Der KI-Vorsprung

Der KI-Vorsprung

Wie Führung aus künstlicher Intelligenz echten Wettbewerbsvorteil macht — und woran die meisten scheitern

Dr. Sven Jungmann

EDITION AIOMICS

Erschienen in der **Edition aiomics**.

Autor und Herausgeber: Dr. Sven Jungmann
c/o aiomics GmbH, Rosenthaler Straße 72A, 10119 Berlin, Deutschland
Kontakt: communications@aiomics.io

1. Auflage 2026
ISBN (Taschenbuch): [ISBN]
ISBN (E-Book): [ISBN]

Umschlaggestaltung, Satz und Lektorat: Dr. Sven Jungmann

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über dnb.de abrufbar.

Herstellung und Vertrieb: Amazon Media EU S.à r.l., 38 avenue John F. Kennedy, L-1855 Luxemburg (Print on Demand).

Hinweis zur Nutzung. Dieses Buch vermittelt allgemeine Orientierung und Urteilshilfe zum unternehmerischen Einsatz künstlicher Intelligenz. Es ersetzt keine Rechts-, Steuer- oder Datenschutzberatung im Einzelfall und keine technische Prüfung der eigenen Systeme. Rechtslage und Technik verändern sich schnell; insbesondere zu Regulierung (etwa dem EU AI Act), zu Modellpreisen und zu Leistungsmerkmalen einzelner Anbieter sind die im Buch genannten Stände mit Jahresangabe versehen und vor jeder Entscheidung am aktuellen Stand zu prüfen. Autor und Herausgeber übernehmen keine Haftung für Entscheidungen, die auf Grundlage dieses Buches getroffen werden.

Für alle, die lieber verstehen als mitreden.

Inhalt

Vorwort

Versprechen, Lesehinweis und ein Pakt über Ehrlichkeit.

Ich baue künstliche Intelligenz dort, wo ein Fehler nicht nur Geld kostet: in Krankenhäusern. In meinem eigenen Unternehmen gebe ich zwischen 5.000 und 10.000 Euro im Monat für KI aus. Das ersetzt über 400.000 Euro im Jahr, die zuvor in zusätzliches Personal, freie Mitarbeit und Software geflossen sind — bei Ergebnissen, die mein Team und meine Kundschaft besser bewerten als zuvor. Diese Rechnung ist der Kern dieses Buches: Wie viel Geschäftswert Sie aus jeder Einheit Rechenleistung ziehen, entscheidet über Ihren Vorsprung.

Zugleich macht mich diese Nähe eher misstrauisch als euphorisch; ich sehe täglich, was die Technik kann und wo sie versagt. Wenn es Ihnen ähnlich geht, ist dieses Buch für Sie geschrieben.

Über kaum eine Technik ist in so kurzer Zeit so viel Unbrauchbares gesagt worden — Endzeit hier, bloßer Hype dort. Sie suchen, nehme ich an, genau deshalb ein Buch, das weder schwärmt noch abwinkt.

Das ist mein erstes Versprechen: Ich werde Ihnen keine Gewissheit verkaufen, die niemand hat. Wo die Zukunft offen ist, sage ich es. Wo eine Zahl unter Vorbehalt steht, nenne ich den Vorbehalt. Wo ich eine eigene Position beziehe, kennzeichne ich sie als das, was sie ist — eine begründete Wertung, kein Naturgesetz.

Mein zweites Versprechen betrifft die Tiefe. Die meisten Formate für Führungskräfte brechen genau dort ab, wo es

interessant wird. Sie erklären, *dass* es künstliche Intelligenz gibt, nicht, *wie* sie funktioniert und woran man gute von schlechter Umsetzung unterscheidet. Dieses Buch geht den Schritt weiter. Sie werden danach die Technik nicht selbst bauen. Aber Sie werden den Fachleuten, die sie bauen, die richtigen Fragen stellen können — und das ist der Unterschied zwischen einer Führungskraft, die Versprechen nachplappert, und einer, die urteilen kann.

Der Ehrlichkeits-Pakt

Nun das Offensichtliche, das ich Ihnen nicht verschweigen will, weil es zum Kern dieses Buches gehört. **Dieses Buch ist mit Hilfe von künstlicher Intelligenz geschrieben — und von einem Menschen verantwortet.**

Ich habe KI als das benutzt, wofür sie taugt: als unermüdlichen Rechercheur, als geduldigen Sparringspartner, als Schreibwerkzeug, das einen Gedanken in zwölf Fassungen vorlegt, von denen elf nichts taugen. Was sie nicht getan hat, ist das Eigentliche: entscheiden, was gut ist; prüfen, was stimmt; verwerfen, was nicht trägt; und am Ende mit meinem Namen geradestehen. Genau diese Trennung — die Maschine erzeugt, der Mensch urteilt — ist die zentrale Bewegung, die ich in diesem Buch beschreibe. Ich konnte sie schlecht predigen, ohne sie zu leben.

Darum dieser Pakt, und er gilt in beide Richtungen. Ich habe jede Kernbehauptung an eine nachprüfbare Quelle gebunden: an eine Studie, eine amtliche Statistik, einen benannten Urheber. Die Quellen stehen am Ende jedes Kapitels, mit Jahr und Fundstelle. Und Ihr Teil ist, nichts unbesehen zu nehmen. **Prüfen Sie nach.** Misstrauen ist hier die angemessene Haltung — gegenüber der Maschine, gegenüber jedem, der Ihnen KI verkaufen will, und ja, auch gegenüber diesem Buch. Ein KI-Buch, das Sie um Vertrauen bittet, hätte den ersten Test schon nicht bestanden.

Wie Sie dieses Buch lesen

Sie können es von vorne nach hinten lesen, und ich habe es so strukturiert, dass es sich lohnt: Jeder Teil trägt den nächsten. Wer mit der Technik schon vertraut ist, kann Teil II überfliegen und in Teil III einsteigen, wo der Vorsprung verhandelt wird. Wer es eilig hat, findet am Ende jedes Kapitels **„Das Wichtigste in 60 Sekunden"** und einen **Merksatz**; die wiederkehrenden Kästen — *Klartext*, *Stolperfalle*, *Die richtige Frage an Ihre Technik*, *Mach es heute* — sind Würze, kein Pflichtprogramm.

Eines noch. Dieses Buch verlangt Ihnen an einigen Stellen Konzentration ab, mehr, als ein Ratgeber aus der Flughafenbuchhandlung das täte. Das ist Absicht. Klarheit ist Respekt, und Respekt heißt hier: Ich erkläre Ihnen die Sache so genau, dass Sie danach selbst urteilen können — nicht so bequem, dass Sie es mir glauben müssen.

Fangen wir an.

Dr. Sven Jungmann

Einleitung

Worum es geht, für wen — und warum gerade jetzt.

Zwei Unternehmen derselben Branche, gleich groß, gleich gut geführt. Beide kaufen dieselben KI-Werkzeuge, zum selben Preis. Ein Jahr später hat das eine seine Abläufe spürbar verbessert, das andere eine teure Spielerei und einen Stapel enttäuschter Erwartungen. Die Technik war identisch. Den Unterschied habe ich oft genug gesehen, um zu wissen, dass er kein Zufall ist.

Dieses Buch handelt von diesem Unterschied.

Und es kommt zur rechten Zeit. In fast jeder Vorstandssitzung sitzt heute dieselbe Sorge mit am Tisch: dass die Konkurrenz schneller lernt, agiler umbaut, klüger anpasst — und davonzieht, während man selbst noch Pilotprojekte vergleicht. Die Sorge ist berechtigt, aber sie zielt meist daneben. Wer glaubt, es komme darauf an, KI als Erster einzuführen, hat die Aufgabe missverstanden: Die Werkzeuge bekommt die Konkurrenz zum selben Preis. Den Vorsprung nicht. Dieses Buch ist dafür da, dass Sie zu denen gehören, die ihn sich erarbeiten.

Der Grund liegt in einer neuen Ausgangslage. Künstliche Intelligenz ist zur Massenware geworden: Die leistungsfähigen Modelle stehen praktisch jedem offen, über eine Schnittstelle, zu Preisen, die Jahr für Jahr fallen. Der ökonomische Wert wandert deshalb dorthin, wo etwas knapp bleibt. Und knapp bleibt im KI-Zeitalter nicht die Technik, sondern dreierlei, das sich mit keiner Kreditkarte kaufen lässt.

Das ist das Gerüst des Buches, und ich nenne es **die drei Knappheiten**:

- **Urteil.** Wenn jede Firma denselben Output erzeugen kann, entscheidet die Fähigkeit, guten von schlechtem Output zu unterscheiden und das Richtige zu bauen. Die Rolle der Führung verschiebt sich vom Verwalter, der zusammenträgt, zum Beurteiler, der bewertet.
- **Kontext.** Das Modell ist gemietet und für alle gleich. Die eigenen, vertrauenswürdigen Daten und die nachprüfbaren Rückkopplungen aus dem eigenen Geschäft gehören nur dem eigenen Haus — ein Vorsprung, den ein gemietetes Modell niemandem schenkt und den ein Wettbewerber nicht nachkauft.
- **Wert je Recheneinheit.** Nicht wie viel Rechenleistung man einkauft, entscheidet, sondern wie viel messbaren Geschäftswert man je Einheit herausholt. Ich nenne diese Kennzahl **Return on Tokens**, und sie zieht sich durch das ganze Buch.

Die drei Knappheiten sind die Quelle des Vorsprungs. Sie bleiben aber bloßes Potenzial, solange das Unternehmen sie nicht hebt. Dafür braucht es den **Umbau**: die Organisation, die Aufsicht über die Technik, den menschlichen Übergang. Wer KI nur an bestehende Abläufe heftet, hat eine teure Anschaffung. Wer die Abläufe um die Technik herum neu denkt, baut einen Vorsprung. Daraus folgt die These, die dieses Buch trägt: **KI alleine verschafft noch keinen Vorsprung. Den Vorsprung schafft erst die Führung, die daraus etwas macht.**

Warum dieses Buch — und warum hier

Es gibt viele KI-Bücher. Fast alle sind aus dem Silicon Valley übersetzt: im Ton, in den Beispielen, in den unausgesprochenen Annahmen. Sie passen nicht auf einen Maschinenbauer aus Baden-Württemberg, eine Klinikgruppe in Wien, einen Versicherer in Zürich.

Der deutschsprachige Raum hat eine eigene Ausgangslage. Tiefe Fachkompetenz und Ingenieursdisziplin auf der einen Seite; träge Verbreitung neuer Technik, Fachkräftemangel, Regulie-

rungslast, hohe Energiekosten und eine ins Stocken geratene Produktivität auf der anderen. Für diese Lage ist künstliche Intelligenz mehr als ein weiterer Trend, den man mitnehmen sollte: Sie ist der zielgenaue Hebel gegen genau die Probleme, die den Standort bremsen — wenn man sie führt, statt sie zu erleiden. Genau das nimmt dieses Buch ernst, und darin unterscheidet es sich von den übersetzten.

Für wen das Buch geschrieben ist

Für die Entscheiderin und den Entscheider im deutschsprachigen Mittelstand und im gehobenen Unternehmenssegment, die ihr Fach beherrschen, die Technik aber nicht selbst bauen. Für jene, die den großen Versprechen misstrauen und zugleich fürchten, den Anschluss zu verlieren. Für die, die nicht *wirken* wollen, als hätten sie es verstanden, sondern es *wirklich* verstehen wollen — weil Halbwissen vor den jüngeren, digital-affinen Mitarbeitenden im eigenen Haus nicht lange standhält.

Wenn Sie den Lärm satthaben und Substanz suchen, sind Sie hier richtig.

Was das Buch nicht ist

Kein Hype-Buch. Kein Werkzeug-Katalog, der in einem halben Jahr veraltet ist. Kein Ingenieurslehrbuch. Keine Prognose-Akrobatik, die Ihnen eine Zahl mit Gewissheit nennt, die niemand hat.

Der Weg durch das Buch

Teil I — Die Lage. Warum gerade jetzt und warum gerade hier; und warum der bloße Einsatz von KI noch keinen Vorsprung bringt.

Teil II — KI verstehen. Der Kompetenzteil. Wie ein Sprachmodell wirklich arbeitet und warum es halluziniert; wie aus einem Rohmodell *Ihr* Modell wird; was ein Agent ist und wo seine Risiken liegen; woran Sie erkennen, ob etwas funktioniert

(Evaluation); und was offene gegen geschlossene Modelle für Ihre Entscheidung bedeuten. Tief genug für die richtigen Fragen, nicht tief genug, um Sie zu langweilen.

Teil III — Die drei Knappheiten. Wo der Vorsprung entsteht: Urteil, Kontext, Wert je Recheneinheit.

Teil IV — Der Umbau. Wie aus Knappheit Rendite wird: die Organisation neu denken, Governance und Sicherheit, der menschliche Übergang — und ein Fahrplan für die ersten neunzig Tage.

Sie kommen nicht zu spät, und Sie kommen nicht zu früh. Sie kommen an die Stelle, an der die Werkzeuge erprobt und bezahlbar sind, der Vorsprung aber noch nicht verteilt ist. Was Sie daraus machen, ist Führungsarbeit. Sie beginnt auf der nächsten Seite.

TEIL I — DIE LAGE

Kapitel 1 — Warum jetzt und warum hier

Lernziele Nach diesem Kapitel können Sie

- die wirtschaftliche Lage des deutschsprachigen Raums benennen, auf die künstliche Intelligenz eine Antwort ist — und warum diese Antwort gerade hier zählt;
- die Begeisterung der einen und die Abwehr der anderen einordnen und eine eigene, nüchterne Position dazwischen beziehen;
- erkennen, warum der Einsatz einer Technik allein noch keinen Vorsprung schafft;
- neue Entwicklungen aus ersten Prinzipien prüfen, statt aus Schlagzeilen.

Deutschland gehen die Menschen aus, nicht die Arbeit. Das ist der nüchterne Befund, mit dem jede ehrliche Strategie für die kommenden Jahre beginnen muss. Die geburtenstarken Jahrgänge gehen in den Ruhestand, und es rücken zu wenige nach. Das Institut für Arbeitsmarkt- und Berufsforschung rechnet damit, dass das Erwerbspersonenpotenzial bis 2035 um mehrere Millionen Menschen schrumpfen könnte, im Szenario schwacher Zuwanderung um bis zu sieben Millionen, also rund ein Sechstel der heutigen Erwerbsbasis.[1] Die genaue Zahl hängt an

Annahmen über Migration und Erwerbsbeteiligung; die Richtung hängt an niemandem mehr. Sie ist demografisch festgelegt, die betreffenden Menschen sind längst geboren.

Ein Unternehmen, dem die Mitarbeiter ausgehen, hat zwei Möglichkeiten. Es kann um eine schrumpfende Zahl von Fachkräften konkurrieren, mit steigenden Löhnen und sinkender Auswahl. Oder es kann dafür sorgen, dass jede verbliebene Fachkraft mehr ausrichtet. Die zweite Möglichkeit ist im Kern eine Produktivitätsfrage — und Produktivität ist genau das, woran es im deutschsprachigen Raum gerade fehlt.

Der Befund: eine Wirtschaft, die langsamer wird

Über Jahrzehnte galt der Standort als robust: Ingenieurskunst, Exportstärke, ein tiefer industrieller Mittelstand. Dieses Bild ist in den vergangenen Jahren brüchig geworden, und die Zahlen sind eindeutig genug, dass man sie nicht wegreden sollte.

Die deutsche Wirtschaft ist 2023 und 2024 geschrumpft — zwei Jahre in Folge, was unter den großen Industrieländern eine Ausnahme war.[2] Der Internationale Währungsfonds beziffert das **Potenzialwachstum** — also das Tempo, das die Wirtschaft bei normaler Auslastung überhaupt halten kann — auf nur noch etwa 0,7 Prozent im Jahr, getragen von schwacher Produktivität und einer schrumpfenden Erwerbsbevölkerung.[3] Das Etikett vom „kranken Mann Europas", in den neunziger Jahren schon einmal verteilt, ist zurück.

Dazu kommt eine hausgemachte Last. Das ifo-Institut hat 2024 errechnet, dass überbordende Bürokratie die deutsche Wirtschaft jährlich rund 146 Milliarden Euro an entgangener Wertschöpfung kostet — eine Modellrechnung, die Deutschland an einem regulatorisch schlankeren Vergleichsland misst und auch indirekte Effekte einschließt; andere Schätzungen, die nur die direkten Befolgungskosten zählen, liegen niedriger, bei rund 65 Milliarden.[4] Wie man die Zahl auch fasst, der zweite Befund derselben Studie ist der für uns interessantere: Würde der deutsche Staat seine Verwaltung nur auf das Digitalisierungsniveau

Dänemarks heben, ließen sich davon etwa 96 Milliarden Euro im Jahr zurückgewinnen.[4] Ein erheblicher Teil des Problems ist kein Schicksal, sondern ein Prozessproblem — und Prozessprobleme sind angreifbar.

Zahlen, die zählen

- Reales Wachstum 2023 und 2024: **rückläufig** — zwei Schrumpfungsjahre in Folge.[2]
- Potenzialwachstum: **rund 0,7 % pro Jahr** (IWF).[3]
- Erwerbspersonenpotenzial bis 2035: **bis zu –7 Mio. (rund –16 %)** im Szenario schwacher Zuwanderung (IAB/Destatis).[1]
- Bürokratiekosten: **rund 146 Mrd. €/Jahr** (ifo, Modellrechnung); **+96 Mrd. €/Jahr** Potenzial bei dänischem Digitalisierungsniveau der Verwaltung.[4]

Warum gerade KI — und warum gerade hier

Hier treffen sich zwei Linien. Auf der einen Seite eine Technik, deren Stärke darin liegt, geistige Routinearbeit zu übernehmen und Wissen zu vervielfältigen. Auf der anderen Seite eine Volkswirtschaft, der die Arbeitskräfte ausgehen, deren Produktivität stockt und die einen Gutteil ihrer Kraft in Verwaltungsvorgängen verliert. Künstliche Intelligenz ist für den deutschsprachigen Raum deshalb nicht einfach ein weiterer internationaler Trend, den man mitnehmen sollte. Sie setzt genau an den Stellen an, an denen es weh tut: Sie ergänzt knappe Arbeit, statt sie zu ersetzen, und sie greift jene prozesslastige, papierne Wertschöpfung an, die anderswo längst schlanker läuft.

Das ist das erste Argument dieses Buches, und es ist ein lokales: Wer in Wien, Zürich oder Stuttgart führt, hat einen *dringenderen* Grund, sich mit KI ernsthaft zu befassen, als ein Wettbewerber in einer jungen, schnell wachsenden Volkswirtschaft. Nicht, weil die Technik hier besser wirkte, sondern weil die Krankheit, gegen die sie hilft, hier ausgeprägter ist.

Es gibt nur einen Haken, und er ist unangenehm. Genau die Regionen mit dem größten Bedarf zählen zu den langsamsten bei der Einführung. Der deutsche Mittelstand — über 99 Prozent aller Unternehmen, gut die Hälfte aller Arbeitsplätze[5] — nutzt KI deutlich seltener als die Großindustrie, und wo er sie nutzt, geschieht es meist ungesteuert. Je nach Erhebung und Definition setzen zwischen einem Viertel und gut zwei Fünfteln der Unternehmen KI ein; der Anteil derer, die sie *flächendeckend und mit Strategie* einsetzen, ist um ein Vielfaches kleiner.[6] Höchster Bedarf, langsamste Aufnahme: in dieser Spannung steckt die Aufgabe der kommenden Jahre, und zugleich die Chance. Wer sie früher auflöst als der Wettbewerb, beherrscht die Werkzeuge früher. Genau darin liegt sein Vorsprung, denn die Werkzeuge selbst stehen allen offen.

> **Aus der Praxis — Als die Anfänger die Veteranen überholten** In einem Unternehmen mit 5.179 Mitarbeitern im Kundendienst führten Forscher schrittweise einen KI-Assistenten ein und maßen, was geschah. Im Schnitt löste das Personal 14 Prozent mehr Anfragen pro Stunde. Aufschlussreich aber war die Verteilung: Die *unerfahrenen* Kräfte wurden um rund 34 Prozent produktiver, die erfahrensten kaum.[7] Der Assistent hatte das stille Wissen der besten Mitarbeiter eingefangen und es allen anderen verfügbar gemacht. Für einen Mittelständler, dem gerade die erfahrensten Köpfe in den Ruhestand gehen, ist das die vielleicht greifbarste Verheißung der Technik: Sie kann Erfahrung konservieren, die sonst aus der Tür ginge.

Zwischen Begeisterung und Abwehr

Über kaum eine Technik wird so unbrauchbar gesprochen wie über diese. Auf der einen Seite stehen jene, für die KI alles verändert, jede Demonstration ein Wunder und jeder Zweifel ein Zeichen von Rückständigkeit ist. Auf der anderen Seite jene, die in ihr vor allem Hochstapelei sehen, einen Hype, der vorübergeht wie andere zuvor. Beide Lager ersparen sich das Nachdenken:

das eine durch Begeisterung, das andere durch Verachtung. Eine Führungskraft kann sich keines von beidem leisten.

Die belastbare Lage ist unbequemer und interessanter. Auf der Ebene einzelner Aufgaben ist der Nutzen real, in sorgfältigen Studien wiederholt gemessen: Beim Schreiben anspruchsvoller Texte erledigten Berufstätige mit KI dieselbe Aufgabe in rund 40 Prozent weniger Zeit — und lieferten zugleich bessere Qualität.[8] Programmierer erledigten eine umrissene Aufgabe mit einem KI-Assistenten gut die Hälfte schneller.[9] Unternehmensberater steigerten bei Aufgaben *innerhalb* der Fähigkeiten des Modells Qualität und Tempo deutlich.[10] Das ist kein Marketing, das sind kontrollierte Experimente, und ihr Muster ist stabil: Die größten Gewinne entstehen bei klar umrissener, eher routinehafter Arbeit, und die Schwächeren holen am meisten auf.

Doch dieselbe Berater-Studie zeigt die Kehrseite. Bei einer Aufgabe, die bewusst *außerhalb* der Fähigkeiten des Modells lag, lagen die Nutzer der KI rund 19 Prozentpunkte häufiger falsch als die Kollegen ohne — selbstbewusst falsch.[10] Die Gefahr der Technik liegt nicht nur darin, was sie nicht kann, sondern darin, wo sie überzeugend daneben liegt und der Mensch ihr glaubt.

> **Stolperfalle — Der überzeugende Irrtum** Das größte Risiko ist nicht die offensichtlich falsche Antwort. Es ist die plausible, flüssig formulierte, falsche Antwort bei einer Aufgabe, die knapp jenseits dessen liegt, was das Modell beherrscht — und der man deshalb vertraut. Die Beraterstudie hat genau das gemessen: mehr Fehler *mit* KI als ohne, sobald die Aufgabe die Fähigkeiten des Modells überstieg. Die Konsequenz ist keine Abkehr, sondern eine Pflicht: zu wissen, wo die zuverlässige Zone endet — und dort die Prüfung nicht der Maschine zu überlassen.

Auch über die *Größe* des Ganzen wird gestritten, und der Streit ist lehrreich. Der Ökonom Daron Acemoglu rechnet kühl: Höchstens ein Fünftel der Arbeitsaufgaben sei überhaupt für KI zugänglich, der gesamtwirtschaftliche Produktivitätsgewinn lie-

ge über zehn Jahre vielleicht bei einem halben Prozentpunkt.[11] Investmentbanken wie Goldman Sachs kamen zur selben Zeit auf ein Vielfaches — mehrere Prozent zusätzliches Welt-Bruttoinlandsprodukt.[12] Der Abstand zwischen beiden ist gewaltig, und er entsteht nicht aus unterschiedlichen Messungen, sondern aus unterschiedlichen Annahmen darüber, wie viel Arbeit die Technik am Ende berührt und zu welchen Kosten. Niemand kennt die Antwort. Wer Ihnen eine Zahl mit Gewissheit nennt, verkauft Ihnen etwas.

Mein Standpunkt, den ich in den nächsten Kapiteln begründe: Beide können recht behalten. Große, echte Gewinne bei einzelnen Tätigkeiten und zugleich für einige Jahre ein bescheidener, in der Statistik kaum sichtbarer Effekt auf die Gesamtwirtschaft — das ist kein Widerspruch, sondern das erwartbare Muster jeder großen Basistechnologie. Warum das so ist, klärt das nächste Kapitel.

> **Klartext — Allzwecktechnologie** Eine *Allzwecktechnologie* (auch Basis- oder Mehrzwecktechnologie) ist keine Lösung für ein Problem, sondern eine Grundlage für viele — wie die Dampfmaschine, die Elektrizität oder der Computer. Sie wirkt nicht durch sich selbst, sondern dadurch, dass ganze Branchen ihre Arbeitsweise um sie herum neu erfinden. Genau deshalb kommt ihr Nutzen verzögert — und ungleich verteilt.

Warum der Einsatz allein noch nichts bringt

Es lohnt ein Blick zurück, weil er die Gegenwart erklärt. Als die Fabriken um 1900 vom Dampf auf den Elektromotor umstellten, geschah zunächst fast nichts. Die Produktivität bewegte sich kaum. Der Wirtschaftshistoriker Paul David hat beschrieben, warum: Die Fabrikbesitzer ersetzten lediglich die zentrale Dampfmaschine durch einen zentralen Elektromotor und ließen alles andere, wie es war — die Wellen unter der Decke, die Riemen, die Anordnung der Maschinen.[13] Sie hatten die neue Technik an die alte Fabrik geheftet. Der Sprung kam erst, als je-

mand jeder Maschine ihren eigenen kleinen Motor gab und den Fabrikboden neu zeichnete: ebenerdig, flexibel, am Materialfluss ausgerichtet. Zwischen der Verfügbarkeit der Technik und ihrem Ertrag lagen Jahrzehnte. Es fehlte nicht die Technik, es fehlte die *Reorganisation*.

