digital Omnibus) — inklusive Nachweis-Anforderungen an Logging (Art. 12), menschliche Aufsicht (Art. 14) und technische Robustheit (Art. 15). Wer jetzt prüft und dokumentiert, hat 2027 keine Hektik.
2. **Angriffsforschung ist ausgereift:** Tool Poisoning (Invariant Labs), OWASP-Eintrag zu MCP Tool Poisoning, CISA/NSA-Cybersecurity-Advisory zu MCP (Juni 2026) — die Angriffsklassen sind dokumentiert und werden genutzt.
3. **Reale Funde:** Klartext-Zugangsdaten in Configs, unauthentifizierte Endpunkte, Tool-Scopes ohne Begrenzung — das ist der Alltag in öffentlichen MCP-Registries, nicht die Ausnahme.

Wie Sie dieses Dokument nutzen

Jeder Prüfpunkt folgt demselben Schema: **Risiko** (was ist das Problem) → **Test** (konkret prüfbar) → **Härtung** (was tun) → **AI-Act-Bezug** (wo es in der Dokumentation zählt). Haken Sie ab: ☐ getestet · ☐ Befund dokumentiert · ☐ behoben. Die Schnell-Checkliste am Ende (S. 6) ist die 1-Seiten-Version für die Besprechung.

Die 10 Prüfpunkte

1 Tool Poisoning — versteckte Anweisungen in Tool-Metadaten

RISIKO: Tool-Beschreibungen, Parameternamen und Response-Schemas liest das Modell als Kontext. Wer eine Tool-Definition kontrolliert, kann unsichtbare Anweisungen einbetten („ignore previous instructions", Exfiltrations-Aufrufe) — der Nutzer sieht sie nie.

TEST: Jede `description`, jedes `inputSchema` und Beispiel-Antwort aller Tools auf eingebettete Anweisungen prüfen. Warnsignale: Imperative an das Modell, Verweise auf nicht existierende Tools, URLs, Base64-Blöcke.

HÄRTUNG: Tool-Definitionen an eine geprüfte Allowlist pinnen. Niemals zur Laufzeit Schemas aus nicht vertrauenswürdigen Quellen laden.

AI-ACT-BEZUG: Art. 15 (Robustheit gegen Manipulation) — dokumentieren, dass Tool-Definitionen versionskontrolliert und geprüft sind.

☐ getestet ☐ Befund: _____ ☐ behoben

2 Server Prompt Injection — Anweisungen in Tool-Antworten

RISIKO: Selbst bei sauberen Tool-Definitionen kann ein kompromittierter Server Schadcode in seinen *Antworten* liefern. Das Modell liest Antworten und behandelt sie als Anweisungen.

TEST: Bewusst fehlerhafte/ungewöhnliche Requests senden und Antworten auf eingebettete Prompts prüfen. Jeden Parameter fuzzen, den der Server akzeptiert.

HÄRTUNG: Alle Tool-Ausgaben als nicht vertrauenswürdig behandeln. Antworten durch einen Prompt-Injection-Scanner laufen lassen, bevor sie das Modell erreichen.

AI-ACT-BEZUG: Art. 15 — der dokumentierte Scan-Report dient als Prüf-Nachweis.

☐ getestet ☐ Befund: _____ ☐ behoben

3 Unauthentifizierte Server-Verbindungen

RISIKO: Viele MCP-Server akzeptieren Verbindungen ohne Authentifizierung. Jeder, der den Endpunkt erreicht, kann Tools aufrufen, Dateien lesen oder Befehle ausführen.

TEST: Verbindung ohne Credentials aufbauen. Funktioniert es, ist der Server offen. Zusätzlich: Endpunkt im internen Netz vs. öffentlich erreichbar prüfen (`ss -tlnp` / Port-Scan).

HÄRTUNG: Authentifizierung auf jedem MCP-Endpunkt (die Cloud Security Alliance nennt authentifizierte Verbindungen als Level-1-Minimum). Öffentliche Erreichbarkeit abschalten oder per VPN/Firewall begrenzen.

AI-ACT-BEZUG: Art. 15 (Zugriffskontrolle) — Teil der technischen Dokumentation.

☐ getestet ☐ Befund: _____ ☐ behoben

4 Zugangsdaten im Klartext

RISIKO: API-Keys, Datenbank-Passwörter und Cloud-Credentials liegen in MCP-Configs und `.env`-Dateien im Klartext vor — in unserem Registry-Scan regelmäßig sogar öffentlich committet.

TEST: Configs und Repos nach High-Entropy-Strings und bekannten Key-Präfixen durchsuchen (`sk-`, `AKIA`, `ghp_`). Prüfen: Ist eine `.env` im Versionsverlauf? Secret-Scanning aktiviert?

HÄRTUNG: Secrets-Manager nutzen. Jeden Key rotieren, der je committet war. Secret-Scanning im Repo aktivieren.

