



BLUEPRINT

Souveräne Unternehmens-KI DSGVO-konform einführen.

Eine KI auf Ihrem eigenen Wissen – DSGVO-konform, herstellerneutral und ohne Anbieter-Lock-in. Der 3-Phasen-Blueprint mit OpenWebUI.

Inhalt

1. Auf einen Blick	3
2. Warum KI-Einführung scheitert	4
3. Die Entscheidung kommt vor der Technik	5
4. Der 3-Phasen-Blueprint	6
5. Governance von Anfang an	7
6. Kosten & Wirtschaftlichkeit	8
7. Reifegrad-Selbstcheck	9
8. Referenzarchitektur (Anhang)	10
9. Warum Smart Commerce	11

1.

Auf einen Blick

Das Ziel ist klar: eine souveräne, DSGVO-konforme Unternehmens-KI, die auf Ihrem eigenen Wissen arbeitet – ohne dass sensible Daten das Haus verlassen und ohne Abhängigkeit von einem einzigen Anbieter.

Die meisten Unternehmen scheitern auf dem Weg dorthin nicht an der Technologie, sondern am Modell-Dschungel, an Datenschutzbedenken, an unkontrollierter Schatten-KI und an der Angst vor Anbieter-Lock-in.

Dieser Blueprint zeigt einen strukturierten, herstellerneutralen Weg, KI produktiv und sicher in den Betrieb zu bringen – in drei aufeinander aufbauenden Phasen und auf einer offenen Plattform, die Sie nicht an einen einzigen Anbieter bindet. Das technische Herzstück ist OpenWebUI: eine offene Oberfläche, unter der lokale, Cloud- und hybride Modelle gleichberechtigt laufen.

Die Kernidee: Erst die Entscheidung, dann die Technik. Wer Datensouveränität, Governance und Modell-Unabhängigkeit von Anfang an mitdenkt, vermeidet die teuren Korrekturen, die unkoordinierte KI-Piloten später erzwingen.

Für wen

CIO/CTO (Kontrolle, Sicherheit, TCO), Head of Digital (Use-Cases, Tempo), Geschäftsführung/CFO (Relevanz, Kosten).
Der Blueprint bedient Gatekeeper und Initiator.

Was Sie mitnehmen

Die Entscheidungslogik, den 3-Phasen-Weg, einen Governance-Rahmen, eine Referenzarchitektur im Anhang – und einen Selbstcheck, mit dem Sie Ihren Reifegrad sofort einordnen.

2.

Warum KI-Einführung scheitert

Unternehmen wollen KI einführen – und bleiben trotzdem stecken. Nicht, weil die Modelle fehlen, sondern weil fünf wiederkehrende Hürden den Weg blockieren. Wer sie kennt, kann sie umgehen.

Hürde	Was dahintersteckt
Orientierungslosigkeit im Modell-Dschungel	GPT-4o, Llama 4, Mistral, Phi-4, DeepSeek-R1 – welches Modell für welchen Use-Case? Interne Teams können diese Frage selten beantworten und verschieben die Entscheidung.
Fehlende Datensouveränität	Cloud-Modelle bedeuten, dass sensible Unternehmensdaten den eigenen Server verlassen. Datenschutz und Compliance blockieren den produktiven Einsatz.
Keine einheitliche Oberfläche	Jedes Modell bringt sein eigenes Interface mit. Teams weichen auf Schatten-KI, Insellösungen und Consumer-Tools aus – unkontrolliert und unsicher.
Mangelndes Enablement	Werkzeuge werden eingeführt, aber nicht verankert. Mitarbeitende wissen nicht, wie sie KI sinnvoll in ihren Arbeitsalltag integrieren.
Vendor-Lock-in-Angst	Unternehmen scheuen die langfristige Abhängigkeit von einem einzigen KI-Anbieter – zu Recht, denn der Markt bewegt sich schneller als jeder Vertrag.

Woran Sie erkennen, dass es Sie betrifft

- IT und Fachabteilungen sind sich uneinig, welches KI-Tool offiziell eingesetzt werden soll.
- Mitarbeitende nutzen ChatGPT privat, aber es gibt keine genehmigte Alternative.
- KI-Piloten scheitern an Datenschutzbedenken oder fehlender IT-Integration.
- Die Geschäftsführung fragt nach einer KI-Strategie – niemand hat eine klare Antwort.
- Wettbewerber kommunizieren bereits aktiv über ihren KI-Einsatz.