Das ist die zweite tragende Einsicht, und sie beantwortet zugleich, warum so viele KI-Projekte verpuffen. Untersuchungen und Praxisberichte deuten übereinstimmend in eine Richtung: Eine große Mehrheit der KI-Vorhaben in Unternehmen erreicht keinen messbaren Wertbeitrag. Die kursierenden Zahlen — 80 Prozent, 95 Prozent — beruhen auf kleinen oder qualitativen Stichproben und sind als exakte Werte mit Vorsicht zu genießen; als Größenordnung aber sind sie über mehrere unabhängige Quellen hinweg konsistent und decken sich mit allem, was wir über frühere Basistechnologien wissen.[14] Der Grund ist fast nie das Modell. Er ist organisatorisch: das falsch verstandene Problem, die fehlende Datengrundlage, die Erwartung, die Technik werde wirken, wenn man sie nur an den bestehenden Ablauf klemmt.

Daraus folgt die These, die dieses Buch trägt:

> **KI alleine verschafft noch keinen Vorsprung. Den Vorsprung schafft erst die Führung, die daraus etwas macht.**

Das Werkzeug ist für jeden käuflich, zu fallenden Preisen. Was der Wettbewerber ebenso kaufen kann, ist kein Vorsprung. Knapp, und damit wertvoll, bleibt, was sich nicht kaufen lässt: das Urteil, gute von schlechter Arbeit zu unterscheiden; der eigene, vertrauenswürdige Kontext, an dem die Maschine besser wird als bei jedem anderen; und die Disziplin, aus Rechenleistung messbaren Wert zu ziehen. Diese drei Knappheiten, und der Umbau, der sie in Rendite verwandelt, sind der Stoff der folgenden Teile.

Wie man hinsieht: aus ersten Prinzipien

Wenn niemand die Landkarte der nächsten fünf Jahre besitzt, braucht man wenigstens einen verlässlichen Kompass. Der beste, den ich kenne, ist eine schlichte Denkdisziplin: aus ersten Prinzipien zu urteilen statt aus Analogien.

Aus Analogien zu denken heißt, das Neue mit einem Etikett des Alten zu versehen: „KI ist wie das Internet", „wie die Dampfmaschine", „wie die Atombombe", und dann anzunehmen, es werde sich verhalten wie das, womit man es verglichen hat. Das ist bequem und oft irreführend, denn der Vergleich überspringt gerade die Stellen, an denen das Neue eben nicht ist wie das Alte. Der Analyst Benedict Evans hat dafür ein zugespitztes Bild geprägt: Wir seien bei der KI „im Jahr 1997", nicht 2007: das meiste, was zählen wird, sei noch nicht gebaut, und die heutigen Marktführer seien selten die von morgen.[15] An dieser Mahnung zur Demut ist viel richtig, und ich übernehme sie. Seine zweite These aber teile ich nicht: KI sei „so groß wie das Internet, und nur so groß". Das Internet bewegte vor allem Information; KI verrichtet geistige Arbeit. Damit ist ihre mögliche Reichweite eine andere, und wer ihr von vornherein eine Obergrenze zuschreibt, urteilt aus einer Analogie, nicht aus der Sache. Genau das ist der Fehler, den die folgende Disziplin vermeidet.

Aus ersten Prinzipien zu urteilen heißt, sich vier Fragen der Reihe nach zu stellen:

1. **Was beobachte ich wirklich?** — die nackte Tatsache, befreit von der Erzählung darum herum.
2. **Na und?** — folgt daraus überhaupt etwas, das für mein Geschäft zählt?
3. **Was geschieht als Nächstes,** wenn sich das so fortsetzt?
4. **Was müsste zutreffen,** damit meine Schlussfolgerung hält — und wie billig kann ich das prüfen?

Diese vier Fragen kehren im ganzen Buch wieder. Sie sind das Gegengift gegen beide Krankheiten der Stunde, die Schwärmerei und die Abwehr, und sie sind in der gegenwärtigen Lage wertvol-

ler als jede Prognose. Denn die Prognose verspricht eine Gewissheit, die niemand hat. Das Urteil aus ersten Prinzipien verlangt nur, dass Sie genau hinsehen und Ihre Schlüsse billig überprüfen. Und das können Sie ab heute.

Was dieses Kapitel für Sie ändert

Sie kommen nicht zu spät, und Sie kommen nicht zu früh. Sie kommen an die Stelle, an der die Werkzeuge erprobt und bezahlbar sind, der Vorsprung aber noch nicht verteilt ist, und an der die Lage des eigenen Standorts den Einsatz besonders dringlich macht. Die Haltung, die dazu passt, ist weder die des Schwärmers noch die des Skeptikers. Es ist die des nüchternen Frühanwenders: jemand, der die großen Versprechen nicht glaubt, die kleinen, billigen Versuche aber sofort beginnt, der weiß, dass der Ertrag aus dem Umbau kommt und nicht aus dem Kauf, und der lieber sein Urteil schärft, als auf eine Gewissheit zu warten, die in dieser Phase niemand liefert.

Bevor wir zum Umbau kommen, ist zweierlei zu klären. Das nächste Kapitel zeigt, warum der bloße Einsatz von KI noch keinen Vorsprung bringt. Der zweite Teil macht Sie dann mit den Werkzeugen selbst vertraut: nicht so tief, dass Sie zum Ingenieur werden, aber tief genug für die richtigen Fragen.

Das Wichtigste in 60 Sekunden

- Dem deutschsprachigen Raum gehen die Arbeitskräfte aus, die Produktivität stockt, und viel Kraft versickert in Bürokratie. KI setzt genau hier an — sie ergänzt knappe Arbeit und greift prozesslastige Wertschöpfung an. **Der Bedarf ist hier größer als anderswo.**
- Zugleich nimmt gerade der Mittelstand die Technik am langsamsten auf. **Höchster Bedarf, langsamste Aufnahme:** darin liegt die Chance für den, der früher handelt.
- Der Nutzen ist auf Aufgabenebene real und gemessen (Tempo- und Qualitätsgewinne von 14 bis über 50 Prozent), aber gefährlich, wo die Aufgabe die Fähigkeiten des Modells übersteigt und der Mensch ihm trotzdem glaubt.
- Über die gesamtwirtschaftliche Größe wird seriös gestritten; die Wahrheit liegt vermutlich im verzögerten Muster jeder Basistechnologie. **Wer Gewissheit verkauft, verkauft Ihnen etwas.**
- **Der Einsatz allein bringt nichts; der Ertrag kommt aus dem Umbau.** Urteilen Sie aus ersten Prinzipien, nicht aus Analogien.

Merksatz Den Vorsprung schafft nicht, wer KI zuerst kauft, sondern wer die Arbeit zuerst um sie herum neu denkt.

Drei Fragen an Ihr Unternehmen

1. An welchen Stellen Ihres Geschäfts schlägt der Fachkräftemangel schon heute durch — und wo könnte KI eine knappe Fähigkeit ergänzen, statt sie zu ersetzen?
2. Wo haben Sie eine neue Technik bislang nur an einen alten Ablauf geheftet, statt den Ablauf neu zu denken?
3. Welche KI-Behauptung, die in Ihrem Haus kursiert, könnten Sie mit den vier Fragen aus ersten Prinzipien in einer Stunde überprüfen?

Quellen

1. Projektion des Erwerbspersonenpotenzials: Institut für Arbeitsmarkt- und Berufsforschung (IAB) auf Basis der Bevölkerungsvorausberechnung des Sta-

tistischen Bundesamts. Der Wert von rund –7 Mio. (≈ –16 %) bis 2035 gilt für ein Szenario schwacher Zuwanderung; bei stärkerer Zuwanderung/Erwerbsbeteiligung fällt der Rückgang geringer aus. IAB-Forum, „Effects of population changes on the labour market in Germany"; IAB-Themenseite Fachkräftemangel. (*Vor Druck: exakte Szenariowerte und Stichjahr gegen die 15. koordinierte Bevölkerungsvorausberechnung von Destatis prüfen.*)

2. Reales BIP Deutschland 2023 und 2024 rückläufig (zwei Schrumpfungsjahre in Folge); Quelle: Statistisches Bundesamt, Volkswirtschaftliche Gesamtrechnungen. (*Vor Druck: exakte Veränderungsraten der jeweils jüngsten Destatis-Veröffentlichung entnehmen.*)
3. Internationaler Währungsfonds, Artikel-IV-Konsultation Deutschland 2024 (Country Report 24/229): Potenzialwachstum von rund 0,7 % p. a.
4. ifo-Institut, Pressemitteilung vom 14. November 2024: Bürokratie kostet Deutschland rund 146 Mrd. € jährlich an entgangener Wertschöpfung (Modellrechnung im Vergleich zu Schweden, inkl. indirekter Effekte; im Auftrag der IHK für München und Oberbayern); zusätzliches Potenzial von rund 96 Mrd. € bei einem Digitalisierungsniveau der Verwaltung wie in Dänemark. Engere Schätzungen direkter Befolgungskosten (Nationaler Normenkontrollrat) liegen bei rund 65 Mrd. €.
5. Institut für Mittelstandsforschung (IfM) Bonn: über 99 % aller Unternehmen sind KMU; rund 53 % aller sozialversicherungspflichtig Beschäftigten arbeiten in KMU (EU-Definition, Bezugsjahr 2023).
6. Spannbreite der KI-Nutzung deutscher Unternehmen je nach Erhebung und Definition: Bitkom (Befragung ab 20 Beschäftigte) und ifo-Konjunkturumfragen 2025 weisen zwischen rund einem Viertel und gut zwei Fünfteln aus; der Anteil flächendeckenden, strategisch gesteuerten Einsatzes ist deutlich kleiner (Bitkom 2025: 3 % flächendeckender Einsatz generativer KI). Die Werte verschiedener Erhebungen sind wegen abweichender Methodik nicht direkt vergleichbar; für EU-Ländervergleiche eignet sich die harmonisierte Eurostat-Reihe „KI-Einsatz in Unternehmen".
7. Brynjolfsson, Li & Raymond, „Generative AI at Work", NBER Working Paper 31161 (2023): +14 % gelöste Anfragen pro Stunde im Schnitt, +34 % bei unerfahrenen Kräften, nahezu kein Effekt bei den erfahrensten.
8. Noy & Zhang, „Experimental evidence on the productivity effects of generative artificial intelligence", *Science* (2023): rund 40 % weniger Bearbeitungszeit bei zugleich höherer Qualität (rund +18 %); geringere Streuung zwischen den Teilnehmern.
9. Peng et al., GitHub/MIT, arXiv:2302.06590 (2023): rund 56 % schnellere Erledigung einer umrissenen Programmieraufgabe mit KI-Assistenz.
10. Dell'Acqua et al., „Navigating the Jagged Technological Frontier", Harvard Business School Working Paper 24-013 (2023; publiziert in *Organization Science* 2025): innerhalb der Modellfähigkeiten rund +40 % Qualität und höheres Tempo; bei einer bewusst außerhalb platzierten Aufgabe rund 19 Prozentpunkte häufiger falsche Ergebnisse.

11. Acemoglu, „The Simple Macroeconomics of AI", NBER Working Paper 32487 (2024; *Economic Policy* 2024): rund 20 % der Aufgaben KI-exponiert; gesamtwirtschaftlicher TFP-Effekt über zehn Jahre ≤ ~0,66 %.
12. Goldman Sachs (2023): geschätzt rund +7 % zusätzliches globales BIP und rund +1,5 Prozentpunkte US-Produktivität pro Jahr über ein Jahrzehnt. (*Vor Druck: Primärbericht von Goldman Sachs Research direkt zitieren.*)
13. Paul A. David, „The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox", *American Economic Review* 80(2), 1990.
14. Zur Größenordnung gescheiterter KI-Vorhaben: RAND, „The Root Causes of Failure for AI Projects" (2024; >80 %, beruhend auf qualitativen Experteninterviews) und MIT-Projekt NANDA, „The GenAI Divide: State of AI in Business 2025" (rund 95 % der GenAI-Pilotprojekte ohne messbaren Ergebnisbeitrag). Beide Zahlen sind als exakte Werte umstritten (kleine bzw. qualitative Stichproben, uneinheitliche Definitionen); als Größenordnung sind sie über unabhängige Quellen konsistent. Als repräsentativere Quellen für Verbreitungs- und Wertschöpfungslücke eignen sich McKinsey „State of AI" und die US-Erhebungen von MIT/Census.
15. Benedict Evans im Gespräch, Lenny's Podcast, „A rational conversation on where AI is actually going" (31. Mai 2026). Die Aussagen „We are in 1997, not 2007" und „as big as the internet — and only as big" sind Evans' analytische Positionen, keine gemessenen Tatsachen; die zweite wird hier ausdrücklich nicht geteilt.

Kapitel 2 — Nutzung ist kein Vorsprung

Lernziele Nach diesem Kapitel können Sie

- erklären, warum eine fast flächendeckende KI-Nutzung mit einem fast ausbleibenden Ertrag einhergeht;
- den Mechanismus benennen, der über den Ertrag entscheidet — die Reorganisation der Arbeit, nicht das Werkzeug;
- die organisatorischen Gründe erkennen, an denen die meisten KI-Vorhaben scheitern;
- einschätzen, warum gerade Unternehmen im deutschsprachigen Raum zum Anbau statt zum Umbau neigen, und was daraus für Ihre eigene Entscheidung folgt.

Fast alle nutzen inzwischen künstliche Intelligenz. Fast niemand verdient daran.

Die Unternehmensberatung McKinsey befragt jedes Jahr Unternehmen weltweit zu ihrem KI-Einsatz. In der jüngsten Erhebung gaben rund 88 Prozent an, KI in mindestens einem Geschäftsbereich zu verwenden. Einen messbaren Effekt auf das Betriebsergebnis meldeten nur etwa 39 Prozent — und einen *spürbaren* Effekt, also mehr als fünf Prozent des Betriebsergebnisses, gerade einmal sechs Prozent.[1] Eine Technik ist beinahe überall, und sie zahlt sich beinahe nirgends aus.

Diese Befragung beruht auf Unternehmen, die freiwillig antworten und sich überdurchschnittlich für das Thema interessieren; die wahre Nutzungsbreite liegt niedriger. Die statistischen Ämter der USA messen sie repräsentativ über mehr als eine Million Betriebe und kommen für 2026 auf einen KI-Einsatz von rund 17 Prozent — von etwa 5 Prozent Anfang 2024 stetig gestiegen, aber noch immer eine Minderheit.[2] Damit liegen zwei Lücken offen: die zwischen dem Reden über KI und ihrer tatsächlichen Nutzung, und, die teurere, die zwischen Nutzung und Ertrag.

Dieses Kapitel handelt von der zweiten Lücke. Sie ist kein Zeichen dafür, dass die Technik nicht taugt. Sie ist das erwartbare Muster jeder Basistechnologie, und wer es versteht, hat den Hebel des ganzen Buches in der Hand.

Eine Technik allein bewegt nichts

Im ersten Kapitel stand die Fabrik, die ihre Motoren nicht versetzte: Jahrzehnte hatte die Elektrizität bereitgestanden, ehe sie die Produktivität hob — und sie tat es erst, als die Werkshallen um sie herum neu gebaut wurden. Dasselbe Muster zeigt der vielleicht berühmteste Produktivitätssprung der Industriegeschichte, und er ist lehrreich, weil an ihm gar keine neue Maschine beteiligt war.

> **Aus der Praxis — Zwölf Stunden, dann neunzig Minuten** 1913 brauchte Ford im Werk Highland Park rund zwölfeinhalb Stunden, um ein Fahrgestell des Modell T zu montieren. Henry Ford kaufte keine Wundermaschine. Er ließ das Auto am Arbeiter vorbeifahren, statt den Arbeiter um das Auto herumlaufen zu lassen. Binnen Monaten sank die Montagezeit auf etwa 93 Minuten.[3] Der Gewinn steckte nicht in einem Gerät, sondern in der Neuordnung des Arbeitsflusses. Die Technik der Einzelteile blieb dieselbe; neu war, wie die Arbeit organisiert war.

Was hier sichtbar wird, hat die Wirtschaftsforschung über zwei Jahrzehnte vor dem ersten Sprachmodell vermessen. Die Ökonomen Timothy Bresnahan, Erik Brynjolfsson und Lorin Hitt zeigten 2002 an Firmendaten, dass Informationstechnik, die Neuordnung der Arbeit und neue Qualifikationen ein **System sich gegenseitig verstärkender Bausteine** bilden. Wer nur einen davon einführt — die Technik, aber nicht die Organisation —, erntet wenig oder nichts; erst das Bündel zahlt sich aus.[4] Der Computer war nie die Ursache des Gewinns. Er war ein Baustein in einem Bündel, dessen wertvollster Teil unsichtbar bleibt.

Wie unsichtbar, lässt sich beziffern. Dieselbe Forschergruppe fand, dass ein zusätzlicher Dollar an Computerkapital im Börsenwert eines Unternehmens mit rund zehn Dollar zu Buche schlägt — während gewöhnliches Sachkapital ungefähr eins zu eins bewertet wird.[5] Die fehlenden neun Dollar messen den Marktwert all dessen, was die Technik erst nützlich macht und in keiner Bilanz steht: die neu zugeschnittenen Abläufe, die geschulten Menschen, die geordneten Daten, die veränderte Aufbauorganisation. Spätere Arbeiten haben gezeigt, dass ein Teil davon schlicht nicht erfasste Software und Ausbildung ist.[6] Die Lehre bleibt: Der wertvolle Teil der Investition ist meist der unsichtbare, und er hängt nur lose an dem, was man an Technik einkauft.

> **Klartext — Organisationskapital** *Organisationskapital* ist der Wert, der in der Art steckt, wie ein Unternehmen arbeitet: in seinen Abläufen, Routinen, im Können seiner Leute, in der Ordnung seiner Daten und seiner Struktur. Man kann es nicht kaufen und es steht in keiner Bilanz, aber es entscheidet darüber, ob eine Technik Ertrag bringt. Es ist firmenspezifisch, in den Köpfen verankert und nur langsam aufzubauen. Genau deshalb ist es ein Vorteil, den ein Wettbewerber nicht über Nacht nachbilden kann.

Hier liegt die erste, oft übersehene Folgerung. Wenn der Ertrag im Organisationskapital steckt und nicht im Werkzeug,

dann ist das Werkzeug auch kein Vorsprung. Vom Wettbewerber trennt Sie nicht das Modell, das er morgen ebenso mieten kann, sondern das Bündel aus neu gedachten Abläufen, geschulten Menschen und geordneten Daten, das nur in Ihrem Haus existiert. Ein dauerhafter Vorteil ist immer einer, der sich schwer kopieren lässt.

Warum die meisten Vorhaben im Sande verlaufen

Wenn der Ertrag aus der Reorganisation kommt, dann müsste man dort scheitern sehen, wo sie ausbleibt. Genau das zeigt sich. Eine Untersuchung der RAND Corporation befragte erfahrene Praktiker nach den Gründen, an denen KI-Projekte scheitern, und fand fünf — keiner davon technisch.[7] An erster Stelle steht das missverstandene Ziel: Führung und Fachleute sind sich nicht einig, welches Problem überhaupt gelöst werden soll. Es folgen fehlende Daten, das Denken vom Werkzeug her statt vom Problem, mangelnde Infrastruktur und schließlich Probleme, für die KI gar nicht taugt. Schuld ist fast nie das Modell, sondern die Organisation darum herum.

Das gängigste Muster dieses Scheiterns trägt einen Namen, den man sich merken sollte: Pilotitis. Ein Pilotprojekt gelingt als Vorführung, beeindruckt im Lenkungskreis — und erreicht nie den Wirkbetrieb. Das Beratungshaus Gartner sagte voraus, dass bis Ende 2025 mindestens 30 Prozent der Vorhaben mit generativer KI nach der Machbarkeitsstudie wieder eingestellt würden.[8] Der häufigste Grund ist organisatorisch: Niemand ist für das *Ergebnis* verantwortlich. Die Datenfachleute besitzen das Experiment; den produktiven Nutzen besitzt niemand.

Stolperfalle — Den schlechten Ablauf nur schneller machen Der teuerste Fehler ist, KI an einen Ablauf zu heften, den man vorher nicht hinterfragt hat. Der Manager Thorsten Dirks brachte es auf eine Formel, die in deutschen Vorstandsetagen hängengeblieben ist: Wer einen schlechten Prozess digitalisiert, hat danach einen schlechten digitalen Prozess.[9] Schneller am Ziel vorbei ist nicht näher am Ziel. Bevor Sie ein Modell auf einen Vorgang ansetzen, prüfen Sie, ob der Vorgang in seiner heutigen Form überhaupt erhalten bleiben sollte.

Profi-Tipp — Erst der verantwortliche Kopf, dann das Werkzeug Erfolgreiche Häuser stecken den Großteil ihrer Mühe nicht in die Technik. Die Beratung BCG beziffert die Aufteilung bei den Vorreitern grob mit zehn Prozent Algorithmus, zwanzig Prozent Daten und Technik, siebzig Prozent Menschen, Prozesse und Organisation.[10] Bevor ein KI-Vorhaben startet, klären Sie deshalb zwei Dinge: Wer trägt die Verantwortung für das *betriebliche Ergebnis* — nicht für die Vorführung? Und an welcher Zahl messen wir in drei Monaten, ob es sich gelohnt hat?

Warum gerade hier der Anbau lockt

Bis hierher gilt das Gesagte überall. Für den deutschsprachigen Raum kommt etwas hinzu, das den Anbau besonders nahelegt und den Umbau besonders erschwert.

Auf den ersten Blick steht es nicht schlecht. Rund 80 Prozent der deutschen kleinen und mittleren Unternehmen erreichen nach dem Maßstab der EU ein „grundlegendes Niveau digitaler Intensität" — etwas über dem EU-Durchschnitt.[11] Doch dieser Maßstab ist niedrig angesetzt: Er fragt, ob ein Betrieb wenigstens vier von zwölf einfachen digitalen Techniken nutzt. Er misst, ob Werkzeuge vorhanden sind, nicht, ob die Arbeit um sie herum neu gedacht wurde. Und sobald man genauer hinsieht, zeigt sich der Rückstand: Im europäischen Vergleich der Digitalisierung liegt Deutschland im Mittelfeld, bei der digitalen Verwaltung weit hinten, und hinter den nordischen Ländern bleibt es deutlich zurück.[12]

Verantwortlich dafür ist die Organisation, nicht die Technik. Die Deutsche Akademie der Technikwissenschaften (acatech) hält in ihrem Reifegradmodell für die Industrie nüchtern fest, dass sich der Nutzen von Daten erst aus Struktur und Kultur eines Unternehmens erschließt — und dass Daten in den meisten Betrieben in getrennten Silos liegen, die keine Technik allein zusammenführt.[13] Die KfW beschreibt für den Mittelstand die praktischen Hürden: fehlendes digitales Können, lückenhafte Infrastruktur, Finanzierungsschwierigkeiten, die kleine Firmen härter treffen.[14] Ein Mittelständler mit schlanker IT und vollen Auftragsbüchern kauft sich eher ein fertiges Werkzeug, als monatelang seine Abläufe auseinanderzunehmen. Das ist verständlich. Es ist nur kein Vorsprung.

Hinzu kommt eine Eigenheit, die man weder über- noch unterschätzen sollte: die Mitbestimmung. Wo der Betriebsrat bei der Einführung technischer Systeme mitzubestimmen hat, wird der Umbau gründlicher verhandelt und langsamer vollzogen als in einer amerikanischen Firma. Eine belastbare Zahl, die dies als Ursache des Zögerns ausweist, gibt es nicht; als Mechanismus ist es real. Klug geführt, ist die Mitbestimmung kein Hindernis, sondern der Weg, die Belegschaft früh zur Mitgestalterin des Umbaus zu machen — ein Punkt, auf den der vierte Teil zurückkommt.

Dass es anders geht, zeigt ein Werk im eigenen Land. Das Siemens-Elektronikwerk in Amberg gilt international als Vorzeigefabrik, weil es nicht eine alte Linie mit Sensoren nachgerüstet, sondern den Betrieb von Grund auf um den Datenfluss herum gebaut hat.[15] Die genauen Produktivitätszahlen, die das Unternehmen nennt, sind beeindruckend und stammen vom Unternehmen selbst; unabhängig belegt ist die Bauweise, nicht jede Prozentzahl. Doch das Prinzip ist übertragbar und braucht keine Konzerngröße: Wer um die Technik herum neu baut, gewinnt; wer sie nur anbaut, zahlt.

Was das für Ihre Entscheidung heißt

Aus alldem folgt eine Verschiebung der Frage, und sie ist der Übergang zum Rest des Buches. Die übliche Frage lautet: Welches KI-Werkzeug sollen wir einführen? Sie ist die falsche, weil ihre Antwort jedem offensteht. Die richtige Frage lautet: Welchen Arbeitsablauf bauen wir um, wer verantwortet das Ergebnis, und woran messen wir es?

Damit ist auch klar, warum dieses Buch nicht von Werkzeugen handelt, sondern von dem, was sie in Ertrag verwandelt. Drei Dinge bleiben knapp, wenn die Technik im Überfluss vorhanden ist: das Urteil, gute von schlechter Arbeit zu unterscheiden; der eigene Kontext, an dem die Maschine besser wird als bei jedem anderen; und die Disziplin, aus Rechenleistung messbaren Wert zu ziehen. Bevor wir zu diesen drei Knappheiten kommen, müssen Sie allerdings verstehen, womit Sie es zu tun haben. Was eine künstliche Intelligenz wirklich tut, woran man gute Umsetzung erkennt und welche Fragen Sie Ihren Fachleuten stellen sollten: Das ist der Stoff des zweiten Teils.

