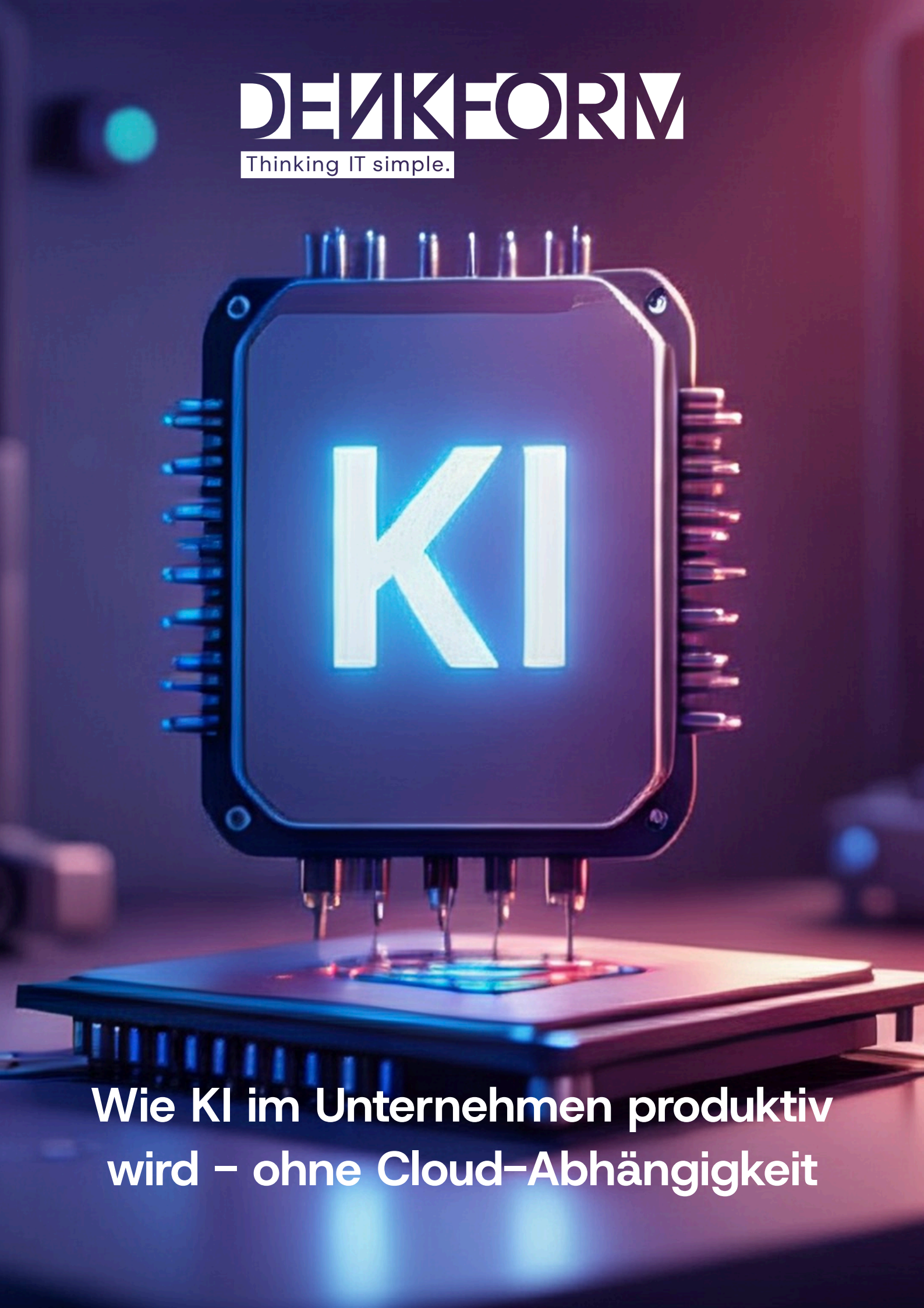
DENKFORM

Thinking IT simple.