3.

Die Entscheidung kommt vor der Technik

Bevor über Tools gesprochen wird, sind zwei Fragen zu klären, die später kaum noch korrigierbar sind: **Wo dürfen Ihre Daten verarbeitet werden?** und

Wie verhindern Sie Abhängigkeit von einem einzigen Anbieter? Aus den Antworten ergibt sich das Betriebsmodell – nicht umgekehrt.

Datensouveränität: drei Betriebsmodelle

Modell	Eignung	Datenlage
Lokal / On-Premise	Sensible Daten, regulierte Branchen, maximaler Schutz. Modelle laufen im eigenen Rechenzentrum.	Daten verlassen das Haus nicht. DSGVO-konform by design.
Cloud (Azure)	Höchste Modellvielfalt und Leistung, schnelle Verfügbarkeit, Enterprise-Authentifizierung.	Verarbeitung in der Cloud, vertraglich und technisch absicherbar.
Hybrid	Der Regelfall: sensible Vorgänge lokal, anspruchsvolle Aufgaben in der Cloud – je Use-Case entschieden.	Steuerbar pro Anwendungsfall, eine Oberfläche.

Warum Open Source die Lock-in-Frage entschärft: Eine offene, herstellerneutrale Oberfläche entkoppelt die Bedienung vom Modell. Sie wechseln das Modell per Klick, ohne neues Tool, ohne neue Schulung und ohne Ihre Daten oder Workflows umzuziehen. Aus einer Anbieter-Wette wird eine revidierbare Entscheidung.

Genau hier setzt OpenWebUI an: Es ist die führende Open-Source-Plattform für KI-Oberflächen (über 123.000 GitHub-Stars, unterstützt von a16z und Mozilla) und vereint lokale, Cloud- und hybride Modelle unter einer einheitlichen, ChatGPT-ähnlichen Bedienung. Die technischen Details dazu finden Sie im Anhang.

4.

Der 3-Phasen-Blueprint

KI sicher einzuführen heißt, drei Dinge nacheinander richtig zu machen: den eigenen Stand verstehen, eine kontrollierte Plattform aufsetzen und die Nutzung im Alltag verankern. Jede Phase hat ein klares Ergebnis.

PHASE 1

Assess – KI-Readiness-Assessment

Inhalt: Ist-Analyse der Systemlandschaft, Priorisierung der Use-Cases, Datenschutz- und Reifegrad-Bewertung.

Ergebnis: eine belastbare KI-Roadmap mit konkreter Architekturempfehlung – die Grundlage, auf der alle weiteren Entscheidungen ruhen.

PHASE 2

Implement – OpenWebUI-Deployment

Inhalt: Setup und Konfiguration der Plattform – lokal (Ollama), über Azure AI Foundry oder auf GPU-Infrastruktur wie RunPod – inklusive rollenbasierter Zugriffssteuerung (RBAC), Single Sign-On und Datenschutz.

Ergebnis: eine laufende, produktionsreife und DSGVO-konforme Plattform statt eines Pilot-Provisoriums.

PHASE 3

Enable – Mitarbeiter-Enablement

Inhalt: Schulungen, Prompt-Engineering-Workshops, eine Use-Case-Bibliothek aus echten Projekten und ein Governance-Framework.

Ergebnis: selbstständige, sichere KI-Nutzung, die im Unternehmen verankert ist – nicht nur ein Tool, das niemand benutzt.

Wichtig: Die Reihenfolge ist kein Zufall. Wer mit Phase 2 beginnt (Technik zuerst), baut auf Annahmen statt auf Analyse. Wer Phase 3 überspringt (kein Enablement), hat eine Plattform, die niemand nutzt. Der Wert entsteht aus der Sequenz.

5.

Governance von Anfang an

Der teuerste Fehler ist, KI erst einzuführen und Governance später nachzuziehen. Dann existieren bereits Schatten-KI, Insellösungen und ungeklärte Datenflüsse. Eine kontrollierte Plattform dreht das um: Steuerung ist Teil des Fundaments, nicht der Nachgedanke.

