

AGENTIC AI PRAXISLEHRBUCH

KI-Agenten im Unternehmen einführen:
Sichere und messbare agentische Workflows



THORSTEN BEHDER

Die Governance-by-Design- Checkliste

Begleitpaket zum Buch „Agentic AI Praxislehrbuch“

von Thorsten Behder

Die Governance-Bausteine aus dem Buch – Autonomiestufen-Modell, Risiko-Scoring und die wichtigsten Prüfpunkte zu EU AI Act, DSGVO, NIST AI RMF und OWASP LLM Top 10 – als Checkliste für den nächsten Piloten.

CHECKLISTE

Elf Vorlagen aus dem Prompt-Anhang des Buchs, wörtlich übernommen – ein Querschnitt durch alle sechs Cluster von der Diagnose bis zur Skalierung.

Diagnose und Plattformdesign

Aus dem Buch — Prompt-Anhang, Cluster 1 und 2

1. Agentische Eignung vorab prüfen

Du bist ein nüchterner Berater für Automatisierung.
Prüfe die folgende Aufgabe entlang des Entscheidungsbaums aus Kapitel 01: Braucht sie wirklich einen Agenten, oder reicht ein festes Skript, ein RPA-Bot oder ein einfacher Chat-Assistent?

Aufgabe: [BESCHREIBUNG]
Wie variabel sind die Eingaben: [ANGABE]
Wie viele Entscheidungen pro Durchlauf: [ANGABE]

Gib eine Empfehlung in vier Stufen zurück: kein KI-Bedarf, einfacher Assistent, fester Workflow, agentischer Workflow.
Begründe in drei Sätzen und nenne das grösste Risiko bei Überautomatisierung.

2. Permission-Matrix ableiten

Du leitest eine Permission-Matrix für eine Agentenplattform ab (Kapitel 07 und 08).
Ordne jeder geplanten Aktion Datenbereich, Tool, Scope, nötige Freigabe und Reviewer zu.

Geplante Aktionen: [AKTION 1; AKTION 2; ...]
Datenbereiche: [ANGABE]
Rollen / Reviewer: [ANGABE]

Gib eine Tabelle mit Aktion, Datenbereich, Tool, Scope (read/write), Approval (keine/Mensch/Vier-Augen) und Reviewer zurück. Setze für jede verändernde Aktion mit sensiblen Daten mindestens Mensch-Freigabe und markiere fehlende Reviewer als offen.

3. Einzelnes Tool sauber definieren

Du schreibst eine praezise Tool-Definition für einen Agenten (Kapitel 08).

Formuliere für das folgende Werkzeug eine eindeutige Beschreibung, die ein Modell zuverlaessig richtig aufruft.

Werkzeug / Zweck: [BESCHREIBUNG]

Eingaben: [PARAMETER]

Erwartete Ausgabe: [ANGABE]

Gib zurueck: Name, einsatzklare Beschreibung (wann nutzen, wann nicht), Parameter mit Typ und Pflicht/optional, Nebenwirkungsklasse (read-only/draft/side-effecting), Fehlerverhalten und ein Negativbeispiel ("nicht aufrufen, wenn ...").

Workflow-Design

Aus dem Buch — Prompt-Anhang, Cluster 3

4. Workflow in vier Bausteine zerlegen

Du bist ein erfahrener Architekt für agentische KI-Workflows. Analysiere den folgenden Geschaeftsprozess und zerlege ihn in die vier Bausteine Modell, Retrieval, Memory und Tool-Use.

Prozess: [PROZESS HIER BESCHREIBEN]

Beteiligte Systeme: [SYSTEME]

Sensible Daten: [JA/NEIN, welche Kategorien]

Gib eine Tabelle mit den Spalten Baustein, Aufgabe, Risiko und Erste Massnahme zurueck. Markiere für Tool-Use jede Aktion als "nur lesend" oder "veraendernd, Freigabe noetig". Schlage am Ende eine geeignete Autonomiestufe vor und begruende sie in einem Satz.

5. AGENTS.md generieren

Du erzeugst eine AGENTS.md für OpenAI Codex (Kapitel 06).
Verbindliche Team-Vorgaben gehören in diese eingetragte Datei,
Memories dienen nur als Recall-Layer.

Projekt / Repo: [BESCHREIBUNG]
Verbindliche Regeln: [REGELN]
Sandbox- und Approval-Präferenzen: [ANGABE]
Was der Agent NICHT tun darf: [VERBOTE]

Gib die AGENTS.md mit Abschnitten Projektkontext, verbindliche
Regeln, Approval-Policy, erlaubte und verbotene Aktionen zurück.
Schreibe keine geheimen Werte hinein, nur Verweise auf den sicheren
Ablageort.

