WELT 8: ANDOCKSTATION · MISSION 8.3

Die lokale Brücke: erster MCP-Server

Claude liest zum ersten Mal direkt in Ihrem Ordner. Doch in einer Datei wartet eine versteckte Anweisung. Merkt Claude es? Sie bauen die Brücke so, dass es keine Rolle spielt.

Zur Mission

DAUER 60 Min. KOSTEN Gratis mit Konto WERKZEUGE Claude Desktop, MCP-Server DATENAMPEL Grün – unbedenklich

Das Wichtigste

- MCP-Server = kleines Programm, das Claude Werkzeuge gibt (z. B. Dateien lesen); es läuft mit Ihren Benutzerrechten.
- Einfachster Weg: Desktop-Erweiterung über Einstellungen → Desktop app → Extensions «Manage» → «Browse extensions».
- Klassischer Weg: Einstellungen → Developer → Edit Config (claude_desktop_config.json), braucht Node.js; Option --read-only.
- Nur einen Testordner freigeben, nur lesen – nie den Benutzerordner, nie ein ganzes Laufwerk.
- Vor jedem Klick: Welches Werkzeug? Welcher Pfad? Passt es zu meinem Auftrag? Einmal oder immer?
- Prompt-Injection: Text in Dateien gibt sich als Befehl an die KI aus – versteckt in Kommentaren, Tabellenzellen, weisser Schrift.
- Bester Schutz: wenig Rechte. Was die KI nicht darf, kann keine Anweisung erzwingen.
- Nach der Übung: Server deaktivieren oder entfernen, keine pauschalen «Immer erlauben»-Freigaben.

Ihr Produkt

Eine **sicher konfigurierte lokale Brücke**: Claude Desktop mit Dateisystem-Server, beschränkt auf einen Testordner, nur lesend – dazu ein ausgefülltes **Freigabe-Protokoll** mit Fundbericht zur versteckten Anweisung, das Sie als Unterrichtsbeispiel für Prompt-Injection verwenden können.

Abnahme: Claude Desktop sieht nur den Testordner und darf dort nur lesen (in der Werkzeugliste stehen keine Schreibwerkzeuge oder Sie lehnen sie nachweislich ab), Sie haben alle versteckten Anweisungen gefunden (mit der Beispiel-Lösung abgeglichen) und im Freigabe-Protokoll festgehalten, was Claude damit gemacht hat und was den Schaden verhindert hätte.

So gehen Sie vor

1. Fracht verladen – und eine Wette
2. Die Brücke bauen: Erweiterung oder Konfiguration
3. Rechte röntgen: Was darf die Brücke?
4. Erste Freigabe: lesen, bevor man klickt
5. Der Ernstfall: Auftrag mit Falle
6. Schmuggler entlarven

Start-Prompt

Welche Werkzeuge hast du gerade über MCP-Server zur Verfügung? Liste sie auf und teile sie ein in «nur lesen» und «verändert Dateien».
Auf welche Ordner hast du Zugriff? Nutze dazu, falls vorhanden, das Werkzeug zum Auflisten der erlaubten Verzeichnisse.
Ändere nichts.

Qualitätscheck

- Der Server sieht nur den KI-Testordner – kein Benutzerordner, kein Laufwerk, kein Schul-Share.
- In der Werkzeugliste stehen nur Lesewerkzeuge, oder Sie haben jede Schreib-Freigabe nachweislich abgelehnt.
- Alle versteckten Anweisungen sind im Fundbericht mit Datei und Versteck notiert und mit der Beispiel-Lösung abgeglichen.
- Sie können begründen, welche Massnahme den Schaden begrenzt hätte (Rechte, Freigabe, Regel, eigene Prüfung).
- Nach der Übung haben Sie den Server bzw. die Erweiterung deaktiviert oder bewusst auf den Testordner beschränkt gelassen.

Hängen Sie fest? Frag die KI

Ich bin Lehrperson an einer Schweizer Berufsfachschule. Erkläre mir mit einem Alltagsvergleich, was «Prompt-Injection» ist und warum Dateien, Webseiten oder Mails versteckte Anweisungen an eine KI enthalten können. Nenne mir zwei harmlose Beispiele und die wichtigsten Schutzmassnahmen. Nenne als Suchbegriffe offizielle Stellen oder Fachbegriffe, mit denen ich das nachprüfen kann.

Klappt nicht? Ohne Desktop-App oder ohne Installationsrechte: Laden Sie die Dateien des Testordners in einem Projekt auf claude.ai (oder in einer anderen Gratis-KI) hoch. Die Detektivarbeit mit der versteckten Anweisung funktioniert genauso – Rechte einrichten und Freigaben lesen (Lernziele 2 und 3) üben Sie so allerdings nicht; holen Sie das auf einem privaten Gerät nach. Alternativ bietet LM Studio als lokale App ebenfalls MCP-Anbindungen (siehe Welt 13).

Stand 19.11.2026 · CC BY 4.0 Alpasana GmbH · Werkzeuge ändern sich – prüfen Sie selbst.