Baustein	Was er sichert
Rollenbasierte Zugriffe (RBAC)	Wer darf welche Modelle, welche Daten und welche Funktionen nutzen – granular steuerbar.
Identität & SSO (Entra ID, LDAP, SCIM)	Anbindung an die bestehende Unternehmens-Identität, keine statischen API-Keys, sauberes On-/Offboarding.
Audit-Logs	Nachvollziehbarkeit, wer wann was getan hat – Voraussetzung für Compliance und Revision.
Verschlüsselte Datenhaltung	Schutz der gespeicherten Inhalte und Gesprächshistorien.
Use-Case- & Prompt-Richtlinien	Klare Leitplanken, was erlaubt ist – verhindert Wildwuchs und reduziert Risiko.

Konsequenz ohne Governance: Wer heute unstrukturiert einführt, zahlt morgen doppelt – für Technischulden, Compliance-Risiken und den verpassten Produktivitätssprung.

6.

Kosten, die man kontrollieren kann

Ein offener, herstellernerutraler Ansatz verändert die Kostenlogik grundlegend. Statt eines aufgeblähten SaaS-Abos mit Pro-Kopf-Lizenz zahlen Sie für Infrastruktur, die Sie steuern – und behalten die Hoheit über die Gesamtkosten.

- **Kein Lock-in-Aufschlag:** Open Source statt langfristiger Lizenzbindung an einen Anbieter.
 - **Skalierung nach Bedarf:** serverlose GPU-Endpoints (z. B. RunPod), die bei Inaktivität auf null skalieren – Sie zahlen Inferenz, nicht Leerlauf.
 - **Ein Tool statt App-Wildwuchs:** eine Oberfläche für alle Modelle senkt Schulungs- und Administrationsaufwand.
 - **Vom Piloten zum Rollout mit demselben Stack:** von 10 Nutzern bis 5.000 – ohne Plattformwechsel.
-

Faustregel: Die Architektur, die Sie früh wählen, bestimmt Ihre Kostenkurve mehr als jeder einzelne Modellpreis. Offenheit und Modell-Unabhängigkeit sind die wirksamsten Kostenhebel.

7.

Reifegrad-Selbstcheck

Eine ehrliche Standortbestimmung in zehn Fragen. Je mehr Sie mit „nein“ oder „unklar“ beantworten, desto eher lohnt sich der Einstieg über Phase 1 (Assessment).

#	Frage
1	Gibt es eine genehmigte KI-Lösung, oder nutzen Teams private Consumer-Tools?
2	Ist geklärt, welche Daten lokal verarbeitet werden müssen und welche in die Cloud dürfen?
3	Haben Sie priorisierte Use-Cases – oder viele Ideen ohne Reihenfolge?
4	Sind Modelle und Oberflächen vereinheitlicht, oder herrscht Insel-Wildwuchs?
5	Gibt es rollenbasierte Zugriffe und ist die KI an Ihre Identitätsverwaltung angebunden?
6	Werden KI-Nutzungen protokolliert (Audit-Logs) für Compliance und Revision?
7	Sind Ihre Mitarbeitenden geschult, KI sinnvoll im Alltag einzusetzen?
8	Existiert ein Governance-Rahmen mit klaren Leitplanken?
9	Können Sie das Modell wechseln, ohne Tool und Workflows zu ändern?
10	Lässt sich Ihr Pilot ohne Plattformwechsel zum unternehmensweiten Rollout skalieren?

Wo stehen Sie? Nutzen Sie diese zehn Fragen als Ausgangspunkt – und besprechen Sie Ihren Reifegrad sowie die nächsten Schritte direkt mit uns.

8.

Referenzarchitektur

Für IT-Verantwortliche: So sieht der Stack konkret aus. OpenWebUI fungiert als herstellerneutrale Oberfläche, unter der drei Modell-Infrastrukturen gleichberechtigt und parallel laufen.

Deployment-Optionen

Quelle	Eigenschaften	Typische Modelle
Lokal / On-Premise (Ollama)	Vollständig offline, keine Datenweitergabe, DSGVO-konform; läuft auf eigener Hardware im eigenen Rechenzentrum.	Llama 4, Mistral, Phi-4, DeepSeek-R1
Microsoft Azure AI Foundry	Enterprise-Anbindung über Azure OpenAI, Entra-ID-Authentifizierung ohne statische API-Keys; Zugang zu über 1.900 Modellen.	GPT-4o, o3-mini, Claude, Llama 4
RunPod (Serverless GPU)	GPU-Endpoints, die bei Inaktivität auf null skalieren; kosteneffizient für leistungsintensive Modelle ohne eigene GPU-Infrastruktur.	Frei wählbare Open-Source-Modelle
Hybrid	Alle drei Quellen parallel in einer Oberfläche; Modellwechsel per Klick, zentrale Nutzerverwaltung, einheitliche Gesprächshistorie.	Kombination je Use-Case