Governance

Aus dem Buch — Prompt-Anhang, Cluster 4

6. Autonomiestufe bestimmen

Du bestimmst die Autonomiestufe für einen Agenten nach dem
Autonomiestufen-Modell aus Kapitel 13.

Use Case und geplante Aktionen: [BESCHREIBUNG]
Risiko-Score (falls vorhanden): [ANGABE]
Reversibilität der Aktionen: [ANGABE]

Ordne den Use Case einer Stufe zu (vorschlagend, ausführend mit
Freigabe, ausführend mit Stichprobe, autonom in engen Grenzen) und
begründe. Beschreibe das passende Human-in-the-Loop-Design und
nenne die Bedingung, unter der die nächsthöhere Stufe später
freigegeben werden darf.

7. Prompt-Injection-Abwehr entwerfen

Du bist ein Security-Engineer für agentische Systeme (Kapitel 12).
Entwirf eine Abwehr gegen Prompt Injection für den folgenden Workflow, der nicht vertrauenswürdige Eingaben verarbeitet.

Workflow: [BESCHREIBUNG]

Untrusted Inputs (woher kommen sie): [ANGABE]

Side-effecting Tools im Zugriff: [LISTE]

Gib zurück: Trennung von Daten und Instruktionen, Eingabe-Filter, Output-Validierung vor side-effecting Aktionen, Least-Privilege für Tools und eine Bestätigungspflicht vor jeder Aktion mit Aussenwirkung.
Nenne einen konkreten Angriffsfall und wie die Abwehr ihn stoppt.

8. Incident-Runbook und Kill-Switch entwerfen

Du entwirfst ein Incident-Runbook mit Kill-Switch für einen produktiven Agenten (Kapitel 12 und 13).

Agent / Workflow: [BESCHREIBUNG]

Schlimmster denkbarer Fehlerfall: [ANGABE]

Beteiligte Rollen: [ANGABE]

Gib zurück: Auslöser (woran erkennt man den Vorfall), Sofort-Stopp (wie wird der Agent abgeschaltet, wer darf das), Eingrenzung, Benachrichtungskette, Wiederanlaufbedingung und Lernschleife.
Mache den Kill-Switch so einfach, dass er unter Stress funktioniert.

Training, Pilot und Skalierung

Aus dem Buch — Prompt-Anhang, Cluster 5 und 6

9. Curriculum-Matrix über vier Rollen aufbauen

Du baust eine Curriculum-Matrix über die vier Leserrollen auf (Kapitel 14): Management, AI-Enabler, Builder, Compliance.

Themenfelder (z.B. Grundlagen, Governance, Tooling, Workflow-Design):
[LISTE]

Verfügbares Gesamtbudget pro Rolle: [ANGABE]

Gib eine Matrix zurück (Zeilen: Themenfelder, Spalten: vier Rollen), in jeder Zelle die Tiefe (Überblick/Anwendung/Vertiefung) und das Prüfprodukt. Markiere die gemeinsamen Pflichtinhalte für alle Rollen.

10. Pilot Charter erstellen

Du erstellst eine Pilot Charter nach Kapitel 18.

Fülle Scope, Nicht-im-Scope, KPIs, Risikoklasse, HIL-Design und Freigabeinstanzen aus.

Use Case: [BESCHREIBUNG]

Pilotdauer: [z.B. 90 Tage]

Beteiligte Rollen: [ANGABE]

Gib die Charter mit den Abschnitten Ziel, Scope, Nicht-im-Scope, KPIs (Prozesswirkung, nicht Aktivität), Risikoklasse, Human-in-the-Loop-Design, Freigabeinstanzen und Abbruchkriterium zurück.

11. Eval Card und Golden Set aufsetzen

Du setzt eine Eval Card mit Golden Set nach Kapitel 09 auf. Schlage Testfälle und Schwellen vor.

Workflow / Aufgabe: [BESCHREIBUNG]

Kritische Fehlerarten: [ANGABE]

Gib eine Eval Card mit Testfall-Kategorien, je 2 bis 3 Beispielfällen, Erfolgskriterium und Schwellen zurück. Setze die Schwelle für den Pilotbetrieb auf größer 92 Prozent und für den Produktivbetrieb auf größer 95 Prozent. Trenne Safety-Tests von Business-Tests.

Der vollständige Prompt-Anhang umfasst 85 Vorlagen über alle sechs Cluster; die zugehörigen Kapitel erklären, wann welche greift.

Dieses Paket gehört zum Buch „Agentic AI Praxislehrbuch“

<https://www.amazon.de/dp/BOH5YNPSL7>
tbai.cloud/buch-agentic-ai-praxislehrbuch.html

Autor und inhaltlich verantwortlich

Thorsten Behder
Weilheimer Str. 21A
81373 München
thorsten@behder.de



Gestaltung und Layout: tbai.cloud