Wie KI im Unternehmen produktiv wird – ohne Cloud-Abhängigkeit

Inhaltsverzeichnis

Executive Summary – On-Prem KI als Betriebsentscheidung

Künstliche Intelligenz ist in Unternehmen kein Erkenntnisproblem mehr, sondern ein Entscheidungsproblem.

Die Frage lautet nicht mehr, ob KI eingesetzt wird, sondern unter welchen Bedingungen sie dauerhaft verantwortbar betrieben werden kann.

Die meisten KI-Projekte scheitern nicht an der Technologie, sondern daran, dass Betrieb, Kosten und Verantwortung zu spät geklärt werden. Was im Test funktioniert, wird im Alltag fragil, sobald Support, Datenschutz oder Budgetverantwortung konkret werden.

Damit verschiebt sich die eigentliche Kernfrage:

Nicht welche KI eingesetzt wird, sondern wo ihr Betrieb verankert ist.

On-Prem und Cloud unterscheiden sich dabei weniger in der Leistungsfähigkeit als in der Verantwortungsstruktur.

Aspekt	Cloud-KI	On-Prem-KI
Modellbetrieb	Extern	Intern
Kosten	Nutzungsabhängig	Planbar
Datenkontrolle	Vertraglich geregelt	Technisch kontrolliert
Betriebsverantwortung	Anbieter	Eigenes Unternehmen

Dieses Playbook betrachtet KI deshalb als Infrastrukturentscheidung.

Ziel ist die dauerhafte Betreibbarkeit im Unternehmensalltag.

Vorwort – Warum dieses Playbook existiert

Viele Unternehmen beschäftigen sich mit KI – doch der Schritt in den regulären Betrieb bleibt oft aus.

Der Grund ist nicht fehlender Nutzen, sondern fehlende Abstimmung zwischen Fachbereichen, IT und Geschäftsführung. Solange diese Perspektiven getrennt bleiben, bleibt KI ein Thema für Workshops und nicht für den Alltag.

Häufig beginnt die Diskussion bei Tools. Erst danach folgen Fragen zu Datenschutz, Integration und Kosten und stellen die ursprünglichen Entscheidungen wieder infrage.

Dieses Dokument wählt bewusst den umgekehrten Weg:

Es beginnt beim Betrieb und leitet daraus Architektur und Einsatzmöglichkeiten ab.

Ziel ist nicht, den Einstieg zu bremsen, sondern ihn überhaupt möglich zu machen.

1. Für wen dieses Playbook ist (Geschäftsführung, IT & Fachbereiche)

Dieses Playbook richtet sich an Personen, die KI nicht ausprobieren, sondern verantworten:

- Geschäftsführung → langfristige Investitions- und Risikobewertung
- IT → Integration, Wartbarkeit und Sicherheit
- Fachbereiche → realistisch nutzbare Anwendungen

Nicht adressiert sind bewusst experimentelle Umgebungen oder Forschungsprojekte. Dort stehen technologische Möglichkeiten im Vordergrund. In diesem Dokument steht dagegen der stabile Betrieb innerhalb einer bestehenden Organisation im Mittelpunkt.

Der Fokus liegt daher nicht auf maximaler Flexibilität, sondern auf Vorhersagbarkeit.

2. Unsere Perspektive: KI ist Betrieb, nicht Experiment

Ein Proof-of-Concept beantwortet die technische Frage, ob ein Modell eine Aufgabe lösen kann.

Mit „Modell" ist in diesem Zusammenhang ein trainiertes KI-System gemeint (häufig sogenannte Large Language Models / LLMs).

Sie bilden die Grundlage moderner KI-Anwendungen und werden je nach Einsatzzweck mit Unternehmensdaten und Prozessen kombiniert.

Ein produktives System muss andere Fragen beantworten:

- Wer betreibt es?
- Wie wird es abgesichert?
- Was passiert im Fehlerfall?