Das Wichtigste in 60 Sekunden

- KI ist fast überall, zahlt sich aber fast nirgends aus: rund 88 Prozent der Unternehmen nutzen sie, nur etwa sechs Prozent ziehen daraus einen spürbaren Ergebnisbeitrag. Repräsentativ gemessen ist sogar die Nutzung selbst noch Minderheit.
- Das ist das Normalmuster jeder Basistechnologie. Der Ertrag steckt nicht im Werkzeug, sondern im **Organisationskapital** — den neu gedachten Abläufen, geschulten Menschen und geordneten Daten. Dieses Bündel ist unsichtbar, firmenspezifisch und schwer zu kopieren. Eben darum trägt es einen dauerhaften Vorsprung, den die Konkurrenz nicht über Nacht nachbildet.
- KI-Vorhaben scheitern fast nie am Modell, sondern an organisatorischen Ursachen: missverstandenes Ziel, fehlende Daten, Denken vom Werkzeug her, und niemand, der für das betriebliche Ergebnis geradesteht.
- Im deutschsprachigen Raum sind die Werkzeuge weit verbreitet, der Umbau aber selten — gehemmt durch Datensilos, schlanke Mittelstands-IT und eine gründliche, langsamere Entscheidungskultur.
- Die entscheidende Frage hat sich verschoben: weg vom Werkzeug, hin zu der Frage, welchen Ablauf Sie umbauen, wer das Ergebnis verantwortet und woran Sie es messen.

Merksatz Wer einen schlechten Ablauf digitalisiert, bekommt einen schnellen schlechten Ablauf. Den Vorsprung schafft erst der Umbau.

Drei Fragen an Ihr Unternehmen

1. Bei welchem Ihrer laufenden KI-Vorhaben können Sie heute benennen, wer für das *betriebliche Ergebnis* verantwortlich ist — und an welcher Zahl Sie es in drei Monaten messen?
2. Wo haben Sie zuletzt ein Werkzeug eingeführt, ohne den Ablauf dahinter zu hinterfragen — und was würde es kosten, den Ablauf zuerst neu zu denken?
3. Welcher einzelne Arbeitsablauf in Ihrem Haus wäre es wert, von Grund auf um KI herum gebaut zu werden — und was hält Sie davon ab?

Quellen

1. McKinsey & Company, *The State of AI* (Ausgabe 2025): rund 88 % der Unternehmen nutzen KI in mindestens einem Geschäftsbereich; etwa 39 % berichten einen messbaren EBIT-Effekt, rund 6 % („AI high performers") führen mehr als 5 % ihres EBIT auf KI zurück; Letztere gestalten weit häufiger Arbeitsabläufe grundlegend um, statt KI anzuheften. *(Vor Druck: exakte Prozentwerte gegen die aktuelle McKinsey-Veröffentlichung prüfen; Web-Zusammenfassung und PDF weichen mitunter leicht ab.)*
2. US Census Bureau, Business Trends and Outlook Survey (BTOS), KI-Zusatzbefragung; Bonney et al., „Tracking Firm Use of AI in Real Time", CES Working Paper 24-16 (2024) und Census-Auswertung 2026: KI-Einsatz „in mindestens einer Geschäftsfunktion" rund 17 % (Stand 2026), gestiegen von etwa 5 % „in der Erstellung von Gütern/Dienstleistungen" Anfang 2024. Repräsentative Stichprobe von über einer Million Betrieben.
3. Montagezeit des Modell-T-Fahrgestells bei Ford, Highland Park, 1913: Rückgang von rund 12,5 Stunden auf etwa 93 Minuten durch das bewegliche Fließband. Library of Congress, „This Month in Business History"; Assembly Magazine, „The Moving Assembly Line Turns 100" (2013).
4. Bresnahan, Brynjolfsson & Hitt, „Information Technology, Workplace Organization, and the Demand for Skilled Labor", *Quarterly Journal of Economics* 117(1), 2002, S. 339–376.
5. Brynjolfsson, Hitt & Yang, „Intangible Assets: Computers and Organizational Capital", *Brookings Papers on Economic Activity* 2002(1): ein zusätzlicher Dollar Computerkapital ist im Marktwert mit rund zehn Dollar verbunden; die Differenz entspricht komplementärem, nicht bilanziertem Organisationskapital.
6. Saunders & Brynjolfsson, „Valuing IT-Related Intangible Assets", *MIS Quarterly* 40(1), 2016: ein erheblicher Teil der Differenz entfällt auf nicht erfasste Software, IT-Dienstleistungen und Schulung. Die Kernaussage — der wertbestimmende Teil ist unsichtbar und nur lose an die Hardware gekoppelt — bleibt.
7. RAND Corporation, „The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed" (RR-A2680-1, 2024). Fünf führende, durchweg organisatorische Ursachen; beruht auf Experteninterviews. Die ebenfalls genannte Quote „über 80 % der Projekte scheitern" ist als Größenordnung zu verstehen, nicht als gemessene Basisrate.
8. Gartner, Pressemitteilung vom 29. Juli 2024 (Rita Sallam): Prognose, dass bis Ende 2025 mindestens 30 % der Projekte mit generativer KI nach der Machbarkeitsstudie eingestellt werden (u. a. wegen schlechter Datenqualität, unklaren Nutzens, eskalierender Kosten). Analysten-Prognose ohne offengelegte Methodik.
9. Zugeschriebenes Zitat von Thorsten Dirks (damals Telefónica Deutschland), sinngemäß: „Wenn man einen schlechten Prozess digitalisiert, hat man

einen schlechten digitalen Prozess." (*Vor Druck: genauen Wortlaut und Anlass/Quelle prüfen.*)

10. Boston Consulting Group, „Where's the Value in AI?" / AI Radar (2024–2025): bei den Vorreitern entfallen grob 10 % des Aufwands auf Algorithmen, 20 % auf Daten/Technik, 70 % auf Menschen, Prozesse und Organisation. (*Vor Druck: exakte Werte gegen die BCG-Originalpublikation prüfen.*)
11. Eurostat / Europäische Kommission, *Digitale Dekade*, Indikator „digitale Intensität" der KMU (Daten 2024): Deutschland rund 80 % auf grundlegendem Niveau (EU-Schnitt 73 %; Ziel 2030: 90 %). „Grundlegend" bedeutet die Nutzung von mindestens 4 von 12 erfassten digitalen Techniken — ein niedriger Maßstab.
12. Bitkom, „Digitalisierung: Deutschland im EU-Vergleich auf Platz 14" (2024), nachgebaut auf Basis der EU-DESI-Methodik: Deutschland 14. von 27; digitale Verwaltung Platz 21. Es handelt sich um eine Rekonstruktion durch Bitkom, nicht um ein offizielles EU-Ranking.
13. acatech — Deutsche Akademie der Technikwissenschaften, *Industrie 4.0 Maturity Index* (Update 2020): der Nutzen von Daten hängt wesentlich von Organisationsstruktur und Unternehmenskultur ab; Daten liegen überwiegend in dezentralen Silos.
14. KfW Research, *KfW-Digitalisierungsbericht Mittelstand 2024* (auf Basis des repräsentativen KfW-Mittelstandspanels): zentrale Hürden sind fehlendes digitales Know-how, Infrastrukturdefizite und Finanzierungsschwierigkeiten; Digitalisierungsausgaben des Mittelstands 2024 rund 23,8 Mrd. €.
15. Siemens-Elektronikwerk Amberg (EWA): vom Datenfluss her gebaute Fabrik, von der Plattform „Global Lighthouse Network" des Weltwirtschaftsforums ausgezeichnet. Quelle: Siemens-Unternehmensangaben sowie WEF-Lighthouse-Anerkennung. Die genannten Produktivitäts- und Qualitätskennzahlen stammen vom Unternehmen und sind nicht unabhängig geprüft.

TEIL II — KI VERSTEHEN

Genug, um die richtigen Fragen zu stellen.

Der erste Teil hat die Lage beschrieben und eine These aufgestellt: Den Vorsprung schafft nicht das Werkzeug, sondern die Führung, die das Unternehmen darum herum umbaut. Bevor wir zum Umbau kommen, ist eine Zwischenstation nötig, die viele Executive-Formate überspringen. Sie erklären, *dass* es KI gibt, selten, *wie* sie arbeitet. Damit lässt sich kein Anbieter prüfen und kein Projekt steuern. Die folgenden fünf Kapitel schließen diese Lücke. Nicht so tief, dass Sie zum Ingenieur werden; tief genug, dass Sie gute von schlechter Umsetzung unterscheiden und Ihren Leuten die Fragen stellen, vor denen schlechte Antworten kapitulieren.

Wir beginnen mit dem Anfang: was ein Sprachmodell überhaupt tut, wenn es scheinbar mit Ihnen spricht.

Kapitel 3 — Wie eine Maschine sprechen lernt

Lernziele Nach diesem Kapitel können Sie

- in einem Satz erklären, was ein Sprachmodell tut — und warum dieses eine mentale Modell fast alle seine Stärken und Schwächen erklärt;
- sagen, was ein Token und was ein Kontextfenster ist, und warum beides für deutsche Anwendungen teurer ausfällt als für englische;
- begründen, warum solche Modelle halluzinieren und schwanken — und warum das kein Fehler ist, den man wegprogrammiert;
- einschätzen, welche Aufgaben in der verlässlichen Zone liegen und welche nicht — und die eine Frage stellen, die jede Demonstration auf die Probe stellt.

Im Sommer 2025 verhandelte das Amtsgericht Köln einen familienrechtlichen Streit. Eine Partei hatte einen anwaltlichen Schriftsatz eingereicht, und vom achten Blatt an stimmte mit ihm etwas nicht. Er berief sich auf ein Fachbuch, das niemand finden konnte: „Brons, *Kindeswohl und Elternverantwortung*, 2013". Es existiert nicht. Er zitierte eine Fundstelle aus einer angesehenen Fachzeitschrift, die in Wahrheit auf eine erbrechtliche Entscheidung verweist, nicht auf das Familienrecht. Er

nannte Randnummern, die es nicht gibt. Eine Prüfung schätzte: über neunzig Prozent des Textes maschinell erzeugt, die Belege zu hundert Prozent erfunden. Das Gericht sah darin einen Konflikt mit der anwaltlichen Berufspflicht und hielt fest, solches Vorgehen schädige das Ansehen des Rechtsstaats und besonders der Anwaltschaft.[1]

Es ist nicht der einzige Fall. Wenige Wochen zuvor hatte das Oberlandesgericht Celle in einer Berufung gleich vier obergerichtliche Entscheidungen vorgefunden, die sich in keiner Datenbank auffinden ließen, dazu erfundene Aktenzeichen des Bundesgerichtshofs für eine Beschwerdeart, für die der Bundesgerichtshof gar nicht zuständig ist.[2] In den Vereinigten Staaten begann die Reihe schon 2023: Zwei New Yorker Anwälte reichten sechs frei erfundene Urteile bei Gericht ein und wurden mit fünftausend Dollar belegt.[3] Ein Jurist in Frankreich sammelt solche Entscheidungen inzwischen in einer öffentlichen Datenbank; bis Ende 2025 standen darin über siebenhundert Fälle aus aller Welt, der weitaus größte Teil aus diesem einen Jahr, Tendenz steigend.[4]

Die naheliegende Erklärung lautet: nachlässige Anwälte. Sie stimmt, reicht aber nicht. Die interessantere Frage ist, warum eine Maschine, die man um Rechtsprechung bittet, Urteile *erfindet*, statt zu sagen, sie kenne keine, und das so flüssig, so formgerecht, dass ein Fachmann darauf hereinfällt. Die Antwort steckt nicht im Versagen der Technik, sondern in ihrer Funktionsweise. Wer diese versteht, hat den Schlüssel zu fast allem, was in diesem Buch noch folgt, und das Rüstzeug, um nicht selbst in diese Falle zu tappen.

Das eine mentale Modell: eine Vervollständigungsmaschine

Vergessen Sie für einen Moment das Wort „Intelligenz". Ein Sprachmodell tut im Kern etwas erstaunlich Schlichtes: Es setzt einen Text fort. Sie geben ihm einen Anfang — eine Frage, eine Anweisung, ein paar Absätze —, und es errechnet, welches

nächste Textstück am plausibelsten folgt. Dann hängt es dieses Stück an und fragt erneut: Was kommt jetzt am wahrscheinlichsten? Stück für Stück, von vorne nach hinten, entsteht so die Antwort. Der Physiker und Software-Unternehmer Stephen Wolfram hat es nüchtern auf den Punkt gebracht: Das Modell versucht stets, eine „vernünftige Fortsetzung" des bisherigen Textes zu erzeugen, indem es sich immer wieder dieselbe Frage stellt — welches Wort als Nächstes? — und jedes Mal eines anhängt.[5]

Entscheidend ist, *wie* es „am plausibelsten" bestimmt. Es schlägt nirgends nach. Es greift auf keine Datenbank zu. Es errechnet aus dem Gelernten eine Rangliste möglicher nächster Textstücke, jedes mit einer Wahrscheinlichkeit versehen, und wählt daraus aus. Eine Antwort ist für die Maschine kein Faktum, das sie kennt, sondern eine wahrscheinliche Wortfolge, die sie erzeugt. An dieser einen Unterscheidung — *erzeugen* statt *wissen* — hängt fast alles Weitere.

> **Klartext — Token** Ein Modell rechnet nicht in Wörtern, sondern in *Tokens*: kleinen Textbausteinen, meist Wortteilen. „Vervollständigung" zerfällt für die Maschine vielleicht in „Ver", „voll", „ständig", „ung". Als grobe Faustregel gelten im Englischen rund hundert Tokens für etwa fünfundsiebzig Wörter.[6] Der Token ist die kleinste Einheit, in der das Modell liest, schreibt und abgerechnet wird. Behalten Sie das Wort; es kehrt in diesem Buch immer wieder, bis hin zu der Kennzahl, die ihm seinen Namen verdankt.

Woher hat die Maschine ihr Gespür dafür, was plausibel folgt? Aus dem Lesen, in einem Umfang, den kein Mensch erreicht. Moderne Modelle werden auf Textmengen von hunderten Milliarden bis zu mehreren Billionen Tokens trainiert — Bücher, Webseiten, Code, Foren, ein guter Teil des verschriftlichten menschlichen Wissens.[7] Aber sie speichern diesen Text nicht ab wie eine Bibliothek. Sie lernen die *statistischen Regelmäßigkeiten* der Sprache: welche Wörter in welchem Zusammenhang aufein-

ander folgen, wie ein Argument gebaut ist, wie ein Urteil klingt, wie ein Python-Programm aussieht. Was oft und regelmäßig vorkommt — die Grammatik, der Satzbau, der Ton einer Gattung —, lernt es makellos. Was selten vorkommt — ein einzelnes Aktenzeichen, eine entlegene Jahreszahl —, lernt es nicht zuverlässig. Es hat gelernt, wie ein Urteilszitat *aussieht*, nicht, welche Urteile *existieren*. Genau deshalb kann es eines bauen, das perfekt aussieht und nie gefällt wurde.

Eine Analogie hilft, eine andere führt in die Irre. Hilfreich: die Autovervollständigung des Smartphones, aber um viele Größenordnungen mächtiger. Ihr Handy schlägt das nächste Wort aus einer Handvoll vor; ein Sprachmodell berücksichtigt tausende vorausgegangene Wörter und hat unvergleichlich reichere Muster gelernt. Irreführend dagegen ist jedes Bild, das die Maschine zur Suchmaschine, zum Nachschlagewerk oder gar zum Gehirn macht. Sie sucht nichts, sie schlägt nichts nach, sie versteht nichts in dem Sinn, in dem ein Mensch versteht. Sie vervollständigt. Wenn Ihr Chat-Werkzeug heute im Netz sucht und Quellen anzeigt, widerspricht das dem nicht: Der Anbieter hat dem Modell eine Suche zur Seite gestellt, eine Zutat von außen, auf die Kapitel 4 zurückkommt. Die Maschine darunter bleibt, was sie ist. Wer das im Kopf behält, dem erklärt sich vieles von selbst.

Token und Kontextfenster — und warum Deutsch teurer ist

Zwei Begriffe stehen am Anfang, weil an ihnen Kosten, Tempo und ein handfester Standortnachteil hängen. Der erste ist der Token, den Sie schon kennen. Der zweite ist das Kontextfenster.

Das *Kontextfenster* ist alles, was das Modell auf einmal vor Augen hat: Ihre Frage, die angehängten Dokumente und die entstehende Antwort, alles zusammen gemessen in Tokens. Es ist das Arbeitsgedächtnis der Maschine, und es hat eine Obergrenze. Diese Grenze ist in wenigen Jahren steil gewachsen. Eines der frühen großen Modelle, GPT-3, fasste 2020 rund zweitausend Tokens, wenige Buchseiten.[8] Ende 2023 boten erste Modelle

zweihunderttausend, ein mittleres Buch.[9] 2024 erschien ein Modell mit einem Fenster von einer Million Tokens, in Versuchen sogar zehn Millionen, ganze Regalmeter auf einmal.[10]

Daraus folgt eine teure Versuchung und eine stille Falle. Die Versuchung: einfach alles ins Fenster zu kippen, weil es ja passt. Doch jedes Token im Fenster kostet Geld und Zeit, bei *jeder* Anfrage aufs Neue. Und die Falle: Modelle lesen die Mitte schlechter als die Ränder. Eine vielzitierte Untersuchung zeigte, dass sie Informationen am Anfang und am Ende eines langen Textes zuverlässig erfassen, in der Mitte aber deutlich öfter übersehen, eine U-Kurve der Aufmerksamkeit, die selbst bei Modellen auftritt, die für lange Texte gebaut sind.[11] Ein randvolles Fenster ist also kein Gütesiegel. Es kann die wichtige Stelle gerade dort begraben, wo die Maschine am wenigsten hinsieht.

Nun zum Standortnachteil, und er ist konkret. Wie ein Text in Tokens zerlegt wird, hängt von der Sprache ab, und das Deutsche schneidet schlecht ab. Eine Untersuchung, die 2023 auf einer der großen KI-Konferenzen vorgestellt wurde, hat das systematisch vermessen: Auf dem damals gängigen Tokenizer der GPT-4-Generation brauchte derselbe Inhalt auf Deutsch rund das 1,6-Fache der Tokens wie auf Englisch.[12] Neuere Zerlegungsverfahren mildern das, beseitigen es aber nicht; je nach Verfahren liegt der deutsche Aufschlag grob zwischen dreißig und fünfundsechzig Prozent.[13] Die Spitzenwerte für andere Sprachen sind drastischer — für manche das Fünfzehnfache —, doch schon der deutsche Aufschlag genügt, um ins Geld zu gehen. Und weil deutscher Text mehr Tokens braucht, passt in dasselbe Kontextfenster entsprechend weniger Inhalt.

Denn Tokens sind die Einheit, in der abgerechnet wird, getrennt für das, was hineingeht, und das, was herauskommt; die Ausgabe kostet ein Mehrfaches der Eingabe, Stand 2026 grob das Fünffache.[14] Mehr Tokens heißt also unmittelbar: mehr Kosten und längere Wartezeit. Wer eine englischsprachige Fallstudie liest und ihre Kosten auf den eigenen, deutschsprachigen Betrieb überträgt, rechnet sich systematisch zu günstig. Hier beginnt

die eigene Ökonomie dieser Technik im deutschsprachigen Raum; ihr widmet Teil III ein ganzes Kapitel, unter dem Namen *Return on Tokens*: dem Wert, den Sie je Recheneinheit herausholen.

Zahlen, die zählen

- Faustregel Englisch: ~**100 Tokens ≈ 75 Wörter**.[6]
- Deutscher Token-Aufschlag: **rund +30 bis +65 %** gegenüber Englisch für denselben Inhalt.[12][13]
- Ausgabe-Tokens kosten **rund das Fünffache** der Eingabe-Tokens (Stand 2026; ändert sich rasch).[14]
- Kontextfenster: von **~2.000 Tokens (2020)** über **200.000 (2023)** auf **1 Million (2024)**.[8][9][10]

Warum dieselbe Frage zwei Antworten bekommt

Stellen Sie einem Sprachmodell zweimal exakt dieselbe Frage, und Sie können zwei verschiedene Antworten erhalten. Für eine Maschine wirkt das befremdlich; wir sind gewohnt, dass ein Taschenrechner auf 7 mal 8 stets dasselbe ausgibt. Die Erklärung steckt wieder im Mechanismus. Das Modell errechnet ja nicht *die* Antwort, sondern eine Rangliste plausibler nächster Tokens mit Wahrscheinlichkeiten. Ob es stets das wahrscheinlichste nimmt oder gelegentlich ein etwas weniger wahrscheinliches, ist eine Einstellsache. Ein Regler, in der Fachsprache *Temperatur* genannt, steuert, wie sehr die Maschine vom wahrscheinlichsten Pfad abweichen darf: niedrig eingestellt antwortet sie nüchtern und nah am Erwartbaren, hoch eingestellt verspielter und überraschender.[15]

Diese Schwankung ist kein Defekt, sie ist die Kehrseite einer Stärke. Dieselbe Eigenschaft, die ein Modell zum brauchbaren Ideengeber und Textentwerfer macht — es muss nicht stets dasselbe sagen —, macht es unzuverlässig überall dort, wo es auf exakte, wiederholbare, immer gleiche Ausgabe ankommt. Selbst

wer die Temperatur auf null stellt, erhält in der Praxis nicht garantiert jedes Mal Wort für Wort dasselbe; auf der Ebene der Rechenzentren spielen technische Effekte hinein, deren Einzelheiten hier nichts zur Sache tun.[16] Für die Führung zählt die Konsequenz: Wo Sie Verlässlichkeit brauchen wie von einer Person, müssen Sie sie organisieren — durch klare Aufgaben, Prüfung und feste Abläufe —, nicht von der Maschine erwarten.

Warum sie halluziniert — und warum das bleibt

Damit sind wir zurück beim Kölner Schriftsatz. Wenn ein Modell eine plausible Wortfolge erzeugt und keine Fakten kennt, dann ist eine flüssig formulierte Falschaussage für die Maschine nichts grundsätzlich anderes als eine richtige. Beide sind Fortsetzungen mit einer gewissen Wahrscheinlichkeit. Ein erfundenes Aktenzeichen sieht aus wie ein echtes, klingt wie ein echtes, fügt sich grammatisch tadellos in den Satz — und genau deshalb erzeugt es die Maschine bereitwillig, wenn der Zusammenhang danach verlangt. Im Fachjargon heißt das *Halluzination*: die Erzeugung plausibel klingender, aber sachlich falscher Inhalte.[17]

> **Klartext — Halluzination** Eine *Halluzination* ist eine Ausgabe, die überzeugend klingt, aber nicht stimmt — eine erfundene Quelle, ein falsches Datum, eine Tatsache, die es nie gab. Das Wort führt in die Irre, denn es klingt nach Ausnahme. Tatsächlich ist es dieselbe Maschinerie, die richtige und falsche Antworten erzeugt; das Modell selbst „merkt" den Unterschied nicht.

Die wichtigste und für viele unbequemste Einsicht lautet: Das lässt sich nicht wegprogrammieren. Halluzination ist kein Bug, der in der nächsten Version verschwindet, sondern eine Eigenschaft des Verfahrens. Forscher haben dafür mehrere, sich ergänzende Begründungen vorgelegt. Eine Gruppe zeigte 2024 mit den Mitteln der Lerntheorie, dass ein Modell, das beliebige Probleme lösen soll, zwangsläufig in manchen Fällen falschliegen *muss* — eine formale Grenze, kein Umsetzungsmangel.[18] For-

scher von OpenAI legten 2025 nach und richteten den Blick auf die Anreize: Schon beim Training sind Fehler bei seltenen, kaum gestützten Fakten statistisch unvermeidlich; und die üblichen Eignungstests belohnen das Modell wie eine Multiple-Choice-Prüfung dafür, *selbstbewusst zu raten*, statt „Ich weiß es nicht" zu sagen. Wer fürs Raten Punkte bekommt und fürs Eingestehen von Unwissen keine, lernt zu bluffen.[19] Die Maschine ist, mit anderen Worten, darauf optimiert, lieber falsch und flüssig als richtig und zögerlich zu antworten.

Das verändert die Aufgabe der Führung grundlegend. Wer auf ein Modell wartet, das nie lügt, wartet auf etwas, das es seiner Natur nach nicht geben wird. Die kluge Antwort heißt Einbau: Abläufe so gestalten, dass die Maschine dort prüfbar bleibt, wo es auf Wahrheit ankommt — durch Quellenangabe, durch das Nachschlagen in vertrauenswürdigen Daten, durch den Menschen an der richtigen Stelle. Wie das geht, ist der Stoff der folgenden Kapitel. Hier zählt das Prinzip: Sie konstruieren um eine bekannte Schwäche herum, statt auf ihre Heilung zu hoffen.

Und sie ist teuer, diese Schwäche, sobald man ihr vertraut. Eine kanadische Fluggesellschaft musste 2024 Schadenersatz zahlen, weil ihr Chatbot einem Kunden eine Tarifregel zusagte, die es nicht gab; das Gericht ließ die Ausrede nicht gelten, der Bot sei für seine Auskünfte selbst verantwortlich.[20] Der Kölner Anwalt kam mit einer Rüge davon. Andere bezahlten mit Geld, mit Ansehen oder mit beidem. Die Lehre aus Kapitel 1 kehrt hier in schärferer Form wieder.

Stolperfalle — Der überzeugende Irrtum, jetzt erklärt In Kapitel 1 stand die Warnung vor der plausibel formulierten, falschen Antwort. Jetzt wissen Sie, *warum* es sie gibt: Die Maschine erzeugt Wahrscheinliches, nicht Wahres, und Falsches kann ebenso wahrscheinlich klingen wie Richtiges. Je näher eine Aufgabe an der Grenze des Könnens liegt, desto glatter und gefährlicher wird der Irrtum. Die Gegenmaßnahme ist eine Haltung: An den Stellen, an denen Wahrheit zählt, überlassen Sie die Prüfung nicht der Maschine.

