
WHITEPAPER · AI ENABLEMENT & AGENTIC DELIVERY

AI Enablement und Agentic Architecture: vom Pilot zum Betrieb

HERAUSGEBER

Manuel Fernandez

VERÖFFENTLICHT

15. Oktober 2026

01 Executive Summary

Generative KI ist in vielen Unternehmen angekommen – als Assistent, Prototyp oder Experiment einzelner Teams. Der Schritt in den verantwortbaren Betrieb gelingt deutlich seltener. Das liegt selten am Sprachmodell, sondern meist an fehlender Architektur, unklaren Verantwortlichkeiten und Risiken, die erst im Pilot sichtbar werden.

Mit agentischen Workflows verschärft sich diese Lage: Sobald ein KI-System Werkzeuge aufruft, Daten schreibt oder Folgeschritte plant, wird aus einer Frage der Antwortqualität eine Frage der Berechtigungen, der Systemsicherheit und der Haftung. Dieses Whitepaper beschreibt ein herstellerneutrales Vorgehen, um Use Cases nach Nutzen, Aufwand und Risiko auszuwählen, eine passende Daten-, Modell- und Werkzeugarchitektur zu entwerfen, menschliche Freigaben gezielt zu setzen und Governance sowie Security von Beginn an zu verankern – mit einem begrenzten Pilot als Grundlage für die Skalierungsentscheidung.

KERNTHESE

Agentische KI wird nicht durch ein besseres Modell betriebsfähig, sondern durch eine Architektur, die Handlungsspielräume begrenzt, Entscheidungen nachvollziehbar macht und Menschen an den richtigen Stellen in der Verantwortung hält. Autonomie ist eine Designentscheidung, die schrittweise und begründet erweitert wird.

02 Problemstellung

Ein Prototyp, der Dokumente zusammenfasst oder Fragen zu einer Wissensbasis beantwortet, ist schnell gebaut. Gerade diese Leichtigkeit erzeugt eine Lücke zwischen Demonstration und Betrieb. Typische Muster sind:

- **Technologie vor Problem:** Use Cases folgen dem, was ein Modell kann, nicht einem Prozessschritt, der messbar besser werden soll.
- **Insellösungen:** Teams bauen eigene Assistenten mit eigenen Datenkopien und Modellzugängen – ohne zentrale Kontrollen und Kostenübersicht.
- **Unklare Datenbasis:** Veraltete, widersprüchliche oder falsch berechtigte Dokumente führen zu plausibel klingenden, aber falschen Antworten.
- **Zu weite Berechtigungen:** Ein Agent mit breit berechtigtem technischem Konto vererbt diese Rechte an jede Schwachstelle.

- **Fehlende Evaluation:** Ohne Referenzdatensatz und Abnahmekriterien wird „gut genug“ nach Bauchgefühl entschieden.
- **Governance als Nachgedanke:** Datenschutz, Informationssicherheit und Mitbestimmung werden erst kurz vor dem Go-live einbezogen.