Solange diese Fragen offen sind, bleibt KI ein Demonstrator – kein System. Genau deshalb verlaufen viele Initiativen im Sande — nicht weil die Technologie ungeeignet wäre, sondern weil sie außerhalb der normalen IT-Logik betrachtet wird.

In der Praxis verhält sich KI wie jede andere Unternehmensanwendung: Sie benötigt Zugriffskontrolle, Monitoring, Backups und klare Zuständigkeiten.

Die entscheidende Veränderung besteht darin, KI nicht nach ihren Fähigkeiten zu bewerten, sondern nach ihrer Betriebsfähigkeit.

3. Warum der Einstieg in KI so oft hängen bleibt

Unternehmen scheitern selten an mangelndem Interesse an KI. Sie scheitern daran, dass zu viele Entscheidungen gleichzeitig getroffen werden sollen.

Wird zunächst ein Tool ausgewählt, entsteht Begeisterung über die Möglichkeiten. Kurz darauf folgen Datenschutz-, Integrations- und Kostenfragen und stellen die ursprüngliche Entscheidung wieder in Frage. Das Projekt kehrt in die Konzeptphase zurück.

Der Ablauf wiederholt sich, weil die zugrunde liegende Frage unbeantwortet bleibt: unter welchen Bedingungen der Einsatz überhaupt verantwortbar ist.

KI erzeugt gleichzeitig rechtliche, technische und wirtschaftliche Auswirkungen. Werden diese erst nachträglich bewertet, blockieren sie sich gegenseitig. Werden sie vorab geklärt, lassen sich Anwendungen realistisch einordnen.

Das Ziel ist daher nicht, schneller zu starten, sondern früher die richtigen Fragen zu stellen.

4. On-Prem, Cloud oder Hybrid: Architektur- und Betriebsentscheidung

Die Diskussion um KI-Betrieb wird häufig ideologisch geführt: flexibel gegen sicher, modern gegen konservativ. In der Praxis geht es um etwas deutlich Konkreteres: wo die Verarbeitung stattfindet und wer sie technisch kontrolliert.

Sobald ein Mitarbeiter eine Anfrage stellt, durchläuft jede KI-Lösung denselben grundsätzlichen Ablauf: Die Eingabe wird verarbeitet, Kontextdaten werden hinzugefügt, das Modell berechnet eine Antwort und das Ergebnis wird ausgeliefert.

Der Unterschied zwischen Cloud und On-Prem liegt nicht im Prinzip, sondern darin, auf welcher Infrastruktur diese Schritte stattfinden.

In einer Cloud-Architektur verlässt die Anfrage das Unternehmensnetzwerk. Sie wird an den Anbieter übertragen, dort verarbeitet und das Ergebnis zurückgegeben. Damit übernimmt der Anbieter nicht nur Rechenleistung, sondern auch wesentliche Teile des Betriebs Modellupdates, Skalierung, Verfügbarkeit und Teile der Sicherheitsarchitektur.

In einer On-Prem-Architektur bleibt die gesamte Verarbeitung im eigenen Netzwerk. Die Anfrage wird intern verarbeitet, der Kontext lokal geladen, das Modell lokal ausgeführt und das Ergebnis intern bereitgestellt. Der Betrieb liegt vollständig beim Unternehmen oder einem beauftragten Dienstleister innerhalb der eigenen Infrastrukturgrenzen. Der technische Unterschied lässt sich auf eine einfache Frage reduzieren: Wo wird die Inferenz berechnet?

Schritt	Cloud-KI	On-Prem-KI
Anfrage	An externen Dienst gesendet	Bleibt im internen Netzwerk
Kontextdaten	Werden übertragen	Bleiben lokal
Modellberechnung	Beim Anbieter	Auf eigener Hardware
Antwort	Zurückübertragen	Lokal erzeugt
Updates	Automatisch extern	Kontrolliert intern

Diese Entscheidung hat direkte Folgen für drei Bereiche: Verantwortung, Komplexität und Kostenlogik.

