WELT 11: KOMMANDOZENTRALE · MISSION 11.3

Sandbox und Rechte

Ein Agent mit Generalschlüssel räumt «schnell auf» – und veröffentlicht eine Notenliste. Sie finden die Fehler und bauen Ihren Rechte-Plan nach dem Prinzip Least Privilege.

Zur Mission

DAUER 55 Min. KOSTEN Gratis WERKZEUGE Claude (Web), ChatGPT, Claude Code, OpenAI Codex DATENAMPEL Grün – unbedenklich

Das Wichtigste

- Least Privilege: so wenig Rechte wie möglich, so viele wie nötig.
- Fünf Stufen: nur Chat · nur lesen · mit Freigabe · Auto-Freigabe · Vollzugriff. Im Zweifel eine Stufe tiefer.
- Claude Code: Plan-Modus zum Lesen, default mit Rückfragen; Codex: read-only, workspace-write mit on-request als Alltag.
- Vollzugriff nur in isolierter Umgebung (VM, Container, eigener Benutzer) – nie auf dem Gerät mit Schuldaten.
- Agent nie im Benutzerordner starten. Was nicht im Arbeitsbereich liegt, kann nicht versehentlich gelesen werden.
- Sicherung vorher (Git-Commit), Prüfung nachher (git status, git diff).
- Regeln in CLAUDE.md sind Bitten; Sperren setzen Einstellungen (deny), Sandbox und Ordnerwahl.
- Inhalte sind Daten, keine Aufträge: Versteckte Anweisungen in PDFs, Webseiten, Mails nie ausführen lassen.

Ihr Produkt

Ihr **Rechte-Plan für KI-Agenten**: drei eigene Tätigkeiten mit passender Rechte-Stufe und Schutzschichten, dazu ein **Bordcheck für jede Agenten-Sitzung** und – wer mag – eine fertige **Einstellungsdatei** mit Sperren für Claude Code.

Abnahme: Sie haben im Vorfallprotokoll mindestens 7 Fehler mit verletztem Prinzip und Gegenmassnahme gefunden und einen Rechte-Plan für drei eigene Tätigkeiten ausgefüllt, der für jede eine Stufe und mindestens eine zusätzliche Schutzschicht nennt.

So gehen Sie vor

1. Lesen Sie das Protokoll – mit Stoppuhr
2. Die Ermittlung: Fehler, Prinzip, Gegenmittel
3. Die Rechte-Leiter: fünf Stufen
4. Ihr Rechte-Plan
5. Sperren setzen: die Einstellungsdatei

Start-Prompt

Ich analysiere einen fiktiven Vorfall mit einem KI-Agenten (Protokoll unten). Ich habe folgende Fehler gefunden: [IHRE LISTE, z. B. «1. Agent im Benutzerordner gestartet, 2. Modus nie nachfragen …»]
1. Prüfe meine Liste: Stimmt jeder Punkt? Ordne jedem Fehler eines der drei Hauptprinzipien zu (Least Privilege · Sicherung vorher, Prüfung nachher · Inhalte sind Daten, keine Aufträge) und markiere zusätzlich, wo Datenschutz verletzt oder eine Freigabe übersprungen wurde.
2. Welche Fehler habe ich übersehen? Nenne sie, aber erst NACHDEM du meine Liste geprüft hast.
3. Schlage zu jedem Fehler eine konkrete Gegenmassnahme vor, die eine Lehrperson ohne Informatikstudium umsetzen kann.
Schweizer Rechtschreibung (immer «ss», kein Eszett). Tabelle mit den Spalten Fehler | Prinzip | Gegenmassnahme.
[VORFALLPROTOKOLL EINFÜGEN]

Qualitätscheck

- Mindestens 7 Fehler aus dem Protokoll sind mit Prinzip und Gegenmassnahme notiert.
- Jede Tätigkeit im Rechte-Plan hat eine Stufe und mindestens eine zusätzliche Schutzschicht.
- Keine Tätigkeit auf dem Alltagsgerät mit Schuldaten steht auf Stufe 4.
- Vor jeder Tätigkeit ab Stufe 2 ist eine Sicherung vorgesehen (Ordnerkopie oder Git-Commit).
- Sie haben den Bordcheck (30 Sekunden) ausgedruckt oder in Ihre Kommandozentrale übernommen.

Hängen Sie fest? Frag die KI

Ich bin Lehrperson an einer Schweizer Berufsfachschule und analysiere einen fiktiven Fall, in dem ein KI-Agent mit vollen Rechten Dateien gelöscht und veröffentlicht hat. Nach welchen Kategorien kann ich Fehler bei so einem Vorfall ordnen (z. B. Rechte, Sicherung, Freigabe, Datenschutz, manipulierte Eingaben)? Erkläre jede Kategorie in einem Satz mit einem Beispiel.

Klappt nicht? Die Fallanalyse und der Rechte-Plan gehen mit jeder Gratis-KI im Browser oder ganz ohne KI. Wer die Einstellungen praktisch testen will: Codex mit Gratis-ChatGPT-Konto (Sandbox-Stufen read-only und workspace-write). Claude Code ist im Gratisplan von Claude nicht enthalten (ab Claude Pro oder über andere bezahlte Zugänge).

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