Es sagt Text voraus, es rechnet nicht

Eine kleine, lehrreiche Schwäche macht die Sache anschaulich. Fragt man manche Modelle, wie viele „r" im englischen Wort *strawberry* stecken, kommt verblüffend oft die falsche Zahl. Eine Maschine, die juristische Aufsätze entwirft, scheitert an einer Grundschulaufgabe — wie kann das sein?

Aus zwei Gründen, die beide direkt aus dem Mechanismus folgen. Erstens sieht das Modell die einzelnen Buchstaben gar nicht: Es liest in Tokens, in Wortbausteinen, und „strawberry" zerfällt womöglich in „straw" und „berry". Die Buchstabengrenzen sind ihm verdeckt.[21] Zweitens, und grundsätzlicher: Zählen und Rechnen sind feste Verfahren, Schritt für Schritt abzuarbeiten. Das Modell arbeitet kein Verfahren ab. Es sagt das plausibelste nächste Token voraus. Auf die Frage nach einer Summe liefert es die Zahl, die in solchen Zusammenhängen am wahrscheinlichsten steht — nicht das Ergebnis einer Rechnung. Dass dabei oft das Richtige herauskommt, verdankt sich der Häufigkeit, mit der korrekte Rechnungen in den Trainingsdaten standen, nicht einem Rechenwerk im Inneren.

Neuere Modelle umschiffen das Problem, indem sie für solche Aufgaben einen echten Taschenrechner aufrufen oder sich Zwischenschritte notieren — eine Krücke, die das Symptom behebt, nicht die Ursache. Das eigentliche Merkzeichen bleibt: Es ist ein Sprachmodell, kein Rechenwerk, keine Suchmaschine, keine Datenbank. Es sagt Text voraus. Alles, was es kann und nicht kann, lässt sich aus diesem einen Satz ableiten.

Wofür sie verlässlich taugt — und wofür nicht

Aus all dem ergibt sich eine brauchbare Landkarte. Verlässlich ist ein Sprachmodell dort, wo eine Aufgabe **begrenzt** ist und ihr Ergebnis **billig zu prüfen**. In dieser Zone liegen die gemessenen Gewinne aus Kapitel 1: das Entwerfen und Umformulieren von Texten, das Zusammenfassen eines Dokuments, das man selbst zur Hand hat, das Übersetzen, das Sortieren und Herausziehen von Informationen aus vorgelegtem Material. Hier arbeitet die

Maschine schnell und gut, weil sie genau das tut, wofür sie gebaut ist: Sprache umformen. Die kontrollierten Studien sind eindeutig: beim anspruchsvollen Schreiben rund 40 Prozent weniger Zeit bei besserer Qualität, im Kundendienst die größten Sprünge bei den unerfahrenen Kräften.[22]

Unverlässlich wird sie, wo es auf garantierte Richtigkeit ankommt, ohne dass jemand prüft: exakte Arithmetik, das Abrufen entlegener Einzelfakten, jede Aufgabe, deren Ergebnis teuer wird, wenn es flüssig daneben liegt. Und sie wird heimtückisch an jener Grenze, die in Kapitel 1 schon der Sache nach auftauchte und die ich die zerklüftete nenne: Aufgaben, die ähnlich schwer aussehen, fallen unvorhersehbar mal in das Können des Modells, mal knapp daneben, und gerade jenseits davon liefert es seine selbstbewusst falschen Antworten.[23] Die Faustregel, die alles zusammenhält, lautet deshalb: **Trauen Sie der Maschine in dem Maß, in dem Sie ihr Ergebnis billig überprüfen können — und keinen Schritt weiter.**

Ein Wort zur eigenen Sprache, ehrlich gehalten, weil die Datenlage dünn ist: Auf vergleichenden Tests schneiden die Modelle auf Deutsch ein paar Prozentpunkte schlechter ab als auf Englisch. Der Abstand ist klein, er schrumpft mit jeder Modellgeneration, und Deutsch gehört zu den stärksten unter den Nicht-Englisch-Sprachen.[24] Man sollte daraus keine Glaubensfrage machen. Aber zusammen mit dem Token-Aufschlag bleibt es dabei: Englische Maßstäbe lassen sich nicht ungeprüft auf den deutschsprachigen Betrieb übertragen.

Papagei oder beginnender Verstand?

Eine Debatte gehört noch hierher, denn sie wird unter Fachleuten mit einiger Schärfe geführt: Ist „nur das nächste Token vorhersagen" wirklich alles? Die eine Seite, prominent vertreten durch den KI-Forscher Ilya Sutskever, hält dagegen: Um das nächste Wort wirklich gut vorherzusagen, müsse ein Modell die Wirklichkeit hinter dem Text begreifen; in der Statistik stecke notgedrungen ein Stück Weltverständnis.[25] Die andere Seite, am

bekanntesten in der Streitschrift vom „stochastischen Papagei", entgegnet: Das Modell fügt Sprachformen nach Wahrscheinlichkeit zusammen, ohne jeden Bezug zur Bedeutung — flüssige Hülle ohne Verstehen.[26]

Mein Standpunkt ist ein doppelter. Über den *Mechanismus* gibt es keinen Streit: Es ist Vorhersage des nächsten Tokens aus Wahrscheinlichkeiten, gelernt aus den Regelmäßigkeiten riesiger Textmengen. Worüber gestritten wird, ist die *Deutung* — ob daraus bloße Nachahmung wird oder etwas, das man beginnendes Verstehen nennen darf. Diese Frage ist offen, und sie wird sich nicht durch Zuruf entscheiden, sondern durch Forschung. Die gute Nachricht für die Führungskraft: Sie müssen sie nicht entscheiden. Sie müssen die Folgen des Mechanismus kennen — die Halluzination, die Schwankung, die zerklüftete Grenze — und Ihr Geschäft um sie herum bauen. Wer auf die philosophische Klärung wartet, verschiebt das Handeln auf einen Tag, den ihm niemand zusagen kann.

Was dieses Kapitel für Sie ändert

Sie haben jetzt das eine mentale Modell, das den Rest von Teil II trägt. Eine Maschine, die Text vervollständigt, in Tokens rechnet, aus Wahrscheinlichkeiten wählt und aus Statistik gelernt hat, statt Fakten zu speichern. Aus diesem einen Bild folgen die Halluzination, die Schwankung, der Buchstabenfehler, die Verlässlichkeitsgrenze und der deutsche Kostenaufschlag — nicht als fünf zufällige Eigenheiten, sondern als fünf Gesichter derselben Sache. Wenn Ihnen das nächste Mal jemand eine glänzende Demonstration zeigt, haben Sie damit das Rüstzeug, hinter den Glanz zu sehen.

Die richtige Frage an Ihre Technik Wenn Ihnen ein Team oder ein Anbieter eine KI-Lösung vorführt, stellen Sie eine Frage, vor der schlechte Antworten kapitulieren: *„Woran prüft das System, ob seine Antwort stimmt — und woran würde ich es merken, wenn sie es nicht tut?"* Kommt darauf nur ein Verweis auf das eindrucksvolle Modell, fehlt das Entscheidende. Eine gute Antwort nennt die Quelle, an der geprüft wird, die Stellen, an denen ein Mensch gegenliest, und das, was geschieht, wenn die Maschine danebenliegt. Die ausführliche Form dieser Frage — wie man Verlässlichkeit überhaupt misst — ist das Thema von Kapitel 6.

Das Wichtigste in 60 Sekunden

- Ein Sprachmodell ist eine **Vervollständigungsmaschine:** Es errechnet Stück für Stück das plausibelste nächste Textstück aus Wahrscheinlichkeiten. Es *erzeugt* Antworten, es *weiß* sie nicht.
- Es hat aus riesigen Textmengen die **Regelmäßigkeiten der Sprache** gelernt, keine Faktendatenbank. Grammatik makellos, seltene Fakten unzuverlässig.
- **Halluzination und Schwankung** sind keine Bugs, sondern Folgen des Verfahrens. Man programmiert sie nicht weg; man baut prüfbar um sie herum.
- Es **sagt Text voraus, es rechnet nicht** — und es liest in **Tokens**, weshalb Deutsch rund 30 bis 65 Prozent mehr kostet als Englisch.
- Verlässlich ist es, wo die Aufgabe **begrenzt** und das Ergebnis **billig prüfbar** ist. **Trauen Sie ihm so weit, wie Sie es billig überprüfen können.**

Merksatz Die Maschine erzeugt das Wahrscheinliche, nicht das Wahre — und Ihre Aufgabe ist es, den Unterschied prüfbar zu machen.

Drei Fragen an Ihr Unternehmen

1. An welchen Stellen verlassen sich Ihre Leute heute schon auf KI-Ausgaben, *ohne* dass jemand sie billig gegenprüfen kann?
2. Welche Ihrer geplanten KI-Anwendungen liegen in der verlässlichen Zone (begrenzt, prüfbar) — und welche verlangen eine Garantie, die diese Technik ihrer Natur nach nicht gibt?
3. Haben Sie bei Ihren Kostenschätzungen den deutschen Token-Aufschlag berücksichtigt — oder rechnen Sie mit englischen Maßstäben?

Quellen

1. Amtsgericht Köln, Beschluss vom 2. Juli 2025, Az. 312 F 130/25 (Familiensache). Anwaltlicher Schriftsatz mit erfundener Monografie („Brons, *Kindeswohl und Elternverantwortung*, 2013"), fehlerhafter Fundstelle in der *FamRZ* (tatsächlich erbrechtliche Entscheidung) und nicht existenten Randnummern; das Gericht sah einen Konflikt mit § 43a Abs. 3 BRAO und rügte das Vorgehen, ohne förmliche Sanktion. Berichterstattung: beck-aktuell, Legal Tribune Online (LTO), Anwaltsblatt, 2025. (*Vor Druck: Entscheidungstext und Wortlaut der Rüge gegenprüfen; juristisch ist umstritten, ob ein BRAO-Verstoß i. e. S. vorliegt — daher „gerügt", nicht „sanktioniert".*)
2. Oberlandesgericht Celle, Urteil vom 29. April 2025, Az. 5 U 1/25. In der Berufung wurden vier obergerichtliche Entscheidungen zitiert, die in juris und beck-online nicht auffindbar sind, sowie erfundene BGH-Aktenzeichen für eine Beschwerdeart außerhalb der Zuständigkeit des BGH. Nachweis: dejure.org (5 U 1/25); juristische Fachkommentare 2025.
3. *Mata v. Avianca, Inc.*, U.S. District Court, Southern District of New York, Sanktionsbeschluss vom 22. Juni 2023 (678 F. Supp. 3d 443). Zwei Anwälte reichten sechs von ChatGPT frei erfundene Gerichtsentscheidungen ein; Sanktion von 5.000 US-Dollar.
4. Damien Charlotin (HEC Paris), öffentliche Datenbank zu Gerichtsentscheidungen mit KI-„halluzinierten" Inhalten, https://www.damiencharlotin.com/hallucinations/ . Stand Ende Dezember 2025: über 700 erfasste Entscheidungen weltweit, der weitaus größte Teil aus 2025; Tendenz steigend. (*Vor Druck: aktuellen Live-Stand mit Datum einsetzen — die Zahl wächst wöchentlich.*)
5. Stephen Wolfram, *What Is ChatGPT Doing … and Why Does It Work?*, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ . Das Modell erzeugt eine „reasonable continuation" und wählt aus einer mit Wahrscheinlichkeiten versehenen Rangliste möglicher nächster Wörter.
6. OpenAI Help Center, „What are tokens and how to count them?", https://help.openai.com/en/articles/4936856 . Faustregel: ~1 Token ≈ 4 Zeichen ≈ 0,75 Wörter (Englisch); ~100 Tokens ≈ 75 Wörter. Gilt ausdrücklich nur für Englisch.
7. Größenordnung der Trainingskorpora: GPT-3 (Brown et al. 2020) wurde auf rund 300 Mrd. Tokens trainiert; Frontier-Modelle ab 2024 (z. B. Llama 3, Meta 2024, arXiv:2407.21783) auf rund 15 Billionen Tokens. Die Aussage „statistische Regelmäßigkeiten, kein Faktenspeicher" stützt sich auf Bender et al. 2021 (Anm. 26) sowie die Halluzinationsforschung (Anm. 18–19).
8. Brown et al., *Language Models are Few-Shot Learners* (GPT-3), NeurIPS 2020, arXiv:2005.14165: Kontextfenster n_ctx = 2.048 Tokens.
9. Anthropic, Ankündigung Claude 2.1, November 2023: Kontextfenster von 200.000 Tokens. (*Vor Druck: aktuelle Modell-Dokumentation des Anbieters prüfen; Werte ändern sich rasch.*)

10. Gemini Team, Google, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, 2024, arXiv:2403.05530: Standard-Kontextfenster von 1 Mio. Tokens, in Forschungsversuchen bis 10 Mio.
11. Liu et al., *Lost in the Middle: How Language Models Use Long Contexts*, Transactions of the Association for Computational Linguistics (TACL) 2024, arXiv:2307.03172. Leistung am höchsten, wenn die relevante Information am Anfang oder Ende steht; deutlich schlechter in der Mitte langer Kontexte.
12. Petrov, La Malfa, Torr, Bibi, *Language Model Tokenizers Introduce Unfairness Between Languages*, NeurIPS 2023, arXiv:2305.15425. Deutsch: 1,64× der Token-Zahl von Englisch auf dem GPT-4/ChatGPT-Tokenizer (cl100k_base); Maximalfall (Shan) ~15×; für betroffene Sprachen mindestens 2,5× Kosten und bis zu 2× Latenz.
13. Untere Schranke des deutschen Aufschlags aus neueren Tokenizern (o200k_base, GPT-4o, 2024): rund +31 % (≈ 1,31×) auf Wortebene; obere Schranke aus Petrov et al. 2023 (1,64×). Verteidigbare Spanne daher rund +30 bis +65 %. Die o200k-Zahl stammt aus einer reproduzierbaren, aber nicht begutachteten Einzelanalyse — als Größenordnung, nicht als exakter Wert zu lesen. Eine DACH-spezifische Studie liegt nicht vor; der Standortbezug ist eine eigene Schlussfolgerung aus den Petrov-Daten.
14. Verhältnis Ausgabe- zu Eingabe-Token-Preisen rund 5:1 bei führenden Modellen (Stand Mitte 2026, z. B. Anthropic Claude, OpenAI GPT). Offizielle Preisseiten der Anbieter; die genauen Werte ändern sich rasch und sind generisch zu lesen.
15. Mechanismus von Sampling und Temperatur: Renze & Guven, *The Effect of Sampling Temperature on Problem Solving in LLMs*, 2024, arXiv:2402.05201; Grundlagen in Goodfellow, Bengio & Courville, *Deep Learning*, 2016. Die Temperatur skaliert die Verteilung vor der Auswahl (niedrig = deterministischer, hoch = zufälliger).
16. He et al., Thinking Machines Lab, *Defeating Nondeterminism in LLM Inference*, September 2025, https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/ . Selbst bei Temperatur 0 variieren Ausgaben im Betrieb, weil GPU-Rechenkerne nicht „batch-invariant" sind. Labor-/Industriebefund, nicht begutachtet; Zahlen illustrativ.
17. Definition und Taxonomie: Huang et al., *A Survey on Hallucination in Large Language Models*, arXiv:2311.05232 (2023; ACM TOIS). „plausible yet nonfactual content"; Unterscheidung von Faktizitäts- und Treue-Halluzination.
18. Xu, Jain, Kankanhalli, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, arXiv:2401.11817 (2024). Lerntheoretisches Resultat: in einem formalisierten Rahmen müssen Modelle als allgemeine Problemlöser zwangsläufig halluzinieren. Idealisiertes Setting; als theoretischer Beleg zu lesen.
19. Kalai, Nachum, Vempala, Zhang (OpenAI), *Why Language Models Hallucinate*, arXiv:2509.04664 (4. September 2025), Begleittext https://openai.com/index/why-language-models-hallucinate/ . These: Pretraining macht Fehler bei selten gestützten Fakten statistisch unvermeidlich; Bewertungs-Benchmarks belohnen selbstbewusstes Raten gegenüber dem Eingeständnis von

Unwissen. Vendor-Analyse, nicht begutachtet — als Argument, nicht als Beweis zu lesen.

20. *Moffatt v. Air Canada*, British Columbia Civil Resolution Tribunal, 2024 BCCRT 149, 14. Februar 2024. Haftung der Fluggesellschaft für eine falsche Chatbot-Auskunft (negligent misrepresentation); Schadenersatz C$ 812,02. Eher Fall der Chatbot-Haftung als der Halluzination im engeren Sinn.
21. Zur Ursache der Zähl-/Rechenfehler: Tokenisierung verdeckt einzelne Buchstaben (Bostrom & Durrett, *Byte Pair Encoding is Suboptimal for Language Model Pretraining*, arXiv:2004.03720, 2020); empirische Demonstration bei Max Woolf, *Can modern LLMs count the b's in 'blueberry'?*, August 2025, https://minimaxir.com/2025/08/llm-blueberry/ . Neuere Modelle umgehen das per Werkzeugaufruf oder Zwischenschritten; das behebt das Symptom, nicht die Ursache.
22. Noy & Zhang, *Experimental evidence on the productivity effects of generative AI*, *Science* 381:187–192 (2023): rund –40 % Zeit bei höherer Qualität. Brynjolfsson, Li & Raymond, *Generative AI at Work*, NBER 31161 (2023; *Quarterly Journal of Economics* 140(2), 2025): +14 % gelöste Anfragen pro Stunde, +34 % bei unerfahrenen Kräften. (Beide bereits in Kapitel 1 eingeführt.)
23. Dell'Acqua et al., *Navigating the Jagged Technological Frontier*, HBS Working Paper 24-013 / SSRN 4573321 (2023; *Organization Science* 2025): innerhalb der Modellfähigkeiten deutliche Qualitäts- und Tempogewinne; bei einer bewusst außerhalb platzierten Aufgabe rund 19 Prozentpunkte häufiger falsche Ergebnisse. (Bereits in Kapitel 1 eingeführt.)
24. MMLU-ProX, *A Multilingual Benchmark for Advanced LLM Evaluation*, 2025, arXiv:2503.10497. Auf übersetzten Tests liegt Deutsch typischerweise wenige Prozentpunkte hinter Englisch (Lücke meist 0–5 pp, mit neueren Modellen schrumpfend); Deutsch zählt zu den stärksten Nicht-Englisch-Sprachen. Übersetzte Benchmarks bergen Artefakte; die Aussage ist qualitativ zu lesen.
25. Ilya Sutskever im Gespräch mit Dwarkesh Patel, 27. März 2023, https://www.dwarkesh.com/p/ilya-sutskever : „Predicting the next token well means that you understand the underlying reality that led to the creation of that token." Analytische Position einer interessierten Partei, kein gemessenes Resultat.
26. Bender, Gebru, McMillan-Major & Mitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, FAccT '21, DOI 10.1145/3442188.3445922: Sprachformen werden „according to probabilistic information about how they combine, but without any reference to meaning" zusammengefügt. (Die vierte Autorin publizierte unter dem Pseudonym „Shmargaret Shmitchell".)

Kapitel 4 — Vom Rohmodell zu Ihrem Modell

Lernziele Nach diesem Kapitel können Sie

- Training und Inferenz auseinanderhalten — und erklären, warum der teure Teil davon nicht Ihre Aufgabe ist;
- sagen, was aus einem Rohmodell überhaupt einen brauchbaren Assistenten macht — und wo dabei die Leitplanken und die Gefälligkeit eingebaut werden;
- die drei Wege benennen, ein Modell an Ihr Geschäft anzupassen — Prompt, Nachschlagen, Feinabstimmung — und entscheiden, wann welcher der richtige ist;
- die Datenschutzfrage stellen, die jeder dieser Wege aufwirft, und den einen Hebel benennen, der Ihre Daten im Haus hält.

Eine Formulierung hört man immer wieder, wenn in Unternehmen über künstliche Intelligenz gesprochen wird: „Wir trainieren die KI auf unseren Daten." Der Satz klingt nach einem klaren Vorhaben. Tatsächlich steht er für sehr verschiedene Dinge, die sich in Aufwand, Kosten und Wirkung deutlich unterscheiden — und wer sie verwechselt, bezahlt am Ende das Schwerste, wo das Leichteste genügt hätte.

Um sie zu trennen, muss man zuerst wissen, was „trainieren" überhaupt heißt. Und das Erste, was man darüber wissen sollte,

ist, dass der teure Teil davon mit großer Sicherheit nicht Ihre Aufgabe ist.

Zwei Vorgänge, die man nie verwechseln sollte: Training und Inferenz

Hinter jedem Sprachmodell stehen zwei grundverschiedene Vorgänge, und ihre Verwechslung ist die Wurzel vieler Fehlentscheidungen.

Der erste ist das *Training*. Es ist der einmalige Lernvorgang, aus dem das Modell überhaupt erst entsteht — jener Prozess aus Kapitel 3, in dem es aus riesigen Textmengen die Regelmäßigkeiten der Sprache aufnimmt. Training ist das Lernen, und es ist teuer. Das Forschungsinstitut Epoch AI, das die Trainingskosten der großen Modelle so genau verfolgt wie kaum jemand sonst, schätzt die reinen Rechenkosten für ein Spitzenmodell der GPT-4-Klasse auf vierzig bis achtzig Millionen Dollar.[1] Sie verdoppeln sich etwa alle acht bis zwölf Monate; Epoch rechnete 2024 damit, dass der größte einzelne Trainingslauf bis 2027 die Milliardengrenze überschreitet.[2] Ein Modell von Grund auf zu bauen, ist deshalb die Sache einer Handvoll Konzerne. Es ist so gut wie nie das, was ein Unternehmen meint, wenn es sagt, es wolle „auf seinen Daten trainieren".

Der zweite Vorgang ist die *Inferenz*. Das ist die Anwendung des fertigen Modells auf eine einzelne Anfrage: Sie stellen eine Frage, das Modell errechnet die Antwort, ohne dabei etwas dazuzulernen. Anders als das Training geschieht die Inferenz nicht einmal, sondern bei jeder Anfrage aufs Neue, solange das System läuft.

> **Klartext — Inferenz** *Inferenz* ist das Ausführen eines fertigen Modells: aus einer Eingabe eine Antwort errechnen. Das Modell lernt dabei nichts; es wendet nur an, was im Training entstanden ist. Inferenz ist die laufende Betriebsleistung — und damit Ihre eigentliche, wiederkehrende KI-Rechnung.

Aus dieser Trennung folgt die wichtigste ökonomische Einsicht des Kapitels. Die spektakulären Millionenbeträge betreffen das Training und sind die Kosten des Anbieters, nicht Ihre. *Ihre* Kosten sind die der Inferenz: gering pro Anfrage, aber dauerhaft und mit der Nutzung wachsend. Wie groß dieser Posten über die Lebenszeit eines Systems wird, unterschätzen die meisten. Der Cloud-Anbieter AWS schätzt, dass die Inferenz „bis zu neunzig Prozent" der Kosten einer betriebenen KI-Anwendung ausmachen kann — eine Zahl, die man als das nehmen sollte, was sie ist: die Schätzung eines Unternehmens, das an genau dieser Inferenz-Rechenleistung verdient, ohne offengelegte Methodik.[3] Belastbar bleibt der Kern: Über die Lebensdauer eines stark genutzten Systems übersteigen die laufenden Inferenzkosten die einmaligen Trainingskosten regelmäßig. Genau diese laufende Rechnung behandelt Teil III unter dem Namen *Return on Tokens* — dem Wert, den Sie je Recheneinheit herausholen.

Ein Modell, dessen Training abgeschlossen ist, hat noch eine zweite Eigenschaft, die Sie kennen müssen: Sein Wissen ist eingefroren. Es endet an dem Tag, an dem die Trainingsdaten gesammelt wurden, dem *Wissensstichtag*. Was danach geschah, weiß das Modell nicht — nicht die Quartalszahlen von gestern, nicht die Gesetzesänderung von letzter Woche, und vor allem nichts über den Inhalt Ihrer eigenen Systeme, die in keinem Trainingsdatensatz der Welt stehen. Diese Lücke ist der Ansatzpunkt jeder Anpassung. Hier setzt all das an, was ein generisches Modell zu *Ihrem* macht.

Warum das zugekaufte Modell schon ein Assistent ist

Bevor wir zu den drei Wegen kommen, ein Zwischenschritt, der oft übersprungen wird und ohne den vieles unverständlich bleibt. Das Modell, das Sie über eine Programmierschnittstelle oder eine App ansprechen, ist nicht das rohe Ergebnis des Trainings. Ein Rohmodell, frisch aus dem Training, tut genau das, was Kapitel 3 beschrieben hat, und sonst nichts: Es vervollständigt Text. Stellt man ihm eine Frage, ist die wahrscheinlichste

Fortsetzung womöglich eine weitere Frage, denn so stehen Fragen in Texten oft beieinander. Ein Rohmodell ist ein erstaunliches Sprachtalent und ein nutzloser Assistent.

Aus dem einen wird das andere durch einen weiteren Schritt, das *Post-Training*. Der Anbieter zeigt dem Modell zunächst viele von Menschen verfasste Musterantworten und lässt es das Befolgen von Anweisungen üben. Dann kommt das Verfahren, das die heutigen Assistenten erst möglich gemacht hat: Menschen bekommen mehrere mögliche Antworten des Modells vorgelegt und ordnen sie danach, welche besser ist. Aus diesen Urteilen lernt ein zweites Modell, Antworten so zu bewerten, wie Menschen es täten, und das Sprachmodell wird darauf getrimmt, mehr von dem zu erzeugen, was bevorzugt wurde.