Verantwortung

- Cloud → Betrieb teilweise beim Anbieter
- On-Prem → Betrieb vollständig im eigenen Unternehmen

Cloud reduziert Aufwand, erfordert aber Vertrauen.
On-Prem erhöht Verantwortung, reduziert Abhängigkeiten.

Komplexität

- Cloud → einfacher Einstieg
- Hybrid → in der Praxis oft komplex (Datenklassifizierung, Routing, Fehlersuche)

Ein Hybrid-Modell lohnt sich nur bei klar getrennten Datenflüssen.

Kostenlogik

- Cloud → Kosten steigen mit Nutzung
- On-Prem → Kosten entstehen vorab, Nutzung wird günstiger

Einordnung

Keine Architektur ist grundsätzlich überlegen – sie lösen unterschiedliche Anforderungen:

- Cloud → schnelle Einführung, geringe interne Verantwortung
- On-Prem → Kontrolle, Planbarkeit, Integration

Die Entscheidung ist keine technologische, sondern eine betriebliche:
Wer trägt im Alltag Verantwortung für Betrieb, Kosten und Risiken?

5. Architektur: Warum Standardhardware der unterschätzte Erfolgsfaktor ist

In vielen KI-Diskussionen beginnt die Architekturfrage bei Modellgrößen und Rechenleistung. Für den produktiven Betrieb ist jedoch nicht entscheidend, wie leistungsfähig ein System maximal sein kann, sondern ob es dauerhaft beherrschbar bleibt.

Ein KI-System besteht im Alltag nicht nur aus einem Modell. Es umfasst mehrere klar getrennte Komponenten: Zugriffsschicht, Kontextspeicher, Inferenzprozess und Protokollierung. Diese Bausteine müssen in bestehende IT-Strukturen integrierbar sein — Netzwerk, Authentifizierung, Backup und Monitoring eingeschlossen.

Je spezieller die Hardware gewählt wird, desto stärker entfernt sich das System von der üblichen IT-Logik. Spezialcluster, experimentelle GPU-Konfigurationen oder individuell zusammengestellte Server erhöhen zwar die theoretische Leistung, erzeugen aber Abhängigkeiten von Einzelpersonen und erschweren Wartung sowie Austausch.

GPUs (Graphics Processing Units) sind spezialisierte Hochleistungsprozessoren, die besonders für KI-Berechnungen geeignet sind. Sie ermöglichen hohe Verarbeitungsgeschwindigkeit, erhöhen jedoch häufig Komplexität, Energiebedarf und Anforderungen an den Betrieb.

Standardisierte Hardware verfolgt ein anderes Ziel: reproduzierbares Verhalten. Performance wird nicht maximiert, sondern stabil gehalten. Updates lassen sich planen, Komponenten austauschen und Systeme im Fehlerfall ersetzen, ohne die Architektur neu zu entwerfen.

Ein typischer produktiver Aufbau folgt deshalb keiner HPC-Logik, sondern einer Service-Logik (HPC steht für „High Performance Computing“ und beschreibt hochspezialisierte Rechenumgebungen, die auf maximale Leistung optimiert sind):

- ein System für Zugriff und Authentifizierung
- ein System für Kontextindex (z. B. Dokumente/Embeddings)
- ein System für Modellinferenz
- ein System für Protokollierung und Monitoring

Für viele Unternehmen ist jedoch nicht die maximale Rechenleistung entscheidend, sondern ein stabiler, wartbarer und langfristig beherrschbarer Betrieb.

Diese Trennung verhindert, dass ein Modellupdate gleichzeitig Zugriffskontrolle oder Datenhaltung verändert. KI wird damit zu einem klar abgegrenzten Dienst innerhalb der bestehenden Infrastruktur — vergleichbar mit einem Mail- oder Datenbankdienst. Standardhardware ist deshalb kein Kompromiss. Sie begrenzt technische Extreme, um organisatorische Stabilität zu ermöglichen.

