Der Vollmachts-Check

Begleitpaket zum Buch „Wenn KI handelt"

von Thorsten Behder

Die Leitmetapher der Vollmacht aus dem Buch als Checkliste, um Kontrolle als Architektur für den eigenen Umgang mit handelnder KI zu prüfen.

CHECKLISTE

Drei Checklisten aus dem Buch, wörtlich übernommen. Sie klären, wie viel Vollmacht ein Agent bekommt, welche Rechte er dafür braucht und was mitgeschrieben werden muss.

Die fünf Autonomiestufen

Aus dem Buch — Kapitel „Vom Chat zur Vollmacht"

- ☐ **Stufe 1 — Operator:** — System schlägt vor, Mensch entscheidet und führt aus
- ☐ **Stufe 2 — Collaborator:** — System handelt, Mensch arbeitet aktiv mit und kann Veto einlegen
- ☐ **Stufe 3 — Consultant:** — System arbeitet selbständig, fragt nur bei Ausnahmen
- ☐ **Stufe 4 — Approver:** — System arbeitet autonom, Mensch genehmigt Ergebnisse
- ☐ **Stufe 5 — Observer:** — System arbeitet vollständig autonom, Mensch beobachtet

Werkzeugrechte vergeben

Aus dem Buch — Kapitel „Hände, Augen, Werkzeuge"

- ☐ **Read-only zürst:** — Welche Werkzeuge können auf reinen Lesezugriff beschränkt werden? Beginnen Sie dort.
- ☐ **Schreibrechte differenzieren:** — Welche Schreiboperationen sind unkritisch (Termineintrag), welche kritisch (Finanztransaktion)? Definieren Sie Freigabestufen.
- ☐ **Audit-Spur einrichten:** — Wird jeder Werkzeugaufruf protokolliert, mit Zeitstempel, Parametern und Ergebnis?
- ☐ **Not-Aus definieren:** — Gibt es einen Mechanismus, der alle Agentenaktivitäten sofort unterbricht, wenn ein Fehler erkannt wird?
- ☐ **Sandbox prüfen:** — Läuft Codeausführung in einer isolierten Umgebung, die keinen ungeplanten Netzwerk- oder Dateizugriff erlaubt?
- ☐ **GUI-Isolation sicherstellen:** — Arbeitet ein CUA nur in dedizierten Fenstern oder virtuellen Desktops, nicht auf Bildschirmen mit parallelen vertraulichen Inhalten?
- ☐ **Connector-Konfiguration dokumentieren:** — Sind die Berechtigungen jedes MCP-Connectors explizit festgelegt und regelmässig überprüft?

Kontrollarchitektur-Audit: zehn Fragen vor dem Start

Aus dem Buch — Kapitel „Kontrolle ist Architektur"

- ☐ **1. Vollmacht** — Welche Rechte hat der Agent? Sind sie auf das für die Aufgabe erforderliche Minimum beschränkt?
- ☐ **2. Freigabestufen** — An welchen Schwellen muss ein Mensch entscheiden? Sind die Schwellen nach Irreversibilität, Schadenshöhe und regulatorischer Relevanz differenziert?
- ☐ **3. Informationsschnitt** — Erhält der Mensch an der Schwelle die richtigen Informationen, um eine qualifizierte Entscheidung zu treffen?
- ☐ **4. Sandboxing** — Ist der Agent in einer isolierten Umgebung eingebettet? Kann er das Produktivsystem direkt verändern, oder nur über kontrollierte Schnittstellen?
- ☐ **5. Audit-Spur** — Wird jede Aktion protokolliert — mit Zeitstempel, Modellversion, Entscheidungsgrundlage und Freigabestatus?
- ☐ **6. Budgetgrenzen** — Hat der Agent ein maximales Transaktionsvolumen pro Zeiteinheit? Was passiert bei Überschreitung?
- ☐ **7. Not-Aus** — Gibt es einen definierten und getesteten Abschaltmechanismus? Wer darf ihn auslösen? Wie schnell wirkt er?
- ☐ **8. Anomalieerkennung** — Gibt es automatische Monitoring-Systeme, die ungewöhnliche Muster erkennen und den Agenten präventiv pausieren können?
- ☐ **9. Rollen** — Sind Agent-Owner, Prozess-Owner, Compliance und Revision klar benannt? Wissen alle Beteiligten, wer im Ernstfall entscheidet?
- ☐ **10. Regelmässige Prüfung** — Ist ein Rhythmus für die Überprüfung der Kontrollarchitektur definiert? Kontrollmechanismen, die nicht regelmässig geprüft werden, veralten.

> Warum Agenten scheitern, wie Prompt Injection funktioniert und wie eine belastbare Audit-Spur aussieht, erklärt das Buch.

Dieses Paket gehört zum Buch „Wenn KI handelt”

https://www.amazon.de/dp/B0GZCJ6JC3

tbai.cloud/buch-wenn-ki-handelt.html

Autor und inhaltlich verantwortlich

Thorsten Behder

Weilheimer Str. 21A

81373 München

thorsten@behder.de

Gestaltung und Layout: tbai.cloud