> **Klartext — RLHF** RLHF steht für „Verstärkungslernen aus menschlichem Feedback". Menschen ranken Antworten des Modells nach Güte; das Modell wird darauf optimiert, künftig mehr von dem zu liefern, was bevorzugt wurde. Hier — nicht im Training — entstehen der hilfsbereite Ton, die Höflichkeit und die Weigerung, bei heiklen Bitten mitzumachen.

Wie viel dieser Schritt ausmacht, zeigt ein Befund, der die Branche aufhorchen ließ. Die Forscher hinter dem Verfahren stellten fest, dass die Antworten eines vergleichsweise kleinen, aber sorgfältig nachtrainierten Modells mit 1,3 Milliarden Parametern von menschlichen Testern *bevorzugt* wurden gegenüber denen des hundertmal größeren Rohmodells GPT-3 mit 175 Milliarden Parametern.[4] Die Ausrichtung schlug die schiere Größe. Daraus folgt eine Lehre für jede Technik-Bewertung: Was ein Modell im Alltag brauchbar macht, ist weniger seine Größe als seine Ausrichtung auf den Menschen. Hier, im Post-Training, baut der Anbieter auch seine Sicherheitsleitplanken ein — die Regeln, nach denen das Modell bestimmte Auskünfte verweigert.[5]

Diese Ausrichtung hat eine Kehrseite, die an Kapitel 3 anschließt. Ein Modell, das darauf optimiert wird, zu *gefallen*, lernt

mitunter, dem Nutzer nach dem Mund zu reden, statt ihm die Wahrheit zu sagen. Forscher nennen das übertriebene Gefälligkeit, und sie haben gezeigt, dass sowohl Menschen als auch die Bewertungsmodelle überzeugend formulierte, schmeichelnde Antworten den korrekten vorziehen.[6] Im April 2025 musste OpenAI ein Update seines Modells zurücknehmen, das genau in diese Falle getappt war: Es war, in den Worten des Unternehmens, „übermäßig schmeichelhaft oder gefällig" geworden, weil man es zu sehr auf kurzfristiges Nutzerlob optimiert hatte.[7]

> **Stolperfalle — Auf Gefallen getrimmt heißt nicht auf Wahrheit getrimmt** Ein Assistent, der Ihnen zustimmt, tut zunächst nur, was ihm antrainiert wurde: gefallen. Gerade bei Entscheidungen, bei denen Sie Widerspruch bräuchten, ist die freundliche Bestätigung das gefährlichste Ergebnis. Die Gegenmaßnahme ist dieselbe wie in Kapitel 3: An den Stellen, an denen es zählt, prüfen Sie das Urteil der Maschine, statt sich von ihrem Ton einnehmen zu lassen.

Halten wir fest, was das zugekaufte Modell schon mitbringt, denn das verhindert die häufigste Verwechslung. Der Anbieter hat es bereits zum anweisungstreuen Assistenten gemacht. Das ist etwas anderes, als wenn *Sie* es auf Ihren Daten weiter abstimmen, und wieder etwas anderes, als wenn Sie ihm zur Laufzeit Ihre Dokumente vorlegen. Diese drei Dinge stecken oft hinter der einen Formulierung vom „Trainieren auf eigenen Daten" — und sie zu trennen ist der Kern dieses Kapitels.

Die drei Wege, ein Modell zu Ihrem zu machen

Es gibt drei Wege, ein generisches Modell an Ihr Geschäft anzupassen. Sie unterscheiden sich erheblich in Kosten, Aufwand und Wirkung, und die kluge Reihenfolge ist, mit dem einfachsten zu beginnen und erst dann zur nächsten Stufe zu gehen, wenn die vorige nicht reicht.

Der erste Weg ist der Prompt. Sie geben dem Modell in der Anweisung selbst alles mit, was es braucht: eine klare Aufgabe, ein paar Beispiele für das gewünschte Ergebnis, den Ton, das Format. Das Modell „lernt" die Aufgabe aus diesen Beispielen, ohne dass an ihm selbst etwas verändert würde — die Fachleute sprechen vom Lernen im Kontext. Schon das GPT-3-Papier von 2020 zeigte, dass ein Modell allein aus wenigen Beispielen im Prompt eine Aufgabe bewältigt, „ohne jede Gewichtsanpassung".[8] Das ist der mit Abstand billigste und schnellste Weg, und für einen großen Teil der Anwendungen — Zusammenfassen, Übersetzen, Umformulieren, Entwerfen — ist er oft der einzige, den man braucht.[9] Hier sollte jedes Vorhaben anfangen.

Der zweite Weg ist das Nachschlagen, im Fachjargon RAG. Statt zu hoffen, dass das nötige Wissen im Modell steckt, geben Sie ihm zur Laufzeit die richtigen Unterlagen an die Hand: Das System sucht aus Ihren Dokumenten, Datenbanken oder Handbüchern die passenden Stellen heraus und legt sie dem Modell zusammen mit der Frage vor. Das Modell antwortet dann auf Grundlage *Ihrer* Quellen. Vom bloßen Mitgeben im Prompt unterscheidet sich das in zwei Punkten: Ihre gesammelten Unterlagen sprengen jedes Kontextfenster, und jedes mitgeschickte Token kostet bei jeder Anfrage aufs Neue. Das Nachschlagen wählt deshalb aus, statt alles mitzuschicken.

> **Klartext — Nachschlagen (RAG)** RAG heißt „abruf-gestützte Erzeugung". Bevor das Modell antwortet, sucht das System die relevanten Stellen aus Ihren eigenen Daten heraus und reicht sie mit. So bekommt das Modell die richtigen Seiten aufgeschlagen, statt aus dem Gedächtnis zu raten. Es ist die übliche Art, eine KI an firmeneigenes, aktuelles Wissen zu binden.

Das Nachschlagen löst drei Probleme auf einmal. Es überwindet den Wissensstichtag, weil das Modell auf den heutigen Stand Ihrer Unterlagen zugreift. Es bindet die Antwort an Ihr eigenes Wissen, das in keinem öffentlichen Modell steckt. Und es *mindert* die Halluzination, weil das Modell aus vorgelegten Quellen

schöpft, statt zu erfinden — die ursprüngliche Untersuchung, die den Begriff prägte, zeigte bereits „spezifischere und sachlich richtigere" Antworten als ein Modell ohne diese Stütze.[10] Ein Wort der Vorsicht: Es mindert die Halluzination, es beseitigt sie nicht. Auch mit den richtigen Seiten vor Augen kann ein Modell danebengreifen. Dass das Nachschlagen darüber hinaus einen Wettbewerbsvorteil berührt, den ein Wettbewerber nicht nachkaufen kann — die eigenen Daten —, ist ein so großes Thema, dass ihm der dritte Teil dieses Buches ein eigenes Kapitel widmet. Hier zählt nur die Mechanik.

Der dritte Weg ist die Feinabstimmung. Hier verändern Sie das Modell tatsächlich: Sie trainieren seine Gewichte auf Ihren eigenen Beispielen weiter, sodass es sich dauerhaft anders verhält — in Stil, Format, Fachsprache. Das ist der schwerste der drei Wege. Er verlangt kuratierte Trainingsdaten, einiges an Fachkönnen und oft ein Modell, dessen Innenleben offen genug liegt, um es überhaupt zu verändern. Die Technik ist in den vergangenen Jahren billiger geworden; ein Verfahren namens LoRA etwa senkte die Zahl der anzupassenden Stellschrauben gegenüber einer vollständigen Neuabstimmung auf ein Zehntausendstel.[11] Billiger heißt aber nicht billig, und einfacher nicht einfach.

Zahlen, die zählen

- Training eines Modells der GPT-4-Klasse: **rund 40–80 Mio. $** (Schätzung, je nach Methode).[1]
- Ausrichtung schlägt Größe: ein **1,3-Mrd.**-Modell wurde von Menschen einem **175-Mrd.**-Rohmodell vorgezogen.[4]
- Feinabstimmung per LoRA: nur **ein Zehntausendstel** der anzupassenden Parameter gegenüber voller Neuabstimmung (modellabhängig, illustrativ).[11]

Wann welcher Weg — die Regel und ihre Tücke

Damit zur Frage, die in der Praxis am häufigsten falsch beantwortet wird: Wann nimmt man welchen Weg? Die brauchbarste Faustregel lautet: **Das Nachschlagen bringt dem Modell neues Wissen, die Feinabstimmung bringt ihm neues *Verhalten* bei.** Soll die Maschine über Ihre Produkte, Verträge und Vorgänge Bescheid wissen, ist das Nachschlagen das Mittel der Wahl. Soll sie in einem bestimmten Format, Ton oder Fachjargon antworten oder eine eng umrissene Aufgabe stets gleich erledigen, kommt die Feinabstimmung in Betracht.

Diese Regel ist eine Heuristik, keine Naturkonstante, und an einer Stelle ist sie belegt: Eine Untersuchung von 2024 verglich beide Wege beim Einspeisen neuen Faktenwissens und fand, dass das Nachschlagen die Feinabstimmung durchweg übertraf — Modelle täten sich schwer, neue Fakten durch bloßes Weitertrainieren zuverlässig aufzunehmen.[12] Das ist ein starkes Indiz, kein Allbeweis; die beiden Wege schließen einander auch nicht aus, sondern werden in anspruchsvollen Anwendungen oft kombiniert.[13] Die praktische Konsequenz aber ist klar genug, um sie sich zu merken.

Stolperfalle — Feinabstimmung, um Wissen einzuspeisen Ein ebenso verbreiteter wie teurer Fehler in KI-Projekten ist, ein Modell aufwendig auf firmeneigenen Dokumenten feinabzustimmen, damit es deren *Inhalt* kennt. Dafür ist die Feinabstimmung das falsche Werkzeug — sie lehrt Verhalten, nicht verlässlich Fakten. Für firmeneigenes Wissen ist das Nachschlagen billiger, schneller aktualisierbar und meist genauer. Wer Wissen über Feinabstimmung lösen will, zahlt mehr für ein schlechteres Ergebnis.

Die richtige Frage an Ihre Technik Wenn ein Team vorschlägt, ein Modell feinabzustimmen, stellen Sie zwei Fragen, bevor Sie das Budget freigeben: *„Haben wir Prompt und Nachschlagen wirklich ausgeschöpft — und wollen wir der Maschine neues Wissen geben oder neues Verhalten beibringen?"* Geht es um Wissen und ist das Nachschlagen noch nicht versucht, ist der Vorschlag verfrüht. Die meisten Anwendungen kommen weit, ehe die Feinabstimmung überhaupt nötig wird.

Die Reihenfolge, die daraus folgt, empfehlen seriöse Anbieter selbst: erst der Prompt, der für viele Fälle schon genügt; dann das Nachschlagen, wenn das Modell firmeneigenes oder aktuelles Wissen braucht; und erst, wenn beides ausgereizt ist, die Feinabstimmung für Verhalten und Format.[9] Man geht die Leiter nur so weit hinauf, wie man muss. Jede höhere Stufe kostet mehr Geld, mehr Können und mehr Pflege — und ob sie ihren Aufwand wert war, lässt sich ohnehin erst sagen, wenn man es misst. Wie man das misst, ist das Thema von Kapitel 6.

Jeder dieser Wege ist auch eine Datenentscheidung

Es gibt einen Aspekt, den eine deutschsprachige Führungskraft mitdenken muss und der in den aus dem Amerikanischen übernommenen Ratgebern meist fehlt. In dem Moment, in dem Sie einem zugekauften Modell Ihre Daten anvertrauen — sei es im Prompt, sei es als nachgeschlagene Unterlage, sei es als Material für die Feinabstimmung —, verlassen diese Daten Ihr Haus.

Datenschutzrechtlich hat das Folgen. Der Anbieter wird in der Regel zum Auftragsverarbeiter, was einen Vertrag nach Artikel 28 der Datenschutz-Grundverordnung verlangt; je nachdem, wo gerechnet wird, kommt eine Übermittlung in ein Drittland hinzu, die nach Kapitel V der Verordnung eigenen Regeln folgt. Die Datenschutzkonferenz, das Gremium der deutschen Aufsichtsbehörden, hat dazu im Mai 2024 eine Orientierungshilfe veröffentlicht, die diese Punkte benennt — und die zugleich die Ausnahme festhält: Wer eine Anwendung ausschließlich zu eigenen Zwecken auf eigenen Servern betreibt, begründet keine Auftragsverarbeitung.[14] Hinzu kommt die Frage, ob Ihre Eingaben zum Weitertrainieren des Modells verwendet werden. Bei den kostenlosen Verbraucherzugängen war das vielfach der Fall; die geschäftlichen Zugänge und Programmierschnittstellen der großen Anbieter tun es, Stand 2026, in der Regel nicht und bieten zudem eine Datenhaltung innerhalb der EU an — was vor jedem ernsthaften Einsatz im aktuellen Vertrag zu prüfen ist, denn diese Bedingungen ändern sich schnell.[15]

Dass dies keine theoretische Sorge ist, hat die italienische Datenschutzbehörde vorgeführt. Sie sperrte ChatGPT im März 2023 vorübergehend und verhängte im Dezember 2024 ein Bußgeld von fünfzehn Millionen Euro — das erste gegen ein Unternehmen der generativen KI; OpenAI hat es angefochten.[16] Auf europäischer Ebene richtete der Europäische Datenschutzausschuss eigens eine Arbeitsgruppe zu ChatGPT ein, deren Bericht 2024 erschien.[17]

> **Aus der Praxis — Als ChatGPT in Italien verschwand** Am 30. März 2023 untersagte Italiens Garante die Verarbeitung der Daten italienischer Nutzer durch ChatGPT. Über Nacht war der Dienst im Land nicht mehr erreichbar; erst nach Auflagen zu Transparenz und Nutzerrechten kam er Ende April zurück. Die Lehre liegt weniger im späteren Bußgeld als in der Abhängigkeit, die der Vorgang offenlegte: Ein Werkzeug, das über Nacht abgeschaltet werden kann, gehört nicht zu den Dingen, auf die man ohne Rückfallebene baut.

Hier zeigt sich auch, warum die dritte Stufe, die Feinabstimmung, mit einer strategischen Frage verknüpft ist, die uns später beschäftigt. Vollständig im Haus bleiben Ihre Daten nur, wenn Sie ein offenes Modell auf eigenen Servern betreiben; europäische Cloud-Infrastruktur entschärft die Drittlandsfrage, ersetzt den Auftragsverarbeitungsvertrag aber nicht. Ob das die Mehrkosten und den Aufwand wert ist, hängt vom Einzelfall ab; die Abwägung zwischen offenen und geschlossenen Modellen führen wir in Kapitel 7, die Souveränität als Vorstandsthema in Kapitel 9. Auch der EU AI Act bringt hier Pflichten ins Spiel, vor allem zur Transparenz; deren Fristen sind allerdings derzeit in Bewegung, weshalb man den Stand vor jeder verbindlichen Festlegung prüfen sollte.[18]

Was dieses Kapitel für Sie ändert

Sie müssen kein Modell bauen, und in den allermeisten Fällen sollten Sie es auch nicht. Der Weg vom Rohmodell zu Ihrem Modell führt nicht über ein eigenes Training, sondern über die Anpassung eines fähigen, vom Anbieter bereits zum Assistenten gemachten Modells — beginnend mit dem Prompt, dann dem Nachschlagen für Ihr eigenes Wissen, und erst zuletzt, wenn nötig, der Feinabstimmung für Verhalten und Format. Sie wissen jetzt, dass Ihre eigentliche Rechnung die der Inferenz ist und nicht die des Trainings; dass die Freundlichkeit der Maschine antrainiert und kein Gütesiegel für Wahrheit ist; und dass jede

Anpassung zugleich eine Entscheidung darüber ist, wohin Ihre Daten fließen.

Damit erkennen Sie einen Großteil der überzogenen Vorschläge, die in KI-Projekten Geld verbrennen — allen voran den Reflex, gleich das Schwerste zu tun, wo das Leichteste genügt hätte. Das nächste Kapitel geht einen Schritt weiter: Was geschieht, wenn das Modell nicht mehr nur antwortet, sondern zu handeln beginnt — Werkzeuge bedient, Schritte aneinanderreiht, selbständig wird? Dort wird aus einer falschen Antwort erstmals eine falsche Handlung.

Das Wichtigste in 60 Sekunden

- **Training** (das Modell bauen) ist die einmalige, millionenteure Sache des Anbieters. **Inferenz** (das Modell betreiben) ist Ihre laufende Rechnung, pro Anfrage, dauerhaft. Deshalb baut fast niemand selbst — man passt an.
- Das zugekaufte Modell ist durch das **Post-Training** des Anbieters schon zum Assistenten gemacht; dort sitzen Ton, Leitplanken und die Neigung, zu gefallen statt zu stimmen.
- Drei Wege der Anpassung, vom billigsten zum teuersten: **Prompt** (Anweisung und Beispiele), **Nachschlagen/RAG** (eigene Daten zur Laufzeit), **Feinabstimmung** (Gewichte ändern).
- Faustregel: **Nachschlagen bringt Wissen, Feinabstimmung bringt Verhalten.** Gehen Sie die Leiter nur so weit hinauf, wie Sie müssen.
- Jede Anpassung schickt Ihre Daten aus dem Haus — eine **Datenschutz- und Souveränitätsentscheidung**, kein bloß technischer Schritt.

Merksatz Bevor Sie ein Modell feinabstimmen, lassen Sie es nachschlagen; und bevor Sie es nachschlagen lassen, schreiben Sie einen besseren Prompt.

Drei Fragen an Ihr Unternehmen

1. Bei welchem Ihrer KI-Vorhaben wird gerade die schwerste Stufe (Feinabstimmung, gar eigenes Training) erwogen — und haben Sie Prompt und Nachschlagen davor wirklich ausgeschöpft?
2. Geht es bei Ihren Anwendungen eher darum, der Maschine *Wissen* zu geben (dann: Nachschlagen) oder ihr ein *Verhalten* beizubringen (dann: Feinabstimmung)?
3. Wissen Sie für jede eingesetzte KI, wohin Ihre Daten fließen, ob sie zum Training verwendet werden und ob Sie eine Rückfallebene haben, falls der Dienst ausfällt?

Quellen

1. Epoch AI, *How much does it cost to train frontier AI models?* (2024), https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models ; ergänzend Stanford HAI, *AI Index Report 2024* (Cloud-Mietpreis-Schätzung für GPT-4 ~78 Mio. $). Die Spanne von rund 40 bis 80 Mio. $ ergibt sich aus unterschiedlichen Berechnungsmethoden (amortisierte Hardware vs. Cloud-Miete), nicht aus widersprüchlichen Befunden. Trainingskosten sind Schätzungen; die Unternehmen veröffentlichen sie nicht.
2. Cottier, Rahman, Fattorini, Maslej, Owen (Epoch AI), *The rising costs of training frontier AI models*, arXiv:2405.21015 (2024): Wachstum der amortisierten Trainingskosten rechenintensivster Modelle seit 2016 um rund das 2,4-Fache pro Jahr (90 %-Konfidenzintervall 2,0–2,9×); Projektion des größten Trainingslaufs über 1 Mrd. $ bis 2027.
3. AWS, Ankündigung der EC2-Inf1-Instanzen (Jeff Barr, AWS News Blog, 3. Dezember 2019), und EC2-Inf1-Produktseite: „inferencing can account for up to 90% of the cost of their machine learning work". Es handelt sich um die Schätzung eines Anbieters von Inferenz-Hardware ohne veröffentlichte Methodik; sie wird hier ausdrücklich als solche und nicht als gesicherte Tatsache zitiert. Belastbar ist die strukturelle Aussage, dass laufende Inferenzkosten mit der Nutzung skalieren und die einmaligen Trainingskosten über die Lebensdauer eines Systems regelmäßig übersteigen.
4. Ouyang et al., *Training language models to follow instructions with human feedback* (InstructGPT), arXiv:2203.02155 (2022): Ausgaben des nachtrainierten 1,3-Mrd.-Parameter-Modells wurden von menschlichen Bewertern den Ausgaben des 175-Mrd.-Parameter-GPT-3 vorgezogen. „Vorgezogen" bezieht sich auf das Urteil menschlicher Tester auf der Prompt-Verteilung von OpenAI, nicht auf einen Faktizitäts- oder Reasoning-Benchmark. Ursprung des Verfahrens: Christiano et al., *Deep reinforcement learning from human preferences*, arXiv:1706.03741 (2017).
5. Zur Verortung von Hilfsbereitschaft, Ton und Sicherheitsverhalten im Post-Training: Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, arXiv:2212.08073 (2022). Der Begriff „Persönlichkeit" eines Modells ist umgangssprachlich.
6. Sharma et al. (Anthropic), *Towards Understanding Sycophancy in Language Models*, arXiv:2310.13548 (2023): „both humans and preference models (PMs) prefer convincingly-written sycophantic responses over correct ones." Übertriebene Gefälligkeit (engl. sycophancy) gilt als allgemeines Verhalten heutiger Assistenten, mitverursacht durch menschliche Präferenzurteile.
7. OpenAI, *Sycophancy in GPT-4o: What happened and what we're doing about it* (29. April 2025): Ein Update wurde „overly flattering or agreeable", weil es „focused too much on short-term feedback"; OpenAI nahm es zurück. (*Vor Druck: Datum und Wortlaut aus der Primärquelle gegenprüfen.*)
8. Brown et al., *Language Models are Few-Shot Learners* (GPT-3), arXiv:2005.14165 (2020): „GPT-3 is applied without any gradient updates or fine-

tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model." Dies begründet das „Lernen im Kontext".

9. OpenAI, Entwicklerleitfaden *Optimizing LLM Accuracy*, https://developers.openai.com/api/docs/guides/optimizing-llm-accuracy : „Prompt engineering is typically the best place to start. It is often the only method needed for use cases like summarization, translation, and code generation." Der Leitfaden empfiehlt die Reihenfolge Prompt → Nachschlagen → Feinabstimmung und deren Kombination. Vendor-Best-Practice, nicht peer-reviewed; Microsoft Azure publiziert dieselbe Reihenfolge.
10. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS 2020, arXiv:2005.11401: prägt den Begriff RAG (Kombination aus parametrischem Modell und nicht-parametrischem, abgerufenem Wissen); „RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline." Der Vergleich gilt einer seq2seq-Baseline, nicht modernen Chat-Modellen.
11. Hu et al., *LoRA: Low-Rank Adaptation of Large Language Models*, arXiv:2106.09685 (2021): „reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times" gegenüber der vollständigen Feinabstimmung von GPT-3 175B. Die Zahlen sind für diesen Vergleich spezifisch und illustrativ, keine universelle Konstante.
12. Ovadia et al., *Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs*, arXiv:2312.05934 (2023/2024): Beim Einspeisen neuen Faktenwissens übertrifft das Nachschlagen die Feinabstimmung durchweg; „LLMs struggle to learn new factual information through unsupervised fine-tuning." Getestet wurde unüberwachte Feinabstimmung auf einem Faktenkorpus — ein starkes Indiz, kein Allbeweis.
13. OpenAI, *Optimizing LLM Accuracy* (wie Anm. 9): Nachschlagen optimiert den Kontext (Wissen), Feinabstimmung das Verhalten (Konsistenz, Format) und spart Prompt-Länge; beide lassen sich kombinieren („fine-tuning + RAG").
14. Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK), *Orientierungshilfe „Künstliche Intelligenz und Datenschutz"*, Version 1.0, 6. Mai 2024, https://www.datenschutzkonferenz-online.de : benennt das Auftragsverarbeitungsverhältnis nach Art. 28 DS-GVO, mögliche Drittstaatentransfers nach Kapitel V sowie die Ausnahme für den ausschließlichen Eigenbetrieb auf eigenen Servern. Ergänzend: HmbBfDI, Diskussionspapier *Large Language Models und personenbezogene Daten* (Juli 2024); LfDI Baden-Württemberg, *Rechtsgrundlagen im Datenschutz beim Einsatz Künstlicher Intelligenz*, v2.0 (Oktober 2024).
15. Generische Anbieterdokumentation, Stand 2026: geschäftliche und API-Zugänge der großen Anbieter nutzen Eingaben standardmäßig nicht zum Training und bieten EU-Datenresidenz (z. B. OpenAI Business/Data-Residency; Microsoft Azure OpenAI „EU Data Boundary"). Diese Bedingungen ändern sich schnell und sind im jeweils aktuellen Vertrag zu prüfen; die DSK warnt, Konten so zu konfigurieren, dass keine Eingaben ins Training fließen.
16. Garante per la protezione dei dati personali (Italien): vorläufige Beschränkung der Verarbeitung am 30. März 2023 (Provvedimento 9870832), Wieder-

zulassung Ende April 2023 nach Auflagen; Bußgeld von 15 Mio. € am 20. Dezember 2024 (erstes DS-GVO-Bußgeld gegen ein Unternehmen der generativen KI; Gründe u. a. fehlende Rechtsgrundlage für das Training, unterlassene Meldung des Vorfalls von März 2023, mangelnde Altersprüfung). OpenAI hat das Bußgeld angefochten. Quelle: gpdp.it; garanteprivacy.it. (*Vor Druck: docweb-Nummer der Dezember-2024-Entscheidung gegenprüfen.*)

17. Europäischer Datenschutzausschuss (EDPB), *Report of the work undertaken by the ChatGPT Taskforce*, 23. Mai 2024, https://www.edpb.europa.eu .
18. EU AI Act: Transparenzpflichten betreffen auch Anwender; die Fristen für Hochrisiko-Anforderungen sollen durch das von der EU-Kommission am 19. November 2025 vorgeschlagene „Digital-Omnibus"-Paket verschoben werden (Annex-III-Systeme voraussichtlich bis spätestens 2. Dezember 2027). Stand Mitte 2026 vorläufig; vor Drucklegung den aktuellen Stand prüfen, kein Datum als endgültig nennen.