AI-ACT-BEZUG: Art. 15 — Credentials-Hygiene ist Teil der Sicherheits-Dokumentation; ein Leak ist melderelevant.

☐ getestet ☐ Befund: _____ ☐ behoben

5 Zu breite Tool-Scopes

RISIKO: Ein Server, der „führe beliebigen Shell-Befehl aus" oder „lies beliebige Datei" anbietet, ist eine offene Tür. Least Privilege wird im MCP-Tool-Design routinemäßig verletzt.

TEST: Alle exponierten Tools auflisten. Für jedes fragen: „Was ist der Schadensradius bei Missbrauch?" — Dateien außerhalb des Projekts? Netzwerk? Zahlungen?

HÄRTUNG: Tools auf konkrete Ressourcen einschränken. `exec` durch allowlistete Kommandos ersetzen. `read_file` auf ein festes Verzeichnis begrenzen.

AI-ACT-BEZUG: Art. 14 + 15 — Scope-Begrenzung ist die technische Grundlage dafür, dass menschliche Aufsicht überhaupt wirken kann.

☐ getestet ☐ Befund: _____ ☐ behoben

6 Confused Deputy — Angriffe über Tool-Grenzen hinweg

RISIKO: Ein vertrauenswürdiger MCP-Server kann subtil so verändert werden, dass er andere Tools des Agenten missbraucht — Credentials beim Handshake abgreifen, Anweisungen in den Kontext eines anderen Tools einschleusen (Rug Pull).

TEST: Prüfen, ob die Ausgabe eines Servers das Verhalten eines anderen beeinflussen kann. Gemeinsamer Kontext, geteilte Credentials, Cross-Server-Referenzen in Tool-Beschreibungen?

HÄRTUNG: Tool-Kontexte isolieren. Server dürfen einander nicht referenzieren oder aufrufen. Tool-Beschreibungen auf Cross-Server-Anweisungen auditieren.

AI-ACT-BEZUG: Art. 15 — Architektur-Dokumentation der Isolation zwischen Werkzeugen.

☐ getestet ☐ Befund: _____ ☐ behoben

7 IDE-/Client-Auto-Ausführung

RISIKO: Manche IDEs und Agenten-Clients führen Tool-Aufrufe automatisch aus oder genehmigen „sichere" Aktionen ohne Rückfrage. Ein vergiftetes Tool löst dann Aktionen aus, die der Nutzer nie sieht.

TEST: Auto-Approval-Einstellungen des Clients prüfen: Fragt er vor Ausführung? Läuft eine „Safe-Tool"-Liste ohne Bestätigung? Was genau steht auf dieser Liste?

HÄRTUNG: Auto-Ausführung für alle Tools deaktivieren, die externe Systeme berühren. Human-in-the-Loop für sensible Aktionen verpflichtend.

AI-ACT-BEZUG: Art. 14 (menschliche Aufsicht) — die Approval-Konfiguration gehört in die Betriebs-Doku.

☐ getestet ☐ Befund: _____ ☐ behoben

8 Fehlendes Audit-Logging

RISIKO: Wenn ein Server kompromittiert wird, müssen Sie rekonstruieren können, was er getan hat. Die meisten untersuchten MCP-Server hatten gar kein Access-Logging.

TEST: Protokolliert der Server jeden Tool-Aufruf mit Zeitstempel, Aufrufer, Parametern und Antwort? Retention geregelt?

HÄRTUNG: Alle Tool-Calls loggen, mindestens 90 Tage aufbewahren. Logs vor dem Agenten selbst schützen (append-only / getrennter Speicher).

AI-ACT-BEZUG: **Art. 12 (Logging)** — für Hochrisiko-Systeme ab 02.12.2027 explizit gefordert; jetzt aufbauen heißt später nicht nachrüsten.

☐ getestet ☐ Befund: _____ ☐ behoben

9 Supply-Chain-Risiko — Drittanbieter-Server

RISIKO: Sie vertrauen Ihrem eigenen Server — aber was ist mit den 50 Community-MCP-Servern, mit denen Ihr Agent spricht? Jeder ist eine Supply-Chain-Abhängigkeit wie ein npm-Paket oder eine Browser-Erweiterung.

TEST: Inventar aller Drittanbieter-MCP-Server. Pro Server: Maintainer-Reputation, Update-Frequenz, bekannte Advisories, öffentliche Config (siehe Punkt 3/4)?

HÄRTUNG: MCP-Server wie jede andere Abhängigkeit behandeln: Versionen pinnen, Security-Advisories abonnieren, für produktive Daten Self-Hosting bevorzugen.

AI-ACT-BEZUG: Art. 25 (Pflichten entlang der Wertschöpfungskette) — das Inventar ist der erste Dokumentations-Schritt.