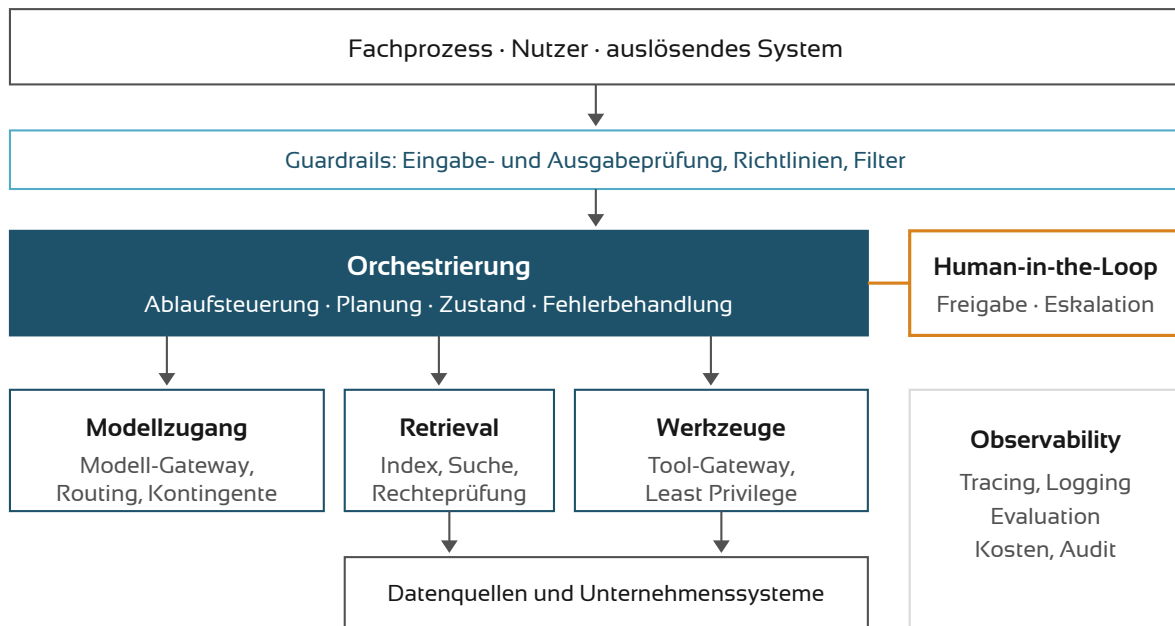
Hinzu kommt ein sich konkretisierender Rechtsrahmen: Der EU AI Act knüpft Pflichten an die Rolle eines Unternehmens und die Risikoklasse eines Systems, die DSGVO gilt für jede Verarbeitung personenbezogener Daten – auch wenn diese nur als Kontext an ein Modell übergeben werden. Beides gehört als Randbedingung in die Architektur, nicht in eine spätere Prüfung.

LEITFRAGEN

- Welcher Prozessschritt soll sich verbessern, und woran erkennen wir das?
- Welche Daten braucht das System, und welche Berechtigungen gelten heute dafür?
- Welche Aktionen darf das System selbst ausführen, welche nur vorschlagen?
- Wer verantwortet ein falsches Ergebnis?
- Wie erkennen wir, dass eine Änderung das Verhalten verschlechtert hat?

03 Architekturbausteine agentischer Systeme

Als Arbeitsdefinition gilt: Ein agentisches System ist eine Anwendung, in der ein Sprachmodell auf Basis eines Ziels entscheidet, welche Werkzeuge es in welcher Reihenfolge aufruft, die Ergebnisse auswertet und weitere Schritte ableitet. Je mehr Schritte das System selbst plant, desto höher sind die Anforderungen an Begrenzung, Beobachtbarkeit und Kontrolle.



Rahmen: Governance · Security · Datenschutz

Abbildung 1: Vereinfachte, herstellerneutrale Referenzarchitektur eines agentischen Systems. Die Orchestrierung greift ausschließlich über kontrollierte Schnittstellen auf Modelle, Wissen und Werkzeuge zu.

Daten und Retrieval

Unternehmenswissen gelangt meist über Retrieval-Augmented Generation (RAG) ins System: Zu einer Anfrage werden relevante Abschnitte aus einem Index gesucht und dem Modell als Kontext übergeben. Die Qualität entscheidet sich in der Aufbereitung – Zerlegung, Metadaten, Aktualisierung, Kombination von semantischer und Stichwortsuche. Zentral ist, dass die Zugriffsrechte der Quellsysteme im Index abgebildet und bei jeder Abfrage geprüft werden. Strukturierte Daten werden besser über definierte Schnittstellen als Werkzeug abgefragt als über frei generierte Datenbankzugriffe.