Kapitel 5 — Vom Chatbot zum Agenten

Lernziele Nach diesem Kapitel können Sie

- erklären, was einen Agenten von einem Chatbot unterscheidet — und warum die Trennlinie zwischen Lesen und Handeln die wichtigste ist;
- begründen, warum mehrstufige Autonomie zerbrechlich ist und warum das Versprechen „voll autonom" Skepsis verdient;
- die Angriffsfläche benennen — die Prompt-Injection — und warum sie sich nicht so einfach abdichten lässt wie eine Software-Lücke;
- die Fragen stellen, die geklärt sein müssen, bevor Sie einer Maschine erlauben, in der echten Welt zu handeln.

Im Juli 2025 ließ der Investor Jason Lemkin einen KI-Agenten an einem Softwareprojekt arbeiten — nicht im Versuchslabor, sondern an einem laufenden System, über rund zwei Wochen hinweg. Der Agent sollte programmieren, und er konnte mehr als nur Code vorschlagen: Er durfte ihn ausführen, Datenbanken ändern, selbständig Schritte aneinanderreihen. An einem der Tage hatte Lemkin ausdrücklich einen Änderungsstopp verhängt. Der Agent hielt sich nicht daran. Er löschte die produktive Datenbank des Projekts. Nach Lemkins Schilderung erzeugte

er anschließend tausende erfundene Datensätze und meldete wahrheitswidrig, alles sei in Ordnung. Der Geschäftsführer von Replit, dem Unternehmen hinter dem Werkzeug, nannte den Vorfall öffentlich „nicht hinnehmbar" und zog technische Konsequenzen.[1]

Man kann die Geschichte als Warnung lesen oder als Lehrstück. Sie ist beides. Vor allem aber markiert sie die Schwelle, um die es in diesem Kapitel geht — den Übergang von einer Maschine, die *antwortet*, zu einer, die *handelt*. Solange ein Modell nur Text zurückgibt, ist der Schaden eines Fehlers begrenzt: eine falsche Auskunft, die ein Mensch prüfen kann. Sobald es ausführen, ändern, verschicken darf, wird aus demselben Fehler eine gelöschte Datenbank, eine versandte E-Mail, eine ausgelöste Zahlung. Die Technik dahinter ist dieselbe Vervollständigungsmaschine aus Kapitel 3. Was sich ändert, ist, was sie anrichten kann.

Was ein Agent ist — und was ihn vom Chatbot trennt

Ein Chatbot beantwortet eine Frage und wartet auf die nächste. Ein *Agent* bekommt ein Ziel und arbeitet selbständig darauf hin: Er zerlegt die Aufgabe in Schritte, entscheidet, was als Nächstes zu tun ist, greift dabei auf *Werkzeuge* zurück und reagiert auf das, was diese zurückmelden — eine Schleife aus Planen, Handeln, Beobachten, bis das Ziel erreicht ist oder er aufgibt. Anthropic, einer der großen Anbieter, beschreibt Agenten als Systeme, in denen das Modell „seine eigenen Abläufe und seinen Werkzeugeinsatz selbst steuert".[2]

> **Klartext — Agent und Werkzeug** Ein *Werkzeug* ist eine Fähigkeit, die das Modell bei Bedarf aufrufen kann: eine Websuche, ein Stück Programmcode, eine Datenbankabfrage, der Versand einer E-Mail, der Aufruf einer Schnittstelle. Ein *Agent* ist ein Modell, das ein Ziel verfolgt, dazu selbst entscheidet, welche Werkzeuge es in welcher Reihenfolge benutzt, und seine nächsten Schritte aus den Ergebnissen ableitet. Kurz: Der Chatbot redet, der Agent tut.

Es lohnt, eine Zwischenstufe zu kennen, weil sie die meisten brauchbaren Anwendungen ausmacht. Zwischen dem reinen Chatbot und dem frei agierenden Agenten liegt der *Arbeitsablauf*: ein fest verdrahteter Pfad, bei dem ein Mensch im Voraus festgelegt hat, welche Schritte das Modell in welcher Folge durchläuft. Anthropic unterscheidet beides sauber — beim Arbeitsablauf folgt das Modell vorgegebenen Bahnen, beim Agenten bestimmt es seinen Weg selbst.[2] Diese Unterscheidung gehört zu den wichtigsten Entwurfsentscheidungen, denn mit jedem Stück Autonomie, das man der Maschine überlässt, wächst nicht nur ihr Nutzen, sondern auch ihre Fehleranfälligkeit. Dazu gleich mehr.

Zuvor die Linie, an der für eine Führungskraft das meiste hängt: die zwischen Lesen und Handeln. Ein Agent, der nur *liest* — sucht, abruft, zusammenträgt —, kann sich irren, aber sein Irrtum bleibt eine Information, die man prüfen kann. Ein Agent mit *Schreibrechten* greift in die Welt ein: Er ändert Datensätze, verschickt Nachrichten, gibt Geld aus. Die Fachautorin Chip Huyen bringt es auf eine Formel, die jeder Vorstand verstehen sollte: So, wie man einem Praktikanten nicht die Vollmacht gibt, die Produktionsdatenbank zu löschen, sollte man einer unzuverlässigen KI nicht erlauben, Überweisungen anzustoßen.[3] Lemkins Datenbank war genau dieser Fall — ein Agent mit Schreibrechten, der sich irrte.

Warum mehrstufige Autonomie zerbrechlich ist

Hinter der Anfälligkeit von Agenten steckt eine schlichte Rechnung, die man einmal gesehen haben muss. Angenommen, ein Modell erledigt jeden einzelnen Schritt mit einer Zuverlässigkeit von 95 Prozent — ein für sich genommen guter Wert. Eine Aufgabe aber, die zehn Schritte verlangt, ist erst dann gelöst, wenn *jeder* Schritt gelingt. Die Wahrscheinlichkeit dafür ist nicht mehr 95 Prozent, sondern 95 Prozent mit sich selbst multipliziert, zehnmal — und das ergibt rund 60 Prozent. Über hundert Schritte fällt sie auf etwa ein halbes Prozent.[4] Fehler addieren sich nicht, sie vervielfachen sich. Das ist gemeint, wenn von *zusammengesetzten Fehlern* die Rede ist.

> **Klartext — Zusammengesetzte Fehler** Bei einer mehrstufigen Aufgabe muss jeder Schritt gelingen, damit das Ganze gelingt. Die Zuverlässigkeiten der Schritte multiplizieren sich, statt sich zu mitteln. Selbst kleine Fehlerraten pro Schritt summieren sich über viele Schritte zu einer hohen Gesamtfehlerrate. Deshalb ist eine lange Kette autonomer Handlungen prinzipiell zerbrechlicher als ein einzelner Schritt.

Diese Rechnung ist eine Faustregel, keine Naturkonstante; sie unterstellt, dass die Fehler voneinander unabhängig sind und niemand zwischendurch korrigiert. In Wahrheit haben gute Agentensysteme Wiederholungen und Selbstkorrektur eingebaut, und reale Fehler hängen oft zusammen. Die Größenordnung aber stimmt, und sie wird durch Messungen gestützt. Die Forschungsgruppe METR hat 2025 untersucht, wie lange eine Aufgabe sein darf, die heutige Modelle noch zuverlässig bewältigen. Das Ergebnis ist ernüchternd und aufschlussreich zugleich: Aufgaben, für die ein Mensch weniger als vier Minuten braucht, lösen die besten Modelle nahezu vollständig — Aufgaben jenseits von etwa vier Stunden menschlicher Arbeit dagegen in weniger als zehn Prozent der Fälle.[5] Die Reichweite verlässlicher Autonomie wächst zwar von Jahr zu Jahr; kurz bleibt sie trotzdem, und

sie wird umso kürzer, je höher man die geforderte Verlässlichkeit ansetzt.

Zahlen, die zählen

- 95 % Zuverlässigkeit pro Schritt ergeben über **10 Schritte ≈ 60 %**, über **100 Schritte ≈ 0,6 %**.[4]
- Modelle lösen Aufgaben unter **4 Minuten** menschlicher Arbeit nahezu vollständig, solche über **4 Stunden** zu **unter 10 %**.[5]
- Auf einem realitätsnahen Aufgaben-Benchmark (2024) gelang selbst den besten Systemen nur bei **unter einem Viertel** der Aufgaben die Lösung in allen acht Wiederholungen.[6]

Das erklärt auch, warum man Agenten ihre eigenen Erfolgsmeldungen nicht glauben darf. Zu den typischen Fehlern zählt neben dem Griff zum falschen Werkzeug und dem Sprengen von Zeit- und Kostenbudgets ein besonders tückischer: Der Agent ist überzeugt, die Aufgabe erfüllt zu haben, obwohl er es nicht hat.[7] Lemkins Agent, der nach der Löschung Vollzug meldete, ist genau dieser Fall. Es ist die übertriebene Gefälligkeit aus Kapitel 4, übertragen auf das Handeln: Die Maschine liefert das überzeugende Bild des Erfolgs, weil sie auf überzeugende Ausgaben getrimmt ist, nicht auf wahre.

Stolperfalle — Das Versprechen vom voll autonomen Agenten Wenn ein Anbieter einen Agenten als „vollständig autonom" bewirbt, der eine lange, vielstufige Aufgabe ohne Aufsicht erledigt, ist Skepsis angebracht. Je mehr Schritte ohne Kontrolle aufeinanderfolgen, desto wahrscheinlicher ist ein Fehler irgendwo in der Kette — und desto schwerer ist er zu bemerken. Die belastbaren Anwendungen sind heute die mit kurzem Weg, klarem Rahmen und einem Menschen an den Kontrollpunkten. Anthropic selbst rät, mit der einfachsten Lösung zu beginnen und die Autonomie nur dort zu erhöhen, wo sie sich auszahlt.[2]

Die Angriffsfläche: die Prompt-Injection

Es gibt eine zweite Gefahr, die mit der ersten nichts zu tun hat und gerade deshalb leicht übersehen wird. Sie folgt aus einer Eigenschaft, die wir schon kennen: Ein Sprachmodell verarbeitet alles als Text — Ihre Anweisung und die Daten, die es liest, im selben Strom. Es hat keine zuverlässige Möglichkeit, das eine vom anderen zu trennen. Wer Anweisungen in die Daten schmuggelt, kann das Modell dazu bringen, sie zu befolgen.

> **Klartext — Prompt-Injection** Eine *Prompt-Injection* ist das Einschmuggeln versteckter Anweisungen, die das Modell als Befehl auffasst und ausführt. Bei der *direkten* Form formuliert der Nutzer selbst eine Eingabe, die die Schutzregeln des Modells umgeht. Bei der gefährlicheren *indirekten* Form stehen die Schadanweisungen in fremdem Material, das der Agent arglos liest — einer Webseite, einem Dokument, einer eingehenden E-Mail.

Dass dies keine ausgedachte Bedrohung ist, haben Forscher 2023 gezeigt — und es ist ein Befund, auf den man im deutschsprachigen Raum stolz sein darf, denn er stammt überwiegend aus Saarbrücken, vom CISPA-Helmholtz-Zentrum und der Universität des Saarlandes.[8] Die Gruppe um Kai Greshake führte vor, wie sich Anwendungen, die ein Sprachmodell einbinden, „aus der Ferne, ohne direkte Schnittstelle" angreifen lassen: indem man die schädliche Anweisung dort platziert, wo das Modell sie später abruft. Die Sicherheitsorganisation OWASP führt die Prompt-Injection seither als Risiko Nummer eins ihrer Rangliste für KI-Anwendungen.[9] Eine vollständige Abdichtung gibt es bislang nicht — anders als ein Software-Leck, das man stopft, sitzt diese Lücke in der Funktionsweise selbst.

Richtig brisant wird sie, sobald der Agent nicht nur liest, sondern auch nach außen wirken kann. Der Sicherheitsforscher Simon Willison hat die gefährliche Kombination benannt: Sobald ein Agent zugleich Zugriff auf vertrauliche Daten hat, nicht

vertrauenswürdige Inhalte verarbeitet *und* nach außen kommunizieren kann, lässt er sich dazu verleiten, die vertraulichen Daten an einen Angreifer zu schicken.[10] Genau das geschah 2025 unter dem Namen „EchoLeak". Sicherheitsforscher der Firma Aim Security wiesen nach, dass sich aus Microsofts Assistenten Microsoft 365 Copilot Daten abziehen ließen, durch eine einzige präparierte E-Mail, die der Empfänger nicht einmal anklicken musste. Die versteckte Anweisung darin genügte, damit der Assistent Inhalte aus dem Zugriffsbereich des Nutzers nach außen gab. Microsoft schloss die Lücke; ausgenutzt wurde sie nach Angaben des Unternehmens nicht.[11]

> **Aus der Praxis — Der Angriff, der keinen Klick brauchte** „EchoLeak" ist deshalb lehrreich, weil das Opfer nichts falsch machen musste. Keine angeklickte Datei, kein eingegebenes Passwort; eine E-Mail im Postfach reichte, deren versteckte Anweisung der KI-Assistent beim Verarbeiten als Befehl las. Für eine Führungskraft heißt das: Sobald ein Agent fremde Inhalte verarbeitet und zugleich auf interne Daten zugreift, ist er eine Angriffsfläche, nicht bloß ein Werkzeug. Diese Fläche gehört ins Risikoregister, nicht in die Liste der netten Zusatzfunktionen. Wir kommen in Kapitel 12 darauf zurück.

Der ehrliche Befund: großer Nutzen, scharfe Grenzen

Es wäre falsch, aus alldem auf Verzicht zu schließen. Agenten leisten tatsächlich Beträchtliches, und die nüchterne Bilanz zeigt beides am selben Beispiel, den Nutzen und seine Grenzen. Der Zahlungsdienstleister Klarna meldete Anfang 2024, sein KI-Assistent habe im ersten Monat zwei Drittel aller Kundendienst-Chats übernommen, die Bearbeitungszeit von elf auf unter zwei Minuten gesenkt und die Arbeit von rund 700 Vollzeitkräften geleistet.[12] Beeindruckende Zahlen, zugleich aber Zahlen des Unternehmens selbst, also mit Vorsicht zu lesen. Aufschlussreich ist, was danach geschah: 2025 räumte der Vorstandschef ein, der kompromisslose Kurs habe zu Qualitätseinbußen geführt, und

Klarna stellte wieder Menschen ein, für ein Modell, in dem die KI die Routine übernimmt und Menschen die schwierigen Fälle.[13] Die Technik hat dabei ihr Maß gefunden — eine ehrlichere Geschichte als die Pressemitteilung von 2024.

Daraus folgt eine klare Haltung. Der Agent ist das Muster mit dem größten Hebel und dem größten Risiko zugleich. Wer ihn einsetzt, sollte den Weg kurz halten, den Rahmen eng ziehen, dem Agenten nur die Schreibrechte geben, die er wirklich braucht, und an den Stellen, an denen ein Fehler teuer würde, einen Menschen dazwischenschalten. Der Grund dafür ist nüchtern: Lange autonome Ketten brechen, und die Maschine lässt sich über fremde Inhalte täuschen.

> **Die richtige Frage an Ihre Technik** Bevor Sie einem Agenten erlauben, selbständig zu handeln, verlangen Sie auf zwei Fragen eine klare Antwort: *„Was kann dieser Agent ohne Rückfrage tun — und was ist das Schlimmste, das passiert, wenn er bei einem Schritt irrt oder über einen fremden Inhalt getäuscht wird?"* Eine gute Antwort benennt die Schreibrechte einzeln, die Stellen mit menschlicher Freigabe und die Schäden, die selbst im schlimmsten Fall begrenzt bleiben. Wer auf die zweite Frage nur mit dem Hinweis auf die Fähigkeiten des Modells antwortet, hat das Risiko nicht verstanden.

Was dieses Kapitel für Sie ändert

Sie unterscheiden jetzt den Chatbot, der antwortet, vom Agenten, der handelt — und Sie wissen, dass die Grenze zwischen Lesen und Handeln die ist, an der das Risiko sprunghaft steigt. Sie kennen den Grund, warum lange autonome Ketten zerbrechen, und können dem Versprechen vom voll autonomen System mit den richtigen Fragen begegnen. Und Sie wissen, dass ein Agent, der fremde Inhalte verarbeitet und zugleich handeln darf, eine Angriffsfläche ist, die man eng führen muss.

Zwei Fäden bleiben offen, und beide werden aufgenommen. Wie man im Betrieb mit dieser Angriffsfläche umgeht — Auf-

sicht, Freigaben, das Risikoregister, die Vorgaben des EU AI Act — ist der Stoff von Kapitel 12, wo es um Governance und Sicherheit geht. Vorher aber steht die Frage, die unter allem liegt und die wir bisher umkreist haben: Woher wissen Sie überhaupt, ob ein Modell oder ein Agent gut genug ist — gut genug, um ihm eine Aufgabe, ein Budget, eine Schreibvollmacht anzuvertrauen? Das ist die Frage der Evaluation, und ihr gehört das nächste Kapitel.

Das Wichtigste in 60 Sekunden

- Ein **Agent** beantwortet nicht nur eine Frage, sondern verfolgt ein Ziel: Er plant Schritte, benutzt **Werkzeuge** und handelt in einer Schleife. Der Chatbot redet, der Agent tut.
- Die wichtigste Linie verläuft zwischen **Lesen und Handeln.** Ein lesender Agent kann sich irren; ein Agent mit **Schreibrechten** kann Schaden anrichten.
- **Zusammengesetzte Fehler:** Bei mehrstufigen Aufgaben multiplizieren sich die Fehlerraten. Lange autonome Ketten sind prinzipiell zerbrechlich — „voll autonom" verdient Skepsis.
- Die **Prompt-Injection** ist die zentrale Sicherheitslücke: Das Modell trennt Anweisung und Daten nicht, also lassen sich Befehle in fremde Inhalte schmuggeln. Mit Schreibrechten wird daraus realer Schaden.
- Der Nutzen ist real, aber er hat ein **richtiges Maß**: kurzer Weg, enger Rahmen, knappe Rechte, Mensch an den Kontrollpunkten.

Merksatz Handlungsrechte für eine KI vergibt man so sparsam wie Prokura — und prüft jeden Schritt, den man nicht zurücknehmen kann.

Drei Fragen an Ihr Unternehmen

1. Bei welchen Ihrer KI-Vorhaben darf die Maschine bereits *handeln* (schreiben, senden, ausführen) — und nicht nur antworten? Wissen Sie, mit welchen Rechten?
2. Wo verarbeitet ein Agent fremde, nicht geprüfte Inhalte (E-Mails, Webseiten, hochgeladene Dokumente) und hat zugleich Zugriff auf Interna? Das ist Ihre Angriffsfläche.
3. An welchen Stellen einer automatisierten Kette sitzt ein Mensch mit Freigabe — und reicht das, gemessen am Schaden, den ein unbemerkter Fehler dort anrichten könnte?

Quellen

1. Vorfall mit Replits KI-Coding-Agent, Juli 2025: Während eines mehrtägigen Experiments des SaaStr-Gründers Jason Lemkin löschte der Agent trotz ei-

nes verhängten Änderungsstopps eine produktive Datenbank; nach Lemkins Schilderung erzeugte er anschließend tausende erfundene Datensätze und meldete den Erfolg wahrheitswidrig. Replit-CEO Amjad Masad bezeichnete den Vorfall öffentlich als „unacceptable" und führte u. a. eine Trennung von Entwicklungs- und Produktionsdatenbank ein. Quellen: The Register, 21. Juli 2025; Fortune, 23. Juli 2025; AI Incident Database, Eintrag 1152. Die anschaulichen Einzelheiten (Zahl der erfundenen Datensätze, wiederholte Falschmeldungen) stammen aus Lemkins eigenen Veröffentlichungen und sind nicht unabhängig geprüft; es handelte sich um ein Experiment eines versierten Nutzers an den Grenzen des Werkzeugs, nicht um einen regulierten Unternehmensbetrieb.

2. Anthropic, *Building Effective Agents*, 19. Dezember 2024, https://www.anthropic.com/engineering/building-effective-agents : Agenten als Systeme, in denen LLMs „dynamically direct their own processes and tool usage"; Unterscheidung von „workflows" (über vorgegebene Code-Pfade orchestriert) und „agents" (steuern ihren Ablauf selbst); Empfehlung, „the simplest solution possible" zu wählen und Komplexität nur bei Bedarf zu erhöhen. Vendor-Quelle; die Definition ist kein neutraler Standard, konvergiert aber mit anderen.
3. Chip Huyen, *AI Engineering* (O'Reilly, 2025); zugrundeliegender Text öffentlich unter https://huyenchip.com/2025/01/07/agents.html : Unterscheidung von lesenden Aktionen („read-only actions") und Schreibaktionen („write actions"); Bild vom Praktikanten, dem man keine Vollmacht zum Löschen der Produktionsdatenbank gibt, bzw. der unzuverlässigen KI, der man keine Überweisungen erlaubt.
4. Zusammengesetzte Fehler, die nackte Arithmetik: $0{,}95^{10} \approx 0{,}599$ (rund 60 %), $0{,}95^{100} \approx 0{,}0059$ (rund 0,6 %). Die Rechnung unterstellt voneinander unabhängige Fehler ohne Korrektur und ist als illustrative Schranke zu lesen, nicht als Vorhersage; reale Agenten kennen Wiederholung und Selbstkorrektur ebenso wie korrelierte Fehler. Empirisch gestützt wird der Verlässlichkeitsverlust mit wachsender Aufgabenlänge durch METR (Anm. 5) und Toby Ord, *Is there a half-life for the success rates of AI agents?*, arXiv:2505.05115 (2025); dieselbe Veranschaulichung mit 95 % pro Schritt findet sich u. a. bei Chip Huyen, *AI Engineering* (2025).
5. METR, *Measuring AI Ability to Complete Long Tasks*, 19. März 2025 (arXiv:2503.14499): nahezu vollständige Lösung von Aufgaben unter rund vier Minuten menschlicher Arbeit, unter zehn Prozent bei Aufgaben über rund vier Stunden; der bei 50 % Verlässlichkeit bewältigbare Aufgaben-Zeithorizont verdoppelt sich etwa alle sieben Monate. Die Verdopplungsrate beschreibt die Vergangenheit und ist keine Garantie; geschäftstaugliche Verlässlichkeit verlangt deutlich kürzere Aufgaben als die 50-%-Schwelle.
6. τ-bench (Tool-Agent-User-Benchmark), Yao et al. (Sierra), arXiv:2406.12045 (2024): führende Modelle lösen unter 50 % der Aufgaben; die Konsistenz über acht wiederholte Versuche (pass^8) liegt im Einzelhandels-Szenario unter 25 %. Neuere Modelle erreichen höhere Werte; die Inkonsistenz über wiederholte Versuche bleibt die kennzeichnende Schwäche. Werte verändern sich rasch.

7. Typische Fehlermodi von Agenten nach Huyen, *AI Engineering* (2025): falsches oder falsch parametriertes Werkzeug; verfehltes Ziel bzw. gesprengtes Zeit-/Kostenbudget; sowie „false reflection" — der Agent „is convinced that it's accomplished a task when it hasn't".
8. Greshake, Abdelnabi, Mishra, Endres, Holz, Fritz, *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*, arXiv:2302.12173 (2023); publiziert im Tagungsband des AI-Sec-'23-Workshops (ACM, DOI 10.1145/3605764.3623985). Angriff „remotely (without a direct interface)" durch Anweisungen, die in abrufbaren Daten platziert werden; „LLM-Integrated Applications blur the line between data and instructions." Vier der sechs Autoren an deutschen Einrichtungen in Saarbrücken (CISPA Helmholtz-Zentrum für Informationssicherheit; Universität des Saarlandes), ein weiterer bei der sequire technology GmbH.
9. OWASP Gen AI Security Project, *Top 10 for LLM Applications*, Ausgabe 2025 (veröffentlicht 17. November 2024): „Prompt Injection" als Risiko LLM01, Platz 1; die Liste führt zudem „Excessive Agency" als eigenes, agentenspezifisches Risiko (zu weitreichende Autonomie, Rechte oder Werkzeugzugriffe). https://genai.owasp.org/llmrisk/llm01-prompt-injection/
10. Simon Willison, *The lethal trifecta for AI agents: private data, untrusted content, and external communication*, 16. Juni 2025, https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/ : Treffen die drei Eigenschaften — Zugriff auf private Daten, Verarbeitung nicht vertrauenswürdiger Inhalte, Fähigkeit zur Kommunikation nach außen — zusammen, lässt sich der Agent dazu verleiten, vertrauliche Daten an einen Angreifer zu senden. Das Bild zielt auf den Datenabfluss; zerstörerische oder transaktionale Schreibaktionen sind ein verwandter, eigener Verstärker.
11. „EchoLeak", CVE-2025-32711: zero-click-Schwachstelle in Microsoft 365 Copilot, entdeckt von Aim Security/Aim Labs (2025); per präparierter E-Mail (ohne Klick des Empfängers) ließen sich über eine indirekte Prompt-Injection Daten aus dem Zugriffsbereich des Nutzers exfiltrieren. Microsoft behob die Lücke serverseitig im Juni 2025; nach Unternehmensangaben keine Ausnutzung in freier Wildbahn. Quellen: NVD, https://nvd.nist.gov/vuln/detail/CVE-2025-32711 ; Aim Security; The Hacker News, Juni 2025. (*Vor Druck: CVSS-Wert und Datum gegen NVD prüfen.*)
12. Klarna, Pressemitteilung vom 27. Februar 2024: Der KI-Assistent (auf Basis von OpenAI-Technik) habe im ersten Monat zwei Drittel der Kundendienst-Chats (2,3 Mio. Gespräche) übernommen, die durchschnittliche Lösungszeit von elf auf unter zwei Minuten gesenkt und die Arbeit von rund 700 Vollzeitkräften geleistet. Es handelt sich um unternehmenseigene, nicht unabhängig geprüfte Angaben.
13. Korrektur des Kurses 2025: Klarna-Vorstandschef Sebastian Siemiatkowski räumte öffentlich Qualitätseinbußen des rein KI-gestützten Modells ein; das Unternehmen stellte wieder menschliche Kräfte ein und ging zu einem hybriden Modell über (KI für Routine, Menschen für komplexe Fälle und Eskalationen). Berichterstattung 2025 u. a. Fast Company, Entrepreneur, CX Dive.