6. Skalierung: Warum klein starten strategisch klug ist

Skalierung wird im KI-Kontext häufig als Planungsaufgabe verstanden: erwartete Nutzerzahlen, maximale Last, zukünftige Modellgrößen. In der Praxis entsteht Skalierung jedoch aus Nutzung — und Nutzung entsteht aus Vertrauen in Stabilität.

Ein System, das theoretisch tausend Anfragen gleichzeitig verarbeiten kann, aber organisatorisch nicht beherrschbar ist, wird nicht produktiv eingesetzt. Ein System mit begrenzter Leistung, das zuverlässig funktioniert, wird dagegen genutzt und anschließend erweitert.

Deshalb beginnt produktive KI nicht mit Kapazitätsplanung, sondern mit einem stabilen Basisknoten. Dieser bildet die reale Nutzung ab: Antwortzeiten, typische Anfragegrößen und tatsächliche Lastprofile. Erst daraus lässt sich sinnvoll ableiten, ob zusätzliche Systeme notwendig sind.

Skalierung erfolgt dann durch Replikation, nicht durch Umbau. Weitere Knoten übernehmen identische Aufgaben und werden über Lastverteilung eingebunden. Architektur und Betrieb bleiben gleich — nur die Kapazität steigt.

Das reduziert zwei typische Risiken: Überdimensionierung vor Nutzung und komplexe Zentralcluster, die bei Ausfall den gesamten Dienst blockieren.

Strategisch bedeutet das: KI wächst nicht durch Vorplanung, sondern durch tatsächliche Verwendung.

7. Datenschutz & Security: Warum On-Prem hier Klarheit schafft

Datenschutzfragen entstehen im KI-Umfeld selten durch fehlende Sicherheitsmechanismen, sondern durch unklare Datenflüsse. Sobald nicht mehr eindeutig nachvollziehbar ist, wo Inhalte verarbeitet oder gespeichert werden, wird jede Bewertung schwierig, intern wie extern.

In einer lokal betriebenen Architektur bleibt der Datenpfad eindeutig: Anfrage, Kontext und Ergebnis bewegen sich ausschließlich innerhalb des eigenen Netzwerks. Sicherheitsmaßnahmen orientieren sich damit an bekannten Prinzipien — Zugriffskontrolle, Segmentierung, Protokollierung und Backup.

Technisch bedeutet das: Das Modell wird wie eine interne Anwendung behandelt, nicht wie ein externer Kommunikationsdienst. Zugriffe können über bestehende Identitätsdienste gesteuert werden, Protokolle bleiben im eigenen Log-System und Test- sowie Produktivumgebungen lassen sich klar trennen.

Ein wesentlicher Unterschied zur Cloud liegt nicht primär in der Verschlüsselung oder Zugriffssicherung — diese existieren in beiden Varianten. Der Unterschied liegt in der Nachvollziehbarkeit: Bei externer Verarbeitung muss Vertrauen vertraglich hergestellt werden, bei interner Verarbeitung technisch.

Das reduziert Diskussionen weniger durch zusätzliche Sicherheit als durch eindeutige Zuständigkeit. Sicherheit wird damit zu einer Frage der Umsetzung, nicht der Interpretation.

8. Make or Buy: Die Entscheidung hinter der Architektur

Ein selbst aufgebautes System bietet maximale Freiheit: Modelle austauschen, Komponenten anpassen, eigene Logik integrieren. Solange Nutzung gering ist und einzelne Personen tief involviert sind, funktioniert das gut. Mit wachsender Nutzung entsteht jedoch ein anderer Effekt — Wissen konzentriert sich auf wenige Verantwortliche und die Wartung wird abhängig von Personen statt von Struktur.

Die Frage ist nicht, ob ein Unternehmen KI selbst bauen kann, sondern ob es sie dauerhaft betreiben will.

Standardisierte Architekturen verfolgen das Gegenteil: weniger Freiheitsgrade, dafür vorhersehbares Verhalten. Änderungen erfolgen kontrolliert, Updates reproduzierbar und Betrieb bleibt auch bei Personalwechsel möglich. Der Fokus verschiebt sich von technischer Optimierung zu organisatorischer Stabilität.