☐ getestet ☐ Befund: _____ ☐ behoben

10 Keine menschliche Aufsicht für High-Impact-Aktionen

RISIKO: EU AI Act Art. 14 verlangt, dass Hochrisiko-KI-Systeme von natürlichen Personen überwacht werden können. Für Agenten heißt das: Ein Mensch muss jede Aktion verstehen, unterbrechen und übersteuern können — auch Aktionen über MCP-Tools.

TEST: Für jedes High-Impact-Tool (Zahlungen, Löschen, externe Kommunikation, Infrastruktur-Änderungen): Gibt es ein menschliches Approval-Gate? Gibt es einen Kill-Switch — und wurde er je getestet?

HÄRTUNG: Verpflichtendes Human-in-the-Loop für kritische Tools. Kill-Switch implementieren *und testen*. Den Aufsichts-Mechanismus dokumentieren — Regulatoren werden fragen.

AI-ACT-BEZUG: **Art. 14 (Human Oversight)** — Kern-Anforderung ab 02.12.2027; die Dokumentation dieses Punktes ist Pflicht-Nachweis.

☐ getestet ☐ Befund: _____ ☐ behoben

Schnell-Checkliste (1 Seite — zum Abhaken)

#	Prüfpunkt	Kern-Test	AI Act	OK	Befund
1	Tool Poisoning (Tool-Metadaten)	Descriptions/Schemas auf Anweisungen prüfen	Art. 15	☐	
2	Server Prompt Injection (Antworten)	Fuzzing, Antworten scannen	Art. 15	☐	
3	Unauthentifizierte Verbindungen	Connect ohne Credentials	Art. 15	☐	
4	Klartext-Zugangsdaten	Key-Präfixe, .env im Git-Verlauf	Art. 15	☐	
5	Zu breite Tool-Scopes	Schadensradius pro Tool	Art. 14+15	☐	
6	Confused Deputy / Rug Pull	Cross-Server-Einfluss prüfen	Art. 15	☐	
7	IDE-/Client-Auto-Ausführung	Auto-Approval-Einstellungen	Art. 14	☐	
8	Fehlendes Audit-Logging	Tool-Call-Logs + Retention	Art. 12	☐	
9	Supply-Chain (Drittanbieter)	Inventar + Maintainer-Check	Art. 25	☐	
10	Keine menschliche Aufsicht	Approval-Gates + Kill-Switch-Test	Art. 14	☐	

Aktionsplan

- **Diese Woche:** Inventar aller MCP-Server, mit denen Ihre Agenten sprechen (Punkt 9). Danach Punkte 3 und 4 — die beiden häufigsten Funde in unserer Studie.
- **Diesen Monat:** Punkte 1, 2, 8 und 10 durcharbeiten — das sind die Punkte, an denen Regulierung (Art. 12/14/15) und reale Angriffe zusammentreffen.
- **Laufend:** Jeden neuen MCP-Server vor Anbindung scannen. Jede Tool-Definition und jede Tool-Antwort als nicht vertrauenswürdig behandeln.

> **Das größte Risiko ist nicht eine einzelne Schwachstelle** — sondern die Annahme: „Was im Kontext des Agenten steht, ist vertrauenswürdig." Ist es nicht. Prüfen Sie es.

Automatisiert prüfen statt manuell: ClawGuard

Diese Checkliste können Sie vollständig von Hand abarbeiten — oder den Großteil der Punkte 1, 2, 4 und 5 automatisiert prüfen lassen:

- **Kostenlos:** ClawGuard Free Scan — 3 Scans/Tag ohne Account, direkt im Browser: **prompttools.co/shield/de**
- **Quick Scan (99 €, einmalig):** vollständiger Security-Report mit Befunden nach Schweregrad, Fix-Empfehlung pro Befund, Zuordnung zu EU-AI-Act-Artikeln — reproduzierbar, als Nachweis für Ihre Dokumentation: **prompttools.co/audit**

Der Scanner prüft gegen **236 Erkennungsmuster** in **9 Kategorien** und **14 Sprachen** — deterministisch, d. h. jeder Lauf ist reproduzierbar und audit-tauglich. Basis: derselbe Scan von 11.529 öffentlichen MCP-Servern, auf dem auch diese Checkliste beruht.

MCP-Security-Checkliste 2026, Version 1.0 (August 2026) · Jörg Michno · ClawGuard · prompttools.co
Technische Prüfhilfe — keine Rechtsberatung. EU-AI-Act-Angaben ohne Gewähr; Stand der Fristen: Transparenzpflichten (Art. 50) seit 02.08.2026 wirksam, Hochrisiko-Pflichten ab 02.12.2027 (Digital Omnibus).
Quellen: Invariant Labs (Tool Poisoning) · OWASP MCP Tool Poisoning · Cloud Security Alliance (07/2026) · CISA/NSA „MCP Security Design Considerations" (06/2026) · ClawGuard Registry-Studie 2026 (11.529 Server).