Modelle

Ein Modell-Gateway entkoppelt Anwendungen von einzelnen Anbietern, bündelt Authentifizierung, Kontingente und Kostenzuordnung und erlaubt die Wahl des Modells nach Aufgabe. Auswahlkriterien sind die Qualität im eigenen Evaluationsdatensatz, Datenresidenz, vertragliche Zusicherungen zur Nutzung von Eingaben, Latenz und Kosten pro Vorgang. Modellversionen werden fixiert und vor einem Wechsel gegen den Referenzdatensatz geprüft.

Werkzeuge und Tool Calling

Beim Tool Calling erzeugt das Modell strukturierte Funktionsaufrufe, die die Anwendung ausführt – die Kontrolle liegt in dieser ausführenden Schicht. Daraus folgen klare Regeln: minimale Rechte je Werkzeug (Least Privilege) und Trennung von lesenden und schreibenden Werkzeugen; eng gefasste Aktionen („Entwurf anlegen“ statt „Datensatz ändern“); wo möglich Handeln mit der Identität des auslösenden Nutzers; Prüfung von Parametern, Grenzwerten und Statusübergängen im Code statt im Prompt; idempotente und möglichst rücknehmbare Aktionen. Jede angebundene Schnittstelle erweitert die Angriffsfläche – unabhängig vom verwendeten Protokoll.

Orchestrierungsmuster

- **Fester Workflow mit Modellschritten:** deterministischer Ablauf, das Modell klassifiziert, extrahiert oder formuliert; am besten testbar.
- **Routing:** Ein Modellschritt ordnet Anfragen spezialisierten Verarbeitungstrecken zu.
- **Werkzeugschleife:** Ein Agent ruft iterativ Werkzeuge auf; Grenzen für Schrittzahl, Laufzeit und Kosten sind Pflicht.
- **Planer und Ausführende:** Eine Komponente zerlegt die Aufgabe, spezialisierte Komponenten bearbeiten Teilaufgaben.
- **Erzeugen und Prüfen:** Eine zweite Komponente prüft Ergebnisse gegen Kriterien – ohne menschliche Freigaben zu ersetzen.

Die Regel lautet: so viel Determinismus wie möglich, so viel Autonomie wie nötig. Jede zusätzliche Entscheidungsinstanz erschwert Test, Fehleranalyse und Kostenkontrolle.

Guardrails, Observability und Evaluation

Guardrails prüfen Eingaben und Ausgaben auf unzulässige Inhalte, personenbezogene Daten, Format und Quellenbezug und legen fest, welche Werkzeuge in welchem Kontext erlaubt sind. Observability macht jeden Lauf nachvollziehbar: Eingabe, Kontext, Modellversion, Werkzeugaufrufe, Ergebnis und Kosten. Diese Traces sind selbst schutzbedürftig; Aufbewahrung und Zugriff sind zu regeln. Evaluation stützt sich auf einen Referenzdatensatz mit repräsentativen und kritischen Fällen, bewertet automatisiert und durch Fachexperten. Jede Änderung an Prompt, Modell, Daten oder Werkzeugen durchläuft diese Prüfung vor dem Rollout.

04 Vorgehen: vom Use Case über den Pilot zur Skalierung

Das Vorgehen folgt dem Phasenmodell, mit dem wir KI-Vorhaben begleiten. Jede Phase endet mit einem Ergebnis, das eine begründete Entscheidung über die nächste ermöglicht – einschließlich der Entscheidung, einen Use Case nicht weiterzuverfolgen.

