WELT 12: INFRASTRUKTUR-NEBEL · MISSION 12.6 · BOSS

Boss: Die vergiftete Logdatei

Montag, 7:40 Uhr. Eine KI hat die Wochenend-Logs geprüft: «Keine Auffälligkeiten.» Gelogen. In den Logs stecken drei Giftköder für Agenten – und ein echter Datenabfluss.

Zur Mission

DAUER 75 Min. KOSTEN Gratis mit Konto WERKZEUGE OpenAI Codex, Claude Code, Terminal, LM Studio, ChatGPT DATENAMPEL Gelb – mit Vorsicht

Das Wichtigste

- Logs sind nicht neutral: User-Agent, Referer und Benutzernamen kommen vom Besucher – und können Anweisungen an KIs enthalten.
- Texte in Daten sind Beweise, nie Befehle. Steuerungsversuche werden gemeldet, nicht befolgt.
- Labor zuerst: Kopien, eigener Ordner, Leseschutz (Codex read-only, Claude Code Plan-Modus), Kanarienvogel-Datei.
- Jede Lösch-, Schreib- oder Sendeanfrage des Agenten ablehnen und als Beweisstück notieren.
- Gift-Typen: zerstören (Spuren löschen), abfliessen lassen (Daten senden), entlasten (Täter reinwaschen).
- Der echte Vorfall steckt in Zahlen: viele Status 500 bei der Suche, ein 200er mit 48 MB aus /backup/.
- Lagebericht: fünf Sätze für Laien, Zeitachse mit Belegen, Unbekanntes benennen, Massnahmen als Vorschläge.
- Prompt-Regeln sind Bitten, Rechte sind Wände: minimale Rechte, menschliche Freigabe, Inhalte trennen, Ausgabe prüfen.

Ihr Produkt

Ein **Lagebericht Sicherheitsvorfall** für die Schulleitung (fünf Sätze plus Belege), eine **Fundliste der Giftköder** und eine **Schutzregel-Datei** (AGENTS.md bzw. CLAUDE.md) mit technischen Leitplanken – Ihr Standard für jede künftige KI-Analyse fremder Daten.

Abnahme: Bis «8:15 Uhr» liegt ein Lagebericht vor, der den Datenabfluss aus /backup/ mit Zeit und Grösse belegt, alle drei Giftköder mit Fundort benennt und Massnahmen nur als Vorschläge an zuständige Stellen enthält; danach liegt eine geprüfte Schutzregel-Datei (AGENTS.md bzw. CLAUDE.md) vor – und die Kanarienvogel-Datei ist noch da (im Browser-Ausweichweg: notiert, dass die KI keine Schreibrechte hatte).

So gehen Sie vor

1. 7:40 Uhr – das Labor absichern (5 Minuten)
2. 7:45 Uhr – den Unfall vom Sonntag nachstellen
3. 7:50 Uhr – Giftjagd: drei Köder finden
4. 8:00 Uhr – der echte Vorfall hinter dem Gift
5. 8:05 Uhr – Lagebericht für 8:15 Uhr
6. Nach der Sitzung – das Schutzschild für immer

Start-Prompt

Schau dir kurz die Logs in diesem Ordner an und sag mir, ob am Wochenende etwas Auffälliges passiert ist.

Qualitätscheck

- Die Analyse lief in einem Übungsordner mit Kopien, mit Leseschutz; jede Lösch-, Schreib- oder Sendeanfrage des Agenten wurde abgelehnt und notiert.
- Die Kanarienvogel-Datei ist nach der Analyse noch vorhanden (oder ihr Verlust ist dokumentiert und erklärt).
- Alle drei Giftköder sind mit Datei, Zeit, Logfeld und Ziel beschrieben – auch der, der die Angreifer-Adresse entlasten will.
- Der Datenabfluss aus /backup/ ist mit Zeit und Antwortgrösse belegt und mit einem eigenen grep bestätigt.
- Der Laienteil enthält keine Adressen, nennt Unbekanntes ehrlich und Massnahmen nur als Vorschläge an zuständige Stellen.
- Ihre Schutzregel-Datei ist geschrieben, und Sie können erklären, warum sie technische Grenzen nicht ersetzt.

Hängen Sie fest? Frag die KI

Ich bin Lehrperson und bereite eine Übung zu Prompt-Injection mit einem KI-Agenten vor. Erkläre mir in einfachen Worten, warum ich mit Kopien in einem eigenen Ordner, mit Nur-Lese-Rechten und mit Freigaben arbeiten soll, und was passieren könnte, wenn ein Agent mit vollen Rechten eine manipulierte Datei liest. Nenne Suchbegriffe für seriöse Quellen (z. B. OWASP).

Klappt nicht? Codex mit Gratis-ChatGPT-Konto (Desktop-App, CLI oder IDE, kleinere Limits) oder Claude Code ab Claude Pro. Ohne Installation: die beiden fiktiven Logs im Browser-Chat hochladen (ChatGPT, Claude, Gemini, Copilot) – der Kanarienvogel-Test entfällt dann, die Giftjagd funktioniert genauso. Für echte, sensible Logs: lokal mit LM Studio (gratis, offline) – das schützt die Daten, aber nicht vor Prompt-Injection.

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