Die Entscheidung lässt sich pragmatisch einordnen:

Ansatz	Vorteil	Risiko
Eigenbau	Maximale Anpassbarkeit	Personenabhängiger Betrieb
Standardisierte Lösung	Stabiler Betrieb	Begrenzte Individualisierung

9. Kostenmodelle: Warum Planbarkeit wichtiger ist als Einstiegspreis

Kosten werden bei KI oft mit Einstiegskosten verglichen: Hardware gegen API-Preis. Relevant ist jedoch die Kostenentwicklung über Zeit.

Cloud-Modelle koppeln Kosten an Nutzung. Anfangs ist das niedrig und kalkulierbar, steigt jedoch mit Akzeptanz und Integration in Prozesse. Die Kosten folgen der Nutzung und damit dem Erfolg.

On-Prem-Modelle kehren diese Logik um: Der Großteil der Kosten entsteht vorab, danach bleibt die Nutzung weitgehend konstant. Je stärker das System verwendet wird, desto geringer werden die Kosten pro Anfrage.

Nutzung steigt	Cloud-KI	On-Prem-KI
Geringe Nutzung	Relativ günstig	Relativ teuer
Mittlere Nutzung	Ähnlich	Stabil
Hohe Nutzung	Stark steigend	Sinkend pro Anfrage

Die wirtschaftliche Bewertung hängt damit weniger vom Preis als vom erwarteten Einsatzgrad ab. Unsicherheit entsteht nicht durch absolute Höhe, sondern durch fehlende Vorhersagbarkeit.

Für Budgetverantwortliche ist daher entscheidend:

Wachsen die Kosten mit der Nutzung oder die Nutzung Kosten amortisiert.

10. Rollen & Know-how: Warum kein KI-Spezialteam nötig ist

Häufig wird angenommen, produktive KI erfordere neue organisatorische Einheiten. In der Praxis entstehen die meisten Probleme jedoch gerade durch Sonderstrukturen. Wird KI als separates Thema behandelt, fehlen klare Zuständigkeiten. Fachbereiche erwarten Ergebnisse, IT sieht Risiken und niemand verantwortet den laufenden Betrieb vollständig.

Stabil wird der Einsatz erst, wenn bestehende Rollen greifen:

- Management definiert Einsatzgrenzen
- IT betreibt das System wie andere Dienste
- Fachbereiche bewerten Ergebnisse fachlich

KI benötigt damit kein neues Expertenteam, sondern klare Zuordnung innerhalb bestehender Verantwortlichkeiten. Die notwendige Kompetenz liegt weniger im Modelltraining als im Umgang mit Ergebnissen. Akzeptanz entsteht nicht durch tiefes technisches Wissen, sondern durch nachvollziehbares Verhalten des Systems.

11. Projektlaufzeit & Quick Wins: Warum frühe Nutzung entscheidend ist

Lange Konzeptphasen erzeugen Erwartungen, aber keine Erfahrung. Produktiver Einsatz entsteht erst, wenn reale Arbeit mit dem System erfolgt. Deshalb beginnen erfolgreiche Einführungen mit begrenzten Anwendungsfällen. Nicht, um den Nutzen klein zu halten, sondern um reale Betriebsparameter zu erhalten: Antwortzeiten, Fehlerfälle und Akzeptanz.

Frühe Anwendungen sind bewusst einfach — Recherche, Zusammenfassungen, Strukturierung vorhandener Informationen. Sie verändern keine Prozesse grundlegend, zeigen aber unmittelbar Nutzen.

Wichtig ist die Reihenfolge:

Erst Nutzung → dann Ausweitung

Nicht umgekehrt. Akzeptanz entsteht aus Alltag, nicht aus Planung.

12. Use Cases: Wo KI im Alltag tatsächlich sinnvoll ist

Die entscheidende Frage für Unternehmen ist nicht, was mit KI möglich ist, sondern wo ihr Einsatz strukturell sinnvoll bleibt.