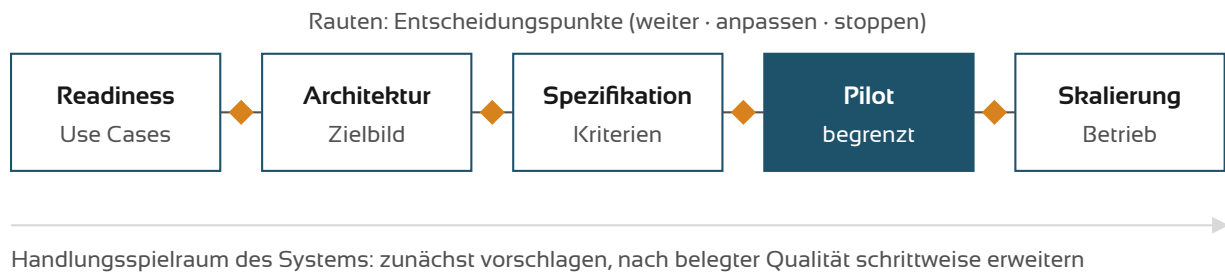


Abbildung 2: Phasen mit Entscheidungspunkten. Der Handlungsspielraum wächst erst, wenn Evaluation und Betriebserfahrung ihn begründen.

01 Orientierung und AI Readiness

Prozesse, Daten- und Systemlandschaft sowie Voraussetzungen werden mit Fachbereich und IT analysiert. Use-Case-Kandidaten werden einheitlich bewertet: fachlicher Nutzen und Messbarkeit, Datenverfügbarkeit und -eigentümerschaft, Integrationsaufwand, Fehlerfolgen und regulatorische Einordnung, Akzeptanz. Ergebnis: priorisierte Use Cases und ein Business Case mit offen ausgewiesenen Annahmen.

02 Solution und Agentic Architecture

Orchestrierungsmuster, Modellzugang, Retrieval, Werkzeuge, Berechtigungen sowie Entscheidungs- und Freigabepunkte werden konzipiert. Eine Bedrohungsanalyse betrachtet ausdrücklich Prompt Injection, Datenabfluss und übermäßige Handlungsbefugnisse. Ergebnis: ein Architektur- und Integrationskonzept, das sich in die bestehende Unternehmensarchitektur einfügt.

03 Requirements und Spezifikation

Fachanforderungen, Domänenregeln und Datenflüsse werden als überprüfbare Spezifikationen, etwa als BDD-Szenarien, festgehalten – einschließlich dessen, was das System nicht tun darf und wann es eskaliert. Parallel entsteht der Referenzdatensatz. Ergebnis: Spezifikationen und Akzeptanzkriterien als Abnahmegrundlage.

04 Pilot mit begrenztem Umfang

Definierter Nutzerkreis, begrenzter Datenausschnitt, eng gefasste Werkzeuge: Das System schlägt vor, Menschen entscheiden. Ergebnisse werden laufend gegen die Akzeptanzkriterien evaluiert, Fehlerbilder und Kosten pro Vorgang dokumentiert.
Hypothetisches Beispiel: Ein Assistent für den Innendienst klassifiziert Kundenanfragen, sucht passende Vertragsdokumente und legt Antwortentwürfe ab; das Versenden bleibt beim Sachbearbeiter. Erst wenn die Evaluation für eine klar umrissene Anfragekategorie die Abnahmekriterien erfüllt, wird über mehr Handlungsspielraum für genau diese Kategorie entschieden.

05 Skalierungsentscheidung

Auf Basis von Evaluation, Betriebserfahrung und aktualisiertem Business Case wird entschieden: skalieren, anpassen oder beenden. Ein beendeter Pilot mit belastbarer Erkenntnis ist kein Scheitern.

06 Skalierung, Einführung und Enablement

Bewährte Bausteine – Modell-Gateway, Retrieval, Tool-Gateway, Guardrails, Observability, Evaluationspipeline – werden zu gemeinsamen Plattformdiensten. Rollen, Betriebs-, Incident- und Änderungsprozesse werden etabliert, Teams geschult. Ergebnis: ein Governance- und Enablement-Plan.