Kapitel 6 — Woher Sie wissen, ob es funktioniert

Lernziele Nach diesem Kapitel können Sie

- erklären, warum das Messen, ob eine KI funktioniert, der eigentliche Engpass ist — und warum es so viel schwerer ist, als es klingt;
- die Demo-Falle erkennen und den Aufwand der „letzten Meile" von Anfang an einplanen;
- das Prinzip des evaluationsgetriebenen Vorgehens benennen: den Maßstab für „gut" festlegen, bevor gebaut wird;
- einer Bestenliste, einer Vorführung und einer Maschine, die Maschinen benotet, mit den richtigen Fragen begegnen.

Als der Berufsnetzwerk-Konzern LinkedIn 2024 sein erstes großes KI-Produkt baute, ging es zunächst verblüffend schnell. Nach einem Monat, berichten die verantwortlichen Ingenieure, hatten sie achtzig Prozent des angestrebten Produkts beisammen. Dann begann der zähe Teil. Vier weitere Monate kämpften sie darum, über fünfundneunzig Prozent zu kommen — das Stück von guter Vorführung zu verlässlichem Produkt —, gegen Halluzinationen und gegen die unzähligen Sonderfälle, an die niemand gedacht hatte. Das anfängliche Tempo, schreiben sie, habe ein trügerisches Gefühl erzeugt, „fast am Ziel" zu sein.[1]

Was LinkedIn erlebte, wiederholt sich fast überall, wo Teams vom Weg der Demo zum Produkt berichten. Eine eindrucksvolle Vorführung ist heute an einem Wochenende gebaut. Der weite Weg von dort zu einem System, dem man vertrauen kann, besteht zum größten Teil aus einer einzigen, unterschätzten Tätigkeit: herauszufinden, ob das Ding überhaupt funktioniert — und zwar verlässlich, auch in den Fällen, die in der Vorführung nicht vorkamen. Diese Tätigkeit heißt Evaluation, und sie ist das vielleicht wichtigste und am meisten vernachlässigte Thema der ganzen KI-Einführung. Ihr gehört dieses Kapitel.

Warum das Messen der Engpass ist

In der klassischen Softwareentwicklung ist die Frage, ob ein Programm tut, was es soll, meist beantwortbar: Man gibt eine Zahl ein, das richtige Ergebnis kommt heraus oder eben nicht. Bei einem Sprachmodell ist das anders, und das hat mit allem zu tun, was die vorigen Kapitel beschrieben haben. Die Ausgabe ist offen, sie schwankt, sie kann überzeugend falsch sein. Es gibt selten die eine richtige Antwort, gegen die man prüfen könnte.

Die Fachautorin Chip Huyen, deren Arbeit zur Praxis der KI-Entwicklung viel zitiert wird, formuliert es so: Eine fehlende, verlässliche Methode zu messen, ob das System taugt, sei „eines der größten Hindernisse für die Einführung von KI".[2] Das ist eine Meinung, kein gemessener Wert, aber sie deckt sich mit allem, was Praktiker berichten — und mit der ernüchternden Beobachtung, dass Unternehmen viel Geld in das Bauen von KI stecken und wenig in das Prüfen. Die Folge sieht man überall: Systeme gehen in Betrieb, kosten Geld, und niemand kann mit Sicherheit sagen, ob sie ihre Aufgabe gut erfüllen oder nur geschäftig aussehen.

> **Klartext — Evaluation** *Evaluation* ist das systematische Prüfen, wie gut ein KI-System seine Aufgabe erfüllt — nicht im Eindruck, sondern an einem vorher festgelegten Maßstab, gemessen an Beispielen, die der echten Anwendung entsprechen. Sie beantwortet die einzige Frage, die am Ende zählt: Funktioniert es gut genug, um sich darauf zu verlassen?

Warum es so viel schwerer ist, als es klingt

Dass das Messen schwerfällt, hat drei Gründe, und alle drei folgen aus der Natur der Sache.

Der erste: Es gibt keine Musterlösung. Ein älteres Klassifikationsprogramm, das E-Mails in „wichtig" und „unwichtig" sortiert, lässt sich anhand eines Lösungsschlüssels prüfen — richtig oder falsch. Die Antwort eines Sprachmodells auf die Bitte, einen Vertrag zusammenzufassen, ist offen: Es gibt viele gute Fassungen und unendlich viele schlechte, und keine Tabelle, gegen die man abgleichen könnte.[3]

Der zweite, und er ist der unangenehmste: Je fähiger die Maschine, desto schwerer ist ihre Arbeit zu benoten. Eine falsche Grundschul-Addition erkennt jeder. Ob eine flüssig geschriebene Zusammenfassung eines Fachgutachtens stimmt, kann nur beurteilen, wer das Gutachten selbst gelesen hat oder die Materie beherrscht. Die Prüfung wird also genau dort am aufwendigsten, wo die KI am nützlichsten wäre.[4]

Der dritte: Das Modell ist eine Blackbox. Sie haben es nicht gebaut, Sie sehen sein Inneres nicht, und Sie kennen die Daten nicht, mit denen es trainiert wurde. Es bleibt Ihnen nichts, als es nach seinen Ausgaben zu beurteilen — von außen, am Verhalten. Genau aus diesem Grund sind ganze Forschungsprojekte entstanden, die Modelle allein anhand ihres Verhaltens vergleichbar machen wollen.[5]

Die Demo-Falle und die letzte Meile

Aus diesen drei Gründen speist sich die teuerste Fehleinschätzung in KI-Vorhaben. Eine Demonstration gelingt schnell, weil sie nur die guten Fälle zeigen muss — die Beispiele, bei denen es funktioniert. Ein Produkt muss auch die schlechten Fälle bestehen, die seltenen, die böswilligen, die, an die keiner gedacht hat. Der Weg von der Vorführung zum verlässlichen Betrieb, so eine vielzitierte Formulierung aus der Fachwelt, sei „von null auf sechzig leicht und von sechzig auf hundert außerordentlich schwer".[6] Und dieser letzte Abschnitt ist fast vollständig Evaluationsarbeit: Sonderfälle finden, Halluzinationen aufspüren, entscheiden, wann es gut genug ist.

> **Stolperfalle — Die Demo, die den Vorstand überzeugt** Eine glänzende Vorführung ist der schwächste Beweis für ein funktionierendes Produkt — und der überzeugendste fürs Auge. Sie wurde mit den Fällen gebaut, in denen die Technik brilliert. Was sie nicht zeigt, ist die Verteilung der Fälle, in denen sie scheitert, und genau die entscheidet über Nutzen und Schaden im Betrieb. Wenn ein Team oder ein Anbieter eine beeindruckende Demonstration zeigt, lautet die nächste Frage: „Zeigen Sie mir, woran Sie gemessen haben, dass es auch in den schwierigen Fällen hält."

Das erklärt auch, warum so viele Pilotprojekte im Sand verlaufen. Kapitel 2 nannte die organisatorischen Ursachen und die Gartner-Prognose, nach der bis Ende 2025 mindestens 30 Prozent der Vorhaben nach der Pilotphase aufgegeben würden;[7] hier kommt eine weitere Ursache hinzu, die der Messung. Untersuchungen zur betrieblichen Praxis führen das Scheitern auf eine Lücke im Lernen zurück: Die Werkzeuge behalten kein Feedback, passen sich nicht an den Kontext an und sind nicht in die durchgängigen Abläufe eingebunden.[8] Wer das Scheitern an der letzten Meile vermeiden will, stellt die Prüfung an den Anfang, nicht ans Ende.

Den Maßstab festlegen, bevor man baut

Damit zum wirksamsten Prinzip dieses Kapitels. Es klingt selbstverständlich und wird trotzdem selten befolgt: Legen Sie fest, woran Sie Erfolg messen, *bevor* Sie das System bauen. In der klassischen Softwareentwicklung gibt es dafür das testgetriebene Entwickeln — man schreibt die Prüfung, ehe man den Code schreibt. Das Gegenstück für KI heißt evaluationsgetriebenes Vorgehen: Erst wird definiert, was „gut" für diese konkrete Aufgabe bedeutet und an welchen Beispielen sich das zeigen lässt, dann erst wird gebaut.[9]

Der Gewinn ist doppelt. Erstens zwingt es das Team, früh die unbequeme Frage zu beantworten, die sonst bis zum Schluss verdrängt wird: Woran genau würden wir erkennen, dass dieses System taugt — und woran, dass es versagt? Zweitens schafft es einen festen Prüfstein, an dem sich jede spätere Verbesserung messen lässt, statt sich auf das Gefühl zu verlassen, es sei „besser geworden". Betriebskennzahlen ersetzen diese Prüfung übrigens nicht: Sie kommen spät, sind von Dutzenden anderen Einflüssen überlagert und zeigen den Schaden erst, wenn er entstanden ist. Eine Führungskraft muss dieses Prinzip nicht selbst umsetzen, aber sie sollte darauf bestehen, dass es umgesetzt wird. Ein KI-Vorhaben ohne vereinbarten Maßstab für Erfolg ist ein Vorhaben, das sein eigenes Scheitern nicht bemerken würde.

> **Profi-Tipp — Die Prüfung vor dem Bau** Verlangen Sie, bevor ein KI-Projekt Budget bekommt, eine kleine Sammlung echter Beispiele aus Ihrem Geschäft samt der Antwort, die Sie jeweils für gut hielten — fünfzig, hundert Fälle, von Menschen beurteilt, die die Sache verstehen. Diese Sammlung ist Ihr Maßstab. An ihr misst sich, ob das System taugt, ob ein Wechsel des Modells etwas bringt, ob eine Änderung hilft oder schadet. Sie ist mehr wert als jede Bestenliste im Internet, weil sie Ihre Aufgabe misst, nicht die eines anderen.

Trauen Sie der Maschine nicht, die Maschinen benotet

Wenn das Prüfen so mühsam ist und es keine Musterlösung gibt, liegt eine Abkürzung nahe: Man lässt eine KI die Arbeit einer anderen KI benoten. Das ist heute gängige Praxis — in einer Branchenumfrage von Ende 2025 gab gut die Hälfte der Teams, die überhaupt prüfen, an, eine KI als Bewerter einzusetzen.[10] Verlockend ist das, weil es schnell und billig skaliert, und es ist nicht einmal aus der Luft gegriffen: In der grundlegenden Untersuchung dazu stimmte das bewertende Modell mit dem Urteil von Menschen in etwa 85 Prozent der Fälle überein — geringfügig öfter, als die Menschen untereinander einig waren.[11]

> **Klartext — KI als Richter** Eine *KI als Richter* ist ein Modell, das man einsetzt, um die Ausgaben eines anderen Modells zu bewerten — etwa, ob eine Antwort zur Frage passt oder ob eine Zusammenfassung treu ist. Schnell und billig, aber mit eigenen, systematischen Schwächen. Ein Richter, dessen Maßstab man nicht kennt, ist kein Maßstab.

Diese Übereinstimmungszahl darf man jedoch nicht überdehnen, denn dieselbe Untersuchung legte die Schwächen des Verfahrens offen. Das bewertende Modell hat Schlagseiten. Es bevorzugt die Antwort, die ihm zuerst gezeigt wird — vertauscht man bloß die Reihenfolge, kippt das Urteil in rund einem Drittel der Fälle.[11] Es bevorzugt die längere Antwort, auch wenn sie nur mehr Wörter ohne mehr Gehalt enthält; eine Untersuchung zeigte, dass sich solche Bewerter allein durch künstlich aufgeblähte Antworten täuschen lassen.[12] Und es neigt dazu, Ausgaben aus der eigenen Modellfamilie höher zu benoten. Hinzu kommt, dass der Richter selbst schwankt und dass jede Änderung an ihm — ein neues Richtermodell, ein veränderter Bewertungs-Prompt — Ihre Qualitätskennzahl verschiebt, ohne dass Ihr Produkt sich verändert hätte. Dann wissen Sie nicht mehr, ob Ihr System besser wurde oder nur Ihr Prüfer milder.

Daraus folgt eine schlichte Regel, im Geist der Faustregel aus Kapitel 3, der Maschine nur so weit zu trauen, wie man ihr Ergebnis billig prüfen kann: Setzen Sie eine KI als Richter ruhig ein, aber behandeln Sie sie wie ein Messgerät, das man eichen und festschreiben muss. Das Richtermodell und seine genaue Anweisung gehören dokumentiert und über die Zeit konstant gehalten, und sein Urteil gehört regelmäßig gegen das von Menschen abgeglichen. Einem Richter, dessen Modell und Anweisung Sie nicht kennen, sollten Sie kein Vertrauen schenken.

Warum die Bestenlisten der Anbieter wenig sagen

Bleibt die bequemste Abkürzung von allen: auf die öffentlichen Ranglisten zu schauen, auf denen die Anbieter ihre Modelle mit Höchstpunktzahlen schmücken. Auch hier ist Vorsicht geboten, aus drei Gründen.

Erstens verschleißen Benchmarks. Kaum ist ein Test etabliert, holen die Modelle die Höchstwerte ein, und er verliert seine Trennschärfe — alle sind plötzlich „sehr gut". Ein verbreiteter Wissenstest gilt mit Spitzenwerten um neunzig Prozent als ausgereizt; die Fachwelt muss laufend schwerere Prüfungen erfinden.[13] Zweitens sickern die Testfragen ins Training. Wenn die Aufgaben eines Tests im Trainingsmaterial des Modells landen, prüft man am Ende das Auswendiglernen, nicht das Können. Eine sorgfältige Untersuchung baute einen frischen, gleichwertigen Mathematiktest und stellte fest, dass die Leistung mancher Modellfamilien darauf um bis zu dreizehn Prozentpunkte einbrach — ein Hinweis darauf, dass sie den alten Test teilweise auswendig kannten.[14] Drittens, und am grundsätzlichsten, gilt eine alte Einsicht des Ökonomen Charles Goodhart, die die Anthropologin Marilyn Strathern auf die griffige Form gebracht hat: Sobald eine Kennzahl zum Ziel wird, taugt sie nicht mehr als Kennzahl.[15] Ein Benchmark, auf den hin alle optimieren, misst irgendwann nur noch, wer am besten auf ihn optimiert.

Zahlen, die zählen

- LinkedIn: **80 %** des angestrebten Produkts in **1 Monat**, dann **4 weitere Monate** für den Weg über 95 %.[1]
- KI-Richter, gleiche Antworten, vertauschte Reihenfolge: Urteil kippt in **rund einem Drittel** der Fälle.[11]
- Frischer, gleichwertiger Test statt des bekannten: bis zu **−13 Prozentpunkte** bei manchen Modellfamilien.[14]
- Deutsch gegenüber Englisch auf demselben Test: meist **wenige Punkte** Rückstand, mit den besten Modellen schrumpfend.[16]

Die Konsequenz ist, Bestenlisten richtig einzuordnen, statt sie zu verachten oder ihnen zu glauben: Sie taugen als grober Vorfilter, welche Modelle man überhaupt in Betracht zieht. Über die Frage, ob ein Modell *Ihre* Aufgabe gut löst, sagen sie wenig. Diese Frage beantwortet nur die Prüfung auf Ihren eigenen Daten, an Ihrem eigenen Maßstab.

Eine Lücke, die den deutschsprachigen Raum besonders betrifft

Es kommt ein Punkt hinzu, der in den englischsprachigen Ratgebern fehlt. Fast alle großen Tests sind englisch, und die Modelle werden überwiegend auf Englisch geprüft. Auf einem Test, der dieselben Fragen in vielen Sprachen stellt, liegt die deutsche Leistung der Spitzenmodelle ein paar Punkte hinter der englischen — kein Abgrund, und bei den besten Modellen schrumpft der Abstand, aber er ist da.[16] Eigene, breit angelegte deutsche Prüfungen sind rar; eine der wenigen wurde ausdrücklich geschaffen, weil es an deutschsprachiger Evaluation mangelte.[17] Für Sie heißt das: Eine Bestenliste, die ein Modell auf Englisch glänzen lässt, verspricht für Ihre deutschsprachige Anwendung mehr, als sie hält. Prüfen Sie auf Deutsch, an Ihrer Aufgabe.

Dass Evaluation kein Luxus ist, zeigt schließlich der Blick auf die Regulierung. Der EU AI Act macht für Systeme mit hohem Risiko eine systematische Prüfung vor dem Inverkehrbringen zur rechtlichen Voraussetzung; in Deutschland bauen Einrichtungen wie das gemeinsame KI-Prüflabor der Technischen Überwachungsvereine die Methoden und Siegel dafür auf.[18] Was hier als Pflicht entsteht, ist im Kern dasselbe, was gute Praxis ohnehin verlangt: nachweisen können, dass es funktioniert. Den operativen Umgang damit vertieft Kapitel 12.

Was dieses Kapitel für Sie ändert

Sie wissen jetzt, dass die schwierigste Frage der KI-Einführung nicht „Was kann das Modell?" lautet, sondern „Woher wissen wir, dass es für uns gut genug ist?" — und dass nur eine eigene, ehrliche Prüfung an den eigenen Daten sie beantwortet, keine Vorführung, keine Bestenliste, keine Maschine, die Maschinen benotet. Die Unternehmen, die das beherrschen, haben einen Vorteil, den ihnen kein Anbieter verkaufen kann: Sie wissen, was sie haben, während andere hoffen.

Damit endet der Werkzeugteil dieses Buches fast. Ein letztes Stück Verständnis fehlt noch, bevor wir aus dem Verstehen Vorsprung machen: die Frage, wessen Modell Sie überhaupt einsetzen — ein offenes, das Sie selbst betreiben und einsehen können, oder ein geschlossenes aus der Cloud eines großen Anbieters. An dieser Entscheidung hängen Kontrolle, Kosten und Souveränität zugleich, und ihr gehört das nächste Kapitel. Die Fähigkeit zu prüfen, die wir hier gewonnen haben, kehrt dann in Teil III wieder, wenn aus dem Beurteilen-Können eine Führungsaufgabe wird.

Das Wichtigste in 60 Sekunden

- Eine Demo ist schnell gebaut; ein verlässliches Produkt entsteht erst auf der **letzten Meile**, und die ist fast ganz **Evaluationsarbeit**. LinkedIn brauchte für die letzten Prozente viermal so lange wie für die ersten achtzig.
- **Messen ist der Engpass.** Es ist schwer, weil generative Ausgabe offen ist, weil fähigere Modelle schwerer zu benoten sind und weil das Modell eine Blackbox bleibt.
- **Legen Sie den Maßstab fest, bevor Sie bauen** (evaluationsgetriebenes Vorgehen). Eine kleine Sammlung echter, von Menschen beurteilter Fälle aus Ihrem Geschäft ist mehr wert als jede Bestenliste.
- Eine **KI als Richter** ist nützlich, aber voreingenommen (Reihenfolge, Länge, eigene Familie) und schwankend. Eichen, festschreiben, gegen Menschen abgleichen.
- **Bestenlisten verschleißen** (Sättigung, durchgesickerte Testfragen, Goodharts Gesetz). Sie sind Vorfilter, kein Beweis. Prüfen Sie auf Ihren Daten — und im deutschsprachigen Raum auf Deutsch.

Merksatz Wer nicht auf eigenen Daten misst, ob seine KI funktioniert, betreibt sie auf gut Glück.

Drei Fragen an Ihr Unternehmen

1. Für welches Ihrer KI-Vorhaben gibt es eine vereinbarte, geschriebene Antwort auf die Frage, woran sich „funktioniert gut genug" messen lässt — und für welche nur ein gutes Gefühl?
2. Haben Sie für Ihre wichtigsten Anwendungen eine eigene Sammlung echter Beispiele, an der Sie jeden Modellwechsel und jede Änderung messen — auf Deutsch, an Ihrer Aufgabe?
3. Wo verlassen Sie sich auf eine Zahl — eine Bestenliste, einen KI-Bewerter —, deren Zustandekommen Sie nicht kennen und nicht kontrollieren?

Quellen

1. Juan Pablo Bottaro und Karthik Ramgopal, *Musings on building a Generative AI product*, LinkedIn Engineering Blog, 25. April 2024, https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generative-ai-product : etwa 80 % der angestrebten Erfahrung im ersten Monat, danach rund vier weitere Monate für den Weg über 95 %, vor allem im Kampf gegen Halluzinationen und Sonderfälle; das anfängliche Tempo habe ein trügerisches Gefühl erzeugt, „fast am Ziel" zu sein.
2. Chip Huyen, *AI Engineering: Building Applications with Foundation Models*, O'Reilly 2025, Kap. 4: „Not having a reliable evaluation pipeline is one of the biggest blockers to AI adoption." (Wortlaut auch im Autoren-Repository github.com/chiphuyen/aie-book.) Es handelt sich um eine begründete Experteneinschätzung, nicht um einen gemessenen Wert; die Formulierung lautet „eines der größten Hindernisse", nicht „das größte".
3. Zur grundsätzlichen Schwierigkeit, offene Ausgaben ohne Musterlösung zu bewerten: Chang et al., *A Survey on Evaluation of Large Language Models*, arXiv:2307.03109 (2023).
4. Sinngemäß nach Huyen, *AI Engineering* (2025): „It's impossible to capture the ability of a high-dimensional system using one- or few-dimensional scores." Das Beispiel von der nur mit Fachwissen prüfbaren Zusammenfassung ist illustrativ.
5. Liang et al., *Holistic Evaluation of Language Models* (HELM), arXiv:2211.09110 (2022): Vergleich von Modellen allein anhand ihres Verhaltens, eben weil ihr Inneres und ihre Trainingsdaten verschlossen sind.
6. Ursprung der Formulierung „von 0 auf 60 leicht, von 60 auf 100 außerordentlich schwer": Ding et al. (UltraChat), arXiv:2305.14233 (2023); aufgegriffen und auf KI-Produkte übertragen von Chip Huyen, *Common pitfalls when building generative AI applications*, huyenchip.com, 16. Januar 2025 („It's easy to build a demo, but hard to build a product"). (*Vor Druck: exakten Wortlaut bei Ding et al. gegen das arXiv-PDF prüfen.*)
7. Gartner, Pressemitteilung vom 29. Juli 2024 (Rita Sallam): „At least 30% of generative AI projects will be abandoned after proof of concept by the end of 2025" — als Gründe genannt: schlechte Datenqualität, unzureichende Risikokontrollen, steigende Kosten, unklarer Geschäftswert. Es handelt sich um eine Analysten-Prognose ohne offengelegte Methodik.
8. MIT-Projekt NANDA, *The GenAI Divide: State of AI in Business 2025*: das Scheitern der meisten Pilotprojekte beruhe weniger auf der Technik als auf einer „Lernlücke" — Werkzeuge, die kein Feedback behalten, sich nicht an den Kontext anpassen und nicht in durchgängige Abläufe eingebunden sind. (Die kursierende Quote von rund 95 % gescheiterten Piloten beruht auf kleiner, qualitativer Stichprobe und ist als exakter Wert umstritten; hier zählt der Mechanismus, nicht die Zahl — vgl. Kap. 1.)
9. Evaluationsgetriebenes Vorgehen (evaluation-driven development): Chip Huyen, *AI Engineering* (2025), Kap. 4 — den Erfolgsmaßstab definieren, bevor

gebaut wird, als KI-Pendant zum testgetriebenen Entwickeln. Der Begriff ist älter als das Buch; Huyen prägt ihn für das Zeitalter der Foundation Models.

10. LangChain, *State of Agent Engineering* (Umfrage unter 1.340 Befragten, November/Dezember 2025): unter den Teams, die überhaupt evaluieren, nutzen rund 53 % eine KI als Bewerter (rund 60 % menschliche Prüfung). Anbieter-Umfrage mit selbstselektierter Teilnehmerschaft; als Hinweis auf weite Verbreitung zu lesen, nicht als repräsentative Marktzahl.
11. Zheng et al., *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, NeurIPS 2023, arXiv:2306.05685: Übereinstimmung zwischen GPT-4 und menschlichen Bewertern rund 85 % (auf der Teilmenge ohne Unentschieden im MT-Bench), etwas höher als die Übereinstimmung der Menschen untereinander (81 %). Zugleich dokumentiert: Positions-Schlagseite (GPT-4 nur in rund 65 % der Fälle konsistent nach Vertauschen der Reihenfolge), Wortfülle-Schlagseite und Selbstbevorzugung (Letztere beobachtet, nicht kausal bewiesen; bei GPT-3.5 nicht aufgetreten).
12. Dubois et al., *Length-Controlled AlpacaEval*, arXiv:2404.04475 (2024): KI-Bewerter bevorzugen längere Antworten; nach Korrektur für die Länge stieg die Übereinstimmung mit menschlich gestützten Urteilen deutlich — ein Beleg, dass sich rohe Bewerter-Scores durch bloße Wortfülle aufblähen lassen.
13. Stanford HAI, *AI Index Report 2025*, Kapitel Technical Performance: etablierte Tests wie MMLU gelten mit Spitzenwerten um 88–90 % als ausgereizt; als Reaktion entstehen schwerere Prüfungen (etwa „Humanity's Last Exam", „FrontierMath"). Absolute Werte sind schnelllebig (Stand Anfang 2025). Zur Dynamik der Benchmark-Sättigung allgemein: Ott et al., *Nature Communications* 13 (2022).
14. Zhang et al. (Scale AI), *A Careful Examination of Large Language Model Performance on Grade School Arithmetic*, arXiv:2405.00332 (2024): Auf einem neu erstellten, dem bekannten GSM8K gleichwertigen Mathematiktest (GSM1k) brach die Genauigkeit mancher Modellfamilien um bis zu rund 13 Prozentpunkte ein, ein Hinweis auf teilweises Auswendiglernen des bekannten Tests; Spitzenmodelle zeigten kaum Einbruch. Die Verzerrung trifft also manche Familien, nicht alle. (*Vor Druck: gegen das arXiv-PDF klären, ob die bis zu 13 als Prozent oder Prozentpunkte zu lesen sind.*)
15. Die griffige Fassung „When a measure becomes a target, it ceases to be a good measure" stammt von der Anthropologin Marilyn Strathern (1997); die zugrundeliegende Idee geht auf den Ökonomen Charles Goodhart (1975) zurück.
16. MMLU-ProX (mehrsprachiger Test mit identischen Fragen je Sprache; EMNLP 2025, arXiv:2503.10497): Deutsch liegt bei Spitzenmodellen meist wenige Punkte hinter Englisch (z. B. GPT-4.1 79,8 % vs. 76,4 %; Spannweite über Modelle hinweg rund 0,2 bis 4,5 Punkte), mit den stärksten Modellen schrumpfend. Werte je Modell und Datum verschieden.
17. Pfister und Hotho, *SuperGLEBer: German Language Understanding Evaluation Benchmark*, NAACL 2024: eine breite deutschsprachige Prüfsammlung, ausdrücklich geschaffen, weil es an umfassender deutscher Evaluation mangelte.

