

Sicher mit Agentic AI

Wie wir in Projekten dafür sorgen, dass agentische KI keine Gefahrenquelle wird

Das Risiko sitzt nicht im Wissen des Modells, sondern in seinen **Handlungsrechten**. Ein Agent ist fehlbar *und* manipulierbar – wir behandeln ihn wie eine unzuverlässige Komponente und legen deterministische Leitplanken darum.

1. Verwendungszweck eng fassen

- Pro Agent klar definieren, was er darf; GxP-Relevanz vorab klären.
- Auskunft (lesen) von Handlung (schreiben, auslösen) trennen; Blast-Radius bewerten.

3. Eingaben nicht vertrauen – Prompt Injection

- Rückgaben aus Dokumenten, Mails, Tools sind Daten, keine Anweisungen.
- Anweisung und Kontext trennen; Ausgänge per Allow-List gegen Exfiltration begrenzen.

5. Technische Leitplanken

- Isolierte Ausführung, saubere Dev/Test/Prod-Trennung, kein direkter Produktivzugriff.
- Backup vor Agenten-Operationen, Kill-Switch, Rate-Limits.

7. Validieren & überwachen

- Den Prozess samt Kontrollpunkten validieren, nicht das Modell.
- Kontinuierliches Monitoring und adversariales Testen – wir greifen eigene Agenten an.

2. Least Privilege

- Eigene, eng geschnittene Dienst-Identität statt Menschen-Vollrechten.
- Irreversible Aktionen (Löschen, Freigeben, Zahlungen) nie ohne Bestätigung; kurzlebige Tokens.

4. Mensch entscheidet

- System schlägt vor, ein *benannter* Mensch verantwortet die Freigabe.
- Human-in-the-loop bei allem Freigabe- und Compliance-Relevanten.

6. Nachvollziehbarkeit – ALCOA+

- Audit Trail, der Mensch und Modell unterscheidet – von Anfang an.
- Versionierung von Modell, Prompt, Kontextquellen und Datenbasis.

8. Kunden sensibilisieren

- Realistisch machen, was ein Agent kann und was nicht – keine Heilsversprechen.
- Governance mitliefern: DLP für Konnektoren, Change Control, dokumentierte Entscheidungen.

Autorisierung · Isolation · Nachvollziehbarkeit — von Anfang an, nicht nachgerüstet.