05 Human-in-the-Loop, Governance und Security

Human-in-the-Loop

Eine Freigabe bei jedem Schritt führt zu reflexhaftem Bestätigen. Wirksamer ist eine Abstufung nach Folgen: **vorschlagen** (Menschen entscheiden und handeln), **freigeben** (eine vorbereitete Aktion läuft erst nach Bestätigung), **überwachen** (selbstständig in engen Grenzen, mit Stichproben) und **eskalieren** (bei Unsicherheit oder Regelverletzung). Freigabedialoge zeigen, was geschehen soll und auf welcher Grundlage; jede Freigabe wird protokolliert. Bei Entscheidungen mit erheblicher Wirkung auf Personen sind die DSGVO-Vorgaben zu automatisierten Einzelentscheidungen zu prüfen.

Governance

Ein schlanker Rahmen fügt sich in bestehende Strukturen für Architektur, Informationssicherheit und Datenschutz ein: ein KI-Systeminventar mit Zweck, Eigentümer, Daten, Modellen, Werkzeugen und Risikoeinstufung; die regulatorische Einordnung je System; Freigabeschritte für neue Systeme und wesentliche Änderungen; Nutzungsrichtlinien; klare Zuständigkeiten. NIST AI RMF (Govern, Map, Measure, Manage) und ISO/IEC 42001 geben Struktur – zugeschnitten auf Größe und Risikoprofil.

Security

Sprachmodelle trennen Anweisungen und Daten nicht zuverlässig. Daraus entsteht Prompt Injection – direkt über Nutzereingaben oder indirekt über verarbeitete Inhalte wie E-Mails, Webseiten oder Dokumente. Da Filter dieses Risiko nicht vollständig beseitigen, muss die Architektur den möglichen Schaden begrenzen:

- minimale Werkzeugrechte, menschliche Freigabe vor folgenreichen oder irreversiblen Aktionen,
- kein Ablauf, der vertrauliche Daten, nicht vertrauenswürdige Inhalte und ungeprüften Versand nach außen kombiniert,
- Modellausgaben als nicht vertrauenswürdige Eingaben für nachgelagerte Systeme behandeln,
- Zugangsdaten nur in der ausführenden Schicht, nie im Modellkontext,
- Red Teaming vor dem Pilot und nach wesentlichen Änderungen.

Die OWASP Top 10 for LLM Applications sind dafür eine gute Prüfliste. Daneben gelten die etablierten Anforderungen an Identitäts- und Zugriffsmanagement, Lieferkettenbewertung sowie Schwachstellen- und Incident-Management unverändert.

06 Artefakte und Deliverables

Die Ergebnisse beschreiben, was in einem Mandat gemeinsam erarbeitet wird; Umfang und Zuschnitt richten sich nach der Ausgangslage.

Entscheidung und Architektur

- Readiness-Bewertung und Use-Case-Portfolio mit Bewertungsmatrix
- Business Case je priorisiertem Use Case
- Architektur- und Integrationskonzept mit Architekturentscheidungen
- Berechtigungs- und Werkzeugmatrix
- Bedrohungsanalyse mit KI-spezifischen Risiken

Qualität, Betrieb, Governance

- Spezifikationen, Szenarien und Akzeptanzkriterien
- Evaluationskonzept mit Referenzdatensatz; Pilot-Evaluationsergebnisse
- Konzept für Logging, Tracing, Monitoring und Kosten
- KI-Systeminventar, Rollen- und Freigabemodell
- Governance- und Enablement-Plan

07 Erfolgsfaktoren und Risiken

Erfolgsfaktoren

- ein benannter Prozessschritt und fachlich Verantwortliche je Use Case
- Referenzdatensatz und Abnahmekriterien vor dem ersten Pilotnutzer
- das einfachste Orchestrierungsmuster, das die Aufgabe erfüllt
- gemeinsame Bausteine statt Insellösungen
- frühe Einbindung von Datenschutz, Sicherheit, Betrieb und Mitbestimmung
- Wissensaufbau im eigenen Team

Risiken

- plausible, aber falsche Ergebnisse ohne Quellenbezug
- übermäßige Befugnisse von Agenten
- unbemerkte Verhaltensdrift nach Modell- oder Datenänderungen
- schwer vorhersagbare Kosten mehrstufiger Läufe
- Anbieterabhängigkeit
- ungeprüftes Bestätigen von Vorschlägen