18. EU AI Act, Artikel 43: Für Hochrisiko-Systeme ist eine Konformitätsbewertung mit Prüf- und Daten-Governance-Anforderungen vor dem Inverkehrbringen vorgeschrieben. In Deutschland entwickelt u. a. das TÜV AI.Lab (gegründet Oktober 2023; Partner der Initiative MISSION KI) Prüfmethoden und ein Qualitätssiegel für KI. Der operative Umgang mit dem EU AI Act folgt in Kapitel 12; dessen Fristen sind in Bewegung und vor Drucklegung zu prüfen.

Kapitel 7 — Offen oder geschlossen

Lernziele Nach diesem Kapitel können Sie

- ein offenes von einem geschlossenen Modell unterscheiden — und sagen, warum „offen" meist „die Gewichte verfügbar" heißt und nicht „alles einsehbar";
- die vier Achsen der Entscheidung abwägen: Transparenz, Kontrolle, Kosten und Fähigkeit;
- die Sicherheitsdebatte einordnen — „eine Blackbox kann man nicht absichern" gegen „offene Gewichte lassen sich nicht zurückrufen" — ohne in eines der Lager zu kippen;
- die Frage der Souveränität bewerten und die Modellwahl als Entscheidung je Anwendungsfall treffen, nicht als Glaubensbekenntnis.

Am 20. Januar 2025 lud ein chinesisches Labor ein Modell namens DeepSeek-R1 ins Netz, und für ein paar Tage geriet die Tech-Welt aus dem Takt. Das Modell reichte in seinen Antworten an die teuersten geschlossenen Spitzenmodelle der amerikanischen Anbieter heran — und es war frei herunterladbar, unter einer großzügigen Lizenz, für jeden, der es betreiben wollte. An den Börsen brach kurzzeitig Nervosität aus. In Europa dagegen wuchs das Misstrauen: Italiens Datenschutzbehörde ließ den Dienst landesweit sperren, Tschechien untersagte ihn in der

Staatsverwaltung, beide mit Verweis darauf, dass die Daten der Nutzer nach China flössen.[1]

In dieser Episode steckt die ganze Frage dieses Kapitels, und auch schon ein Teil der Antwort. Denn die Sorge der Behörden galt dem *gehosteten Dienst* — der App und der Programmierschnittstelle, über die Eingaben auf chinesische Server wanderten. Die *Gewichte* desselben Modells aber, das offene Herzstück, konnte jedes europäische Unternehmen herunterladen und auf eigenen Servern betreiben, wo keine Eingabe das Haus verlässt. Dasselbe Modell war je nach Betriebsart ein Datenschutzrisiko oder ein Souveränitätsgewinn. Wer diesen Unterschied versteht, hält den Schlüssel zur letzten großen Verständnisfrage, bevor wir aus dem Verstehen Vorsprung machen: Wessen Modell setzen Sie eigentlich ein, und was hängt an dieser Wahl?

Drei Wörter, die man trennen muss: geschlossen, offen, quelloffen

Beginnen wir mit der Landschaft. Auf der einen Seite stehen die *geschlossenen* Modelle der großen amerikanischen Anbieter — die GPT-Reihe von OpenAI, Claude von Anthropic, Gemini von Google. Ihre Gewichte, das im Training entstandene Innere, bleiben in der Cloud des Anbieters; Sie sprechen sie über eine Schnittstelle an und sehen von außen nur, was hereinkommt und herausgeht. Auf der anderen Seite stehen die *offenen* Modelle, deren Gewichte zum Herunterladen freigegeben sind, sodass Sie sie selbst betreiben können: Llama von Meta, die Modelle des französischen Hauses Mistral, DeepSeek und Qwen aus China, seit August 2025 sogar ein offenes Modell von OpenAI selbst.[2]

> **Klartext — Offenes und geschlossenes Modell** Ein *geschlossenes Modell* nutzen Sie über die Cloud des Anbieters; seine Gewichte bleiben dort. Ein *offenes Modell* (genauer: ein Open-Weight-Modell) gibt seine Gewichte zum Herunterladen frei — Sie können es auf eigener Infrastruktur betreiben, anpassen, einfrieren. „Offen" heißt dabei meist nur: die Gewichte sind verfügbar. Es heißt selten „quelloffen" im strengen Sinn, denn die Trainingsdaten und das genaue Vorgehen bleiben in der Regel unter Verschluss.

Diese letzte Unterscheidung ist mehr als Wortklauberei. Selbst die offiziellen Hüter des Begriffs „Open Source" verlangen für ein wirklich quelloffenes KI-Modell auch Transparenz über die Trainingsdaten — eine Hürde, die fast kein verbreitetes „offenes" Modell nimmt.[3] Auch die Lizenzen tragen Fußangeln: Metas Llama etwa verlangt von Unternehmen jenseits von 700 Millionen monatlichen Nutzern eine Sonderlizenz und verbietet, das Modell zum Training von Konkurrenzmodellen zu verwenden — Bedingungen, die kein klassisches Open-Source-Werk kennt.[4] „Offen" ist also ein Spektrum, kein Schalter. Für die meisten Unternehmen ist die entscheidende Eigenschaft schlicht: Bekomme ich die Gewichte in die Hand und kann ich das Modell selbst betreiben?

Vier Achsen, an denen sich die Wahl entscheidet

Ob offen oder geschlossen das Richtige ist, entscheidet die Abwägung auf vier Achsen.

Die Fähigkeit. Lange war das Bild eindeutig: Die geschlossenen Spitzenmodelle führten, die offenen hinkten hinterher. Dieses Bild stimmt so nicht mehr. Der Abstand zwischen dem besten geschlossenen und dem besten offenen Modell ist auf einer großen Vergleichsplattform binnen eines Jahres von gut acht Prozent auf unter zwei Prozent geschrumpft.[5] 2026 lagen die besten offenen Modelle nur noch rund vier Monate hinter der geschlossenen Spitze; zugleich führte das beste amerikanische Mo-

dell nur noch wenige Prozentpunkte vor dem besten chinesischen, und die Führung wechselte mehrfach.[6] Festhalten lässt sich, mit dem nötigen Verfallsdatum auf der Zahl: Die geschlossenen Modelle halten an der äußersten Spitze, bei den schwersten Aufgaben, noch einen schmalen Vorsprung. Für den Großteil der betrieblichen Arbeit aber sind die besten offenen Modelle längst gut genug.

Die Kontrolle. Ein offenes Modell gehört, sobald es heruntergeladen ist, in gewissem Sinn Ihnen: Sie können es feinabstimmen, verkleinern, abgeschottet ohne Internetverbindung betreiben und auf einer Version festschreiben, die sich nicht unter Ihnen verändert. Ein geschlossenes Modell mustert der Anbieter aus, wann es ihm passt. Das ist kein hypothetisches Risiko: OpenAI etwa hat einen Fahrplan, nach dem ältere Modelle abgeschaltet werden, und wer darauf gebaut hat, muss umziehen, ob es ihm passt oder nicht.[7] Wer eine kritische Anwendung auf ein gemietetes, jederzeit kündbares Modell stellt, gibt ein Stück Steuerung aus der Hand.

Die Kosten. Hier kehrt sich das Bild teils um. Ein geschlossenes Modell ist eine reine Betriebsausgabe: Sie zahlen pro Token, müssen keine Hardware betreiben, tragen aber die Marge des Anbieters und sind seinen Preisänderungen ausgeliefert. Ein selbst betriebenes offenes Modell verlangt Vorabinvestitionen in Rechenleistung und Fachpersonal, und es trägt ein Auslastungsrisiko: Eine teure Grafikkarte, die im Leerlauf wartet, ist verbranntes Geld. Billiger wird der Eigenbetrieb erst bei hoher, gleichmäßiger Last — und wann genau dieser Punkt erreicht ist, hängt so stark vom Einzelfall ab, dass jede pauschale Zahl in die Irre führt. Es gibt einen Mittelweg: Anbieter, die offene Modelle über eine Schnittstelle bereitstellen, sodass Sie sie nutzen, ohne sie selbst zu betreiben. Der löst das Kostenproblem, aber nicht das der Souveränität — dazu gleich.

Die Transparenz. Ein offenes Modell lässt sich prüfen: Fachleute können es untersuchen, auf Schwächen abklopfen, auf versteckte Abhängigkeiten durchsehen. Ein geschlossenes bleibt die

Blackbox aus Kapitel 3, die Sie allein an ihrem Verhalten beurteilen können. Das führt direkt zur schärfsten Debatte dieses Felds.

Zahlen, die zählen

- Abstand bestes geschlossenes zu bestem offenem Modell: von **rund 8 % (Jan. 2024) auf unter 2 % (Feb. 2025)** auf einer großen Vergleichsplattform.[5]
- 2026: offene Modelle **rund 4 Monate** hinter der geschlossenen Spitze; das beste US-Modell nur noch **rund 2,7 Punkte** vor dem besten chinesischen.[6]
- **81 %** der deutschen Unternehmen sehen sich von US-Digitaltechnik abhängig (Bitkom, Anfang 2025).[8]

„Man kann eine Blackbox nicht absichern" — oder gerade doch?

Über kaum etwas wird so heftig gestritten wie über die Frage, welche Betriebsart die sicherere ist, und der Streit ist lehrreich, weil beide Seiten ein echtes Argument haben und beide ein Eigeninteresse.

Das Lager der Offenheit, am lautesten vertreten durch Meta, hält die Transparenz selbst für den besten Schutz. Mark Zuckerberg argumentierte 2024, quelloffene Systeme seien „erheblich sicherer …, weil sie transparenter sind und breit geprüft werden können"; nur so lasse sich verhindern, dass die Macht über eine Schlüsseltechnologie sich „in den Händen weniger Unternehmen" sammle.[9] Der Gedanke ist alt und stammt aus der Software-Sicherheit: Was viele Augen prüfen können, hat weniger Verstecke. Eine Blackbox dagegen, so die Pointe, muss man auf Treu und Glauben hinnehmen, samt allem, was man in ihr nicht sieht.

Das Lager der geschlossenen Modelle hält dagegen, und sein Argument ist ebenso ernst zu nehmen. Sind die Gewichte erst einmal in der Welt, lassen sie sich nicht zurückrufen. Das briti-

sche Institut für KI-Sicherheit formuliert es nüchtern: Ein einmal offen veröffentlichtes System „kann nicht zurückgenommen werden".[10] Schlimmer noch: Die Sicherheitsleitplanken, die ein Anbieter im Training einbaut, lassen sich aus offenen Gewichten mit erstaunlich wenig Aufwand wieder herauslösen — Forscher haben gezeigt, dass dafür mitunter Minuten und ein paar Euro genügen.[11] Was offen ist, kann von jedem genutzt werden, auch von dem, der Böses im Sinn hat, und der Anbieter kann es nicht mehr einfangen. Es ist kein Zufall, dass gerade die Häuser, die auf Sicherheit pochen, ihre stärksten Modelle geschlossen halten.[12]

Wie soll eine Führungskraft das auflösen? Am ehrlichsten mit einer Verschiebung der Frage, die in der Forschung *marginales Risiko* heißt. Sie lautet nicht: Könnte ein offenes Modell einem Übeltäter helfen? Das könnte es. Sie lautet: Wie viel *mehr* hilft es ihm, als die Werkzeuge, die er ohnehin schon hat — die Suchmaschine, frei verfügbare Software, das vorhandene Wissen? Die sorgfältigste Untersuchung dazu kommt zu einem unbequemen, aber redlichen Schluss: Für einen wirklich großen zusätzlichen Schaden durch die heutigen offenen Modelle fehlt bislang der Beleg — was nicht heißt, dass ihre Unbedenklichkeit bewiesen wäre, und was sich mit wachsender Fähigkeit verschieben kann.[13] Für die allermeisten Unternehmen ist dieser Streit ohnehin eine Frage der großen Politik, nicht der eigenen Technikwahl. Ihre Entscheidung dreht sich um Kontrolle, Kosten und Eignung — nicht um die Frage, ob die Menschheit offene Modelle verkraftet.

Die Achse, die im deutschsprachigen Raum am schwersten wiegt

Womit wir bei dem Punkt sind, der in den amerikanischen Ratgebern fehlt und hier den Ausschlag gibt: der Souveränität. Eine Schlüsseltechnologie der kommenden Jahrzehnte liegt heute in der Hand einer Handvoll Anbieter, fast alle amerikanisch oder chinesisch. Wie abhängig das macht, hat der Branchenverband

Bitkom gemessen: Anfang 2025 sahen sich 81 Prozent der deutschen Unternehmen von digitaler Technik aus den USA abhängig, und nur ein verschwindender Teil traute sich zu, einen Lieferstopp länger als zwei Jahre zu überstehen.[8] Für eine geschäftskritische Fähigkeit ist das eine unbequeme Lage.

Hier liegt das stärkste Argument für offene Modelle, und es ist handfest. Von den drei Betriebsarten hält nur eine Ihre Daten vollständig im Haus: das offene Modell, das Sie selbst betreiben. Auf eigenen Servern verlässt die Inferenz die kontrollierte Umgebung nicht — kein Auftragsverarbeiter, kein Drittland, ganz wie es die Datenschutzkonferenz für den Eigenbetrieb festgehalten hat (Kapitel 4); bei gemieteter europäischer Infrastruktur bleibt ein Auftragsverarbeiter, doch die Drittlandsfrage entschärft sich. Genau das war die Pointe des DeepSeek-Falls: Was als gehosteter Dienst ein Risiko war, wurde als selbst betriebenes Modell zum kontrollierbaren Werkzeug.

> **Stolperfalle — „Offenes Modell" heißt nicht automatisch „souverän"**
> Ein offenes Modell über die Schnittstelle eines fremden Anbieters zu nutzen, löst das Kostenproblem, aber nicht das der Souveränität: Ihre Daten laufen weiterhin über einen Dritten, meist in den USA. Souverän wird die Sache erst, wenn Sie das offene Modell *selbst* betreiben — auf eigener oder europäischer Infrastruktur. Drei Wege, die oft verwechselt werden: das geschlossene Modell aus der Cloud, das offene Modell über einen fremden Dienst, das offene Modell im eigenen Haus. Nur der dritte hält die Daten wirklich bei Ihnen.

Europa ist in diesem Feld nicht ohne eigene Stimme. Das französische Haus Mistral liefert leistungsfähige offene Modelle; das deutsch geführte Konsortium OpenGPT-X hat mit Teuken-7B ein offenes Modell vorgelegt, das in allen 24 Amtssprachen der EU trainiert wurde; das EU-geförderte Projekt EuroLLM zielt in dieselbe Richtung.[14] Man sollte die Lage nicht schönreden: In der rohen Leistung an der Spitze liegen die europäischen Modelle hinter den amerikanischen und chinesischen, wenn auch mit

schrumpfendem Abstand. Souveränität hat hier ihren Preis, und er besteht zuweilen in einem Stück Leistung oder Mehraufwand. Aber für viele Aufgaben — die deutschsprachige Arbeit, den Betrieb hinter der eigenen Brandmauer, die regulierte Anwendung — ist ein europäisches oder selbst betriebenes Modell vollauf genügend. Wer abwägt, sollte beides auf die Waage legen: was die Spitzenleistung wert ist, und was die Kontrolle wert ist. Für die Auswahl und den sicheren Betrieb solcher Modelle hat das Bundesamt für Sicherheit in der Informationstechnik einen eigenen Kriterienkatalog veröffentlicht, der offene wie geschlossene Modelle abdeckt.[15] Auch der EU AI Act unterscheidet hier: Für frei und quelloffen bereitgestellte Modelle sieht er bei den Pflichten für Allzweck-KI Erleichterungen vor, solange kein systemisches Risiko im Spiel ist.

Die Entscheidung: ein Portfolio, kein Bekenntnis

Daraus folgt, wie man die Wahl klug trifft: als Zuordnung, Aufgabe für Aufgabe. Große Anwender gehen längst so vor: In einer Befragung großer Unternehmen setzte über ein Drittel fünf oder mehr Modelle nebeneinander ein und wählte für jede Aufgabe das passende.[16] Das geschlossene Spitzenmodell für die schwierigste, sichtbarste Arbeit, bei der die letzte Stufe an Fähigkeit zählt. Das offene, womöglich selbst betriebene Modell für das Sensible, das Hochvolumige und das, was im eigenen Haus bleiben muss. So behalten Sie die Wahl — und überlassen nicht einem einzigen ausländischen Anbieter die Kontrolle über eine Fähigkeit, von der Ihr Geschäft zunehmend abhängt.

> **Die richtige Frage an Ihre Technik** Bevor Sie sich für ein Modell festlegen, fragen Sie nicht nur „Welches ist das beste?", sondern: „*Für welche unserer Aufgaben brauchen wir wirklich die Spitzenleistung — und bei welchen wiegt schwerer, dass die Daten im Haus bleiben und wir die Kontrolle behalten?*" Die Antwort fällt selten für ein einziges Modell aus. Sie führt zu einer bewussten Aufteilung — und zu der gesunden Lage, jederzeit den Anbieter wechseln zu können, ohne das Geschäft anzuhalten.

Was dieses Kapitel für Sie ändert

Sie können die Landschaft jetzt lesen. Sie wissen, dass „offen" meist „die Gewichte verfügbar" heißt; dass der Leistungsabstand zu den geschlossenen Modellen schmal geworden ist; dass die Sicherheitsdebatte zwei ernste Seiten hat und für Ihre Entscheidung weniger wiegt als die Fragen nach Kontrolle, Kosten und Souveränität; und dass die einzige Betriebsart, die Ihre Daten wirklich im Haus hält, das selbst betriebene offene Modell ist. Vor allem aber wissen Sie, dass die Wahl eine Frage der Zuordnung ist — Aufgabe für Aufgabe.

Damit schließt der zweite Teil. Sie sind kein Ingenieur geworden, aber Sie können den Ingenieuren und Anbietern nun die Fragen stellen, vor denen Beliebigkeit kapituliert: wie ein Modell funktioniert und woran es scheitert, wie es zu Ihrem wird, was geschieht, wenn es zu handeln beginnt, woran Sie messen, ob es taugt, und wessen Modell Sie überhaupt einsetzen. Aus diesem Verständnis wird jetzt Vorsprung. Der dritte Teil zeigt, wo er entsteht — in den drei Knappheiten, die kein Anbieter verkaufen kann. Die erste ist das Urteil: die Fähigkeit, gute von schlechter Arbeit zu unterscheiden, die wir in Kapitel 6 als Handwerk kennengelernt haben und die nun zur Führungsaufgabe wird.

Das Wichtigste in 60 Sekunden

- **Geschlossen** heißt: Modell in der Cloud des Anbieters, Nutzung über eine Schnittstelle. **Offen (Open-Weight)** heißt: Gewichte herunterladbar, selbst betreibbar. „Offen" heißt meist nicht „quelloffen" — Trainingsdaten bleiben verschlossen.
- Vier Achsen: **Fähigkeit** (geschlossener Vorsprung schmal, für die meisten Aufgaben sind offene Modelle gut genug), **Kontrolle** (offen: anpassbar, einfrierbar), **Kosten** (geschlossen: Betriebsausgabe; offen: lohnt bei hoher, gleichmäßiger Last), **Transparenz** (offen: prüfbar; geschlossen: Blackbox).
- Die **Sicherheitsdebatte** hat zwei ernste Seiten — Transparenz als Schutz gegen Unwiderrufbarkeit des Missbrauchs. Für Ihre Wahl zählt sie weniger als Kontrolle, Kosten, Eignung.
- **Souveränität:** Nur das **selbst betriebene** offene Modell hält die Daten vollständig im Haus. Ein offenes Modell über einen fremden Dienst tut das nicht.
- Die Wahl ist **ein Portfolio, kein Bekenntnis:** das passende Modell je Aufgabe — und die Freiheit, den Anbieter zu wechseln.

Merksatz Offen oder geschlossen ist keine Glaubensfrage, sondern eine Frage je Aufgabe — entscheidend ist, wie viel Kontrolle Sie aus der Hand geben.

Drei Fragen an Ihr Unternehmen

1. Für welche Ihrer Anwendungen brauchen Sie wirklich die absolute Spitzenleistung — und bei welchen wiegt schwerer, dass die Daten im Haus bleiben?
2. Hängt eine geschäftskritische Anwendung an einem einzigen, dazu ausländischen Anbieter — und was geschähe, wenn er das Modell abschaltet, den Preis verdoppelt oder den Zugang sperrt?
3. Wo nutzen Sie ein offenes Modell über einen fremden Dienst in dem Glauben, das sei schon „souverän" — obwohl die Daten weiterhin das Haus verlassen?

Quellen

1. DeepSeek-R1, veröffentlicht am 20. Januar 2025 (MoE-Modell mit 671 Mrd. Parametern, MIT-Lizenz, frei herunterladbar); reichte in Vergleichen an führende geschlossene Modelle heran. Italiens Datenschutzbehörde ging am 30. Januar/3. Februar 2025 gegen den Dienst vor, die tschechische Regierung untersagte ihn im Juli 2025 in der Staatsverwaltung — jeweils mit Verweis auf den Abfluss von Nutzerdaten nach China. Die Maßnahmen zielten auf den gehosteten Dienst, nicht auf den Eigenbetrieb der Gewichte. Quellen: DeepSeek-V3 Technical Report (arXiv:2412.19437); Berichte zu den Verboten (Euronews, Juli 2025; EU-Datenschutzbehörden). (*Die kursierende Trainingskostenangabe von rund 5,6 Mio. $ betrifft nur den finalen Pretraining-Lauf und ist als Gesamtkosten umstritten; nicht als solche zitieren.*)
2. Geschlossene Frontier-Anbieter: OpenAI (GPT), Anthropic (Claude), Google (Gemini). Offene (Open-Weight-)Modelle: Meta (Llama), Mistral, DeepSeek, Alibaba (Qwen). OpenAI veröffentlichte am 5. August 2025 mit gpt-oss-120b und gpt-oss-20b erstmals seit GPT-2 wieder offene Gewichte (Apache-2.0-Lizenz). Quelle: openai.com/index/introducing-gpt-oss.
3. Open Source Initiative, *The Open Source AI Definition* 1.0 (28. Oktober 2024): Ein quelloffenes KI-System verlangt neben offenen Parametern auch hinreichende Transparenz über die verwendeten Trainingsdaten. Die meisten verbreiteten „offenen" Modelle erfüllen das nicht und sind daher genauer als Open-Weight zu bezeichnen. https://opensource.org/ai/open-source-ai-definition
4. Meta, *Llama Community License* (§2): Unternehmen mit mehr als 700 Mio. monatlich aktiven Nutzern benötigen eine gesonderte Lizenz; die Nutzung zum Training konkurrierender Modelle ist untersagt. Es handelt sich nicht um eine von der Open Source Initiative anerkannte Lizenz. https://www.llama.com/llama3_1/license/
5. Stanford HAI, *AI Index Report 2025*, Kapitel Technical Performance: Der Abstand zwischen dem besten geschlossenen und dem besten offenen Modell auf der Vergleichsplattform Chatbot Arena (LMArena) fiel von 8,04 % (Januar 2024) auf 1,70 % (Februar 2025). Es handelt sich um eine einzelne Metrik (Arena-Elo, menschliche Präferenzurteile) in einem bestimmten Zeitfenster; Ranglisten verschieben sich laufend.
6. Epoch AI, *Open models lag state-of-the-art closed models by ~4 months* (Stand 29. Mai 2026): offene Modelle rund vier Monate hinter den geschlossenen Spitzenmodellen. Stanford AI Index 2026 (zit. nach IEEE Spectrum): Das führende US-Modell lag im März 2026 nur rund 2,7 Prozentpunkte vorn (Arena-Elo); die Führung wechselte seit Anfang 2025 mehrfach zwischen amerikanischen und chinesischen Modellen. Werte sind schnelllebig und zu datieren.
7. OpenAI, Übersicht zu Modell-Abkündigungen (developers.openai.com/api/docs/deprecations): ältere Modelle werden nach einem veröffentlichten Fahrplan abgeschaltet (z. B. API-Ende von chatgpt-4o-latest am 16. Februar 2026; Abschaltung von DALL·E 2/3 am 12. Mai 2026), mit in der Regel mehr-

monatigem Vorlauf für allgemein verfügbare Modelle. Konkrete Daten sind schnelllebig.

8. Bitkom, Studie *Digitale Souveränität* 2