Geeignet ist KI für:

- Text verstehen und erzeugen
- Inhalte strukturieren
- Zusammenhänge herstellen
- Vorschläge formulieren

Besonders dort, wo Arbeit Zeit kostet (lesen, vergleichen, schreiben)

Weniger geeignet ist KI für:

- eindeutig richtige oder falsche Entscheidungen
- rechtlich bindende Prozesse

KI arbeitet probabilistisch – nicht deterministisch

Der größte Nutzen entsteht nicht durch neue Prozesse, sondern durch die Beschleunigung bestehender Wissensarbeit.

Startpunkt: Aufgaben, bei denen Ergebnisse überprüfbar und korrigierbar sind

13. Praxisbeispiel: Einführung einer On-Prem KI im Unternehmensalltag

In der Praxis zeigt sich: Das Hauptproblem ist nicht die Technik – sondern der Betrieb ohne Sonderstrukturen.

Typische Ausgangslage:

- sensible Daten
- kein dediziertes KI-Team
- IT im laufenden Betrieb ausgelastet

Vorgehen:

- Aufbau einer kleinen, stabilen Basisarchitektur
- Integration in bestehende Benutzerverwaltung
- Trennung von Test- und Produktivumgebung
- Fokus auf wenige, klar definierte Anwendungsfälle

Ergebnis:

- schnelle Nutzung in Fachbereichen
- geringer zusätzlicher Betriebsaufwand für die IT
- schrittweise Erweiterung über neue Use Cases

Die zentrale Erkenntnis:
Akzeptanz entsteht nicht durch Funktionsumfang, sondern durch Stabilität.

14. Warum Denkform in diesem Kontext eine Rolle spielt

An dem Punkt, an dem KI produktiv werden soll, entsteht selten ein Technologieproblem. Es entsteht ein Abstimmungsproblem zwischen Verantwortung, Betrieb und Nutzung.

Die **Geschäftsführung** bewertet Risiken und Investitionsdauer, die **IT** bewertet Wartbarkeit und **Fachbereiche** den praktischen Nutzen. KI funktioniert erst dann reibungslos, wenn diese Perspektiven zusammengeführt werden.

Die Rolle von Denkform liegt daher nicht primär in der Bereitstellung von Software oder Hardware, sondern in der Strukturierung der Entscheidung:

- Architektur
- Betriebsmodell
- organisatorische Zuständigkeiten

Diese Dinge so festgelegt, dass das System nach der Einführung ohne permanente externe Betreuung funktioniert.

Der Ansatz folgt dabei einem klaren Prinzip: Standardisieren, wo Stabilität wichtig ist und individualisieren, wo echter Mehrwert entsteht.

Ziel ist keine Abhängigkeit vom Anbieter zu haben, sondern ein System, das intern getragen werden kann.

15. Kontakt & Austausch

Dieses Dokument soll keine Entscheidung vorgeben, sondern Entscheidungsfähigkeit herstellen. In vielen Fällen entstehen die eigentlichen Fragen erst beim konkreten Abgleich mit der eigenen Organisation.

Ein Austausch dient daher meist der Einordnung: Welche Anforderungen bestehen tatsächlich, welche Annahmen treffen zu und wo liegen organisatorische statt technische Hürden.

Der nächste Schritt ist selten ein Projektstart, sondern Klarheit über den sinnvollen Rahmen.

DENKFORM

Thinking IT simple.

www.denkform.net

Denkform® GmbH

Zanggasse 6

65719 Hofheim am Taunus

0 611 711 85 7-0

hallo@denkform.net

USt-IdNr. DE 217 32 05 84

Steuernummer 04023118203

Sitz der Gesellschaft: Hofheim am Taunus

Amtsgericht: Frankfurt am Main HRB 142152

Bankverbindung:

Frankfurter Volksbank eG

IBAN : DE60 5019 0000 0026 2583 08

SWIFT-BIC: FFVBDEFF

Stand: März 2026