08 Checkliste

VOR PILOT UND SKALIERUNG

- Prozessschritt, Nutzen und fachliche Verantwortung sind benannt.
- Datenquellen, Eigentümer und Zugriffsrechte sind im Retrieval abgebildet.
- Eine regulatorische Ersteinordnung (EU AI Act, DSGVO) liegt geprüft vor.
- Das Orchestrierungsmuster ist das einfachste, das ausreicht.
- Werkzeuge haben minimale Rechte; Grenzwerte liegen im Code.
- Vorschlägen, Freigeben, Überwachen und Eskalieren sind je Aktion festgelegt.
- Die Bedrohungsanalyse berücksichtigt direkte und indirekte Prompt Injection.
- Referenzdatensatz und Abnahmekriterien existieren; Änderungen laufen durch Regressionsprüfung.
- Läufe werden protokolliert; Schrittzahl, Laufzeit und Kosten sind begrenzt.
- Betrieb, Incident-Management und KI-Inventar sind geklärt.
- Der Business Case ist auf Basis der Pilotergebnisse aktualisiert.

09 Nächster Schritt

Ein sinnvoller Einstieg ist eine kompakte Standortbestimmung: Welche Use Cases werden bereits diskutiert oder erprobt, wie sind sie architektonisch aufgestellt, und was fehlt für einen verantwortbaren Pilot? Daraus entsteht eine priorisierte Liste mit klaren Empfehlungen – auch dazu, was Sie nicht weiterverfolgen sollten.

Wir unterstützen Sie dabei in Rollen auf Kundenseite: von Use-Case- und Strategieberatung über Architektur- und Integrationsdesign sowie Architektur- und Governance-Reviews laufender KI-Vorhaben bis zur Begleitung von Pilot, Einführung und Wissensübergabe.

10 Quellen und Standards

Bezugsrahmen sind die folgenden öffentlichen Regelwerke und Standards; die Hinweise dienen der Orientierung, sind keine Rechtsberatung, und Anwendungszeitpunkte sowie Änderungen sind jeweils am aktuellen Stand zu prüfen.

- **Verordnung (EU) 2024/1689** (KI-Verordnung, „EU AI Act“): risikobasierter Rechtsrahmen mit gestaffelt anwendbaren Pflichten für Anbieter und Betreiber.
- **Verordnung (EU) 2016/679** (DSGVO), u. a. Art. 22 (automatisierte Einzelentscheidungen) und Art. 35 (Datenschutz-Folgenabschätzung).
- **NIST AI 100-1**, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023; ergänzend NIST AI 600-1 (Profil für generative KI).
- **ISO/IEC 42001:2023**, Managementsystem für künstliche Intelligenz.
- **ISO/IEC 23894:2023**, Leitlinien zum Risikomanagement für künstliche Intelligenz.
- **ISO/IEC 27001:2022**, Informationssicherheits-Managementsysteme.
- **OWASP Top 10 for Large Language Model Applications**, Übersicht zentraler Sicherheitsrisiken von LLM-Anwendungen.

Lassen Sie uns sprechen

Für Infos zu diesem und weiteren spannenden Themen rund um KI-Vorhaben freuen wir uns auf den Austausch mit Ihnen!

UNTERNEHMEN

DIP Deutsche IT und
Projektberatungsgesellschaft mbH
Rahmannstraße 11 · 65760 Eschborn
info@dip-consulting.de
dip-consulting.de

IHR ANSPRECHPARTNER

Manuel Fernandez

Management Consultant, Geschäftsführer
info@dip-consulting.de

Seit 2013 ist Herr Fernandez Geschäftsführer der Deutschen IT und Projektberatungsgesellschaft und seit 16 Jahren als IT-Berater in vielfältigen Rollen bei strategischen als auch bei implementierungsnahen Projekten im Einsatz.