Integration & Enterprise-Readiness

- **Identität:** SSO / LDAP / SCIM, Entra ID
 - **Schnittstellen:** REST-API, kompatibel mit bestehenden IT-Stacks
 - **Eigenes Wissen:** RAG-Anbindung an ERP, CRM, SharePoint
- **Betrieb:** Multi-Node, HA-Clustering
 - **Sicherheit:** verschlüsselte Datenbanken, Audit-Logs, RBAC
 - **Skalierung:** vom 10-Nutzer-Piloten bis 5.000 Mitarbeitende

Hinweis: RAG (Retrieval-Augmented Generation) macht Unternehmenswissen für die KI nutzbar, ohne es in das Modell zu trainieren – die Daten bleiben in Ihrer Hand.

9.

Warum Smart Commerce

Strategie, Technik und Enablement aus einer Hand – herstellerneutral und datensouverän.

Bei der Smart Commerce SE kombinieren wir jahrelange Erfahrung unserer eCommerce- und Digital-Consulting-Expert:innen mit technischer Expertise in den Kernbereichen eCommerce-Plattformen, eCommerce-Cloud, CMS, CRM und Digital Marketing. So bieten wir nach umfassenden Analysen problemspezifische Lösungen in Kombination mit einer individuellen, ganzheitlichen Strategie.

Unser kompetentes TEC-Team aus Senior Project Consultants, erfahrenen Software-Architekt:innen, Software- und Web-Entwickler:innen sowie Betriebs-Expert:innen mit tiefgehender eCommerce-Erfahrung bietet ein umfassendes Dienstleistungspaket für High-End-eCommerce-Plattformen und die digitale Unternehmenstransformation.

Oberste Priorität hat für uns, unsere Kund:innen von Anfang an mitzunehmen und Projekte gemeinsam aufzubauen – damit Weg und Ziel jederzeit transparent bleiben. Wir haben uns dem langfristigen Online-Erfolg unserer Geschäftspartner:innen verschrieben: mit nachhaltigen, nutzerorientierten und datenzentrierten Lösungen.

Seit mehr als 10 Jahren folgen wir dem Smart Way, der unsere Kultur von Beginn an prägt: Wir sind ein mitarbeiter:innenzentriertes Unternehmen. Unsere 100+ Mitarbeiter:innen sind Mitunternehmer:innen und Mitgestalter:innen – sie halten die Aktien und gestalten das Unternehmen aktiv mit, von arbeitsplatzspezifischen Entscheidungen bis zu Werten, Zielen und Führungsgrundsätzen.

Was uns unterscheidet	Bedeutung für Sie
Herstellerneutral	Wir binden jedes Modell an – kein Vendor-Favorit, keine versteckte Agenda.
Datensouverän	Lokale Modelle auf eigener Infrastruktur als Erstangebot, nicht als Ausnahme.
Ganzheitlich	Strategie, Technik und Enablement in einem Paket statt aus drei Quellen.
Schnell	Lauffähige Umgebung in Tagen, kein monatelanger Rollout.
Praxisnah	Use-Case-Bibliothek aus echten Kundenprojekten statt Theorie.

Proof-Points: OpenWebUI mit über 123.000 GitHub-Stars, unterstützt von a16z und Mozilla; produktiver Einsatz bei Unternehmen wie Samsung Semiconductor; über 1.900 Modelle über Azure AI Foundry aus einer Oberfläche.

SIE HABEN FRAGEN?

Vereinbaren Sie eine Demo

In einem unverbindlichen Termin besprechen wir Ihren Reifegrad, Ihre Ziele und wie ein souveräner KI-Einstieg mit OpenWebUI für Sie aussehen kann.

Demo-Termin vereinbaren



Mario Wentzel

KI-Consulting · +49 341 99 15 36 00 · hello@smartcommerce.de



Standort Jena

Steinweg 10, 07743 Jena
+49 3641 3 16 10 20

Standort Böblingen

Otto-Lilienthal-Straße 36,
71034 Böblingen
+49 711 18 42 01 00

Standort Leipzig

Naumburger Straße 25,
04229 Leipzig
+49 341 99 15 36 00

www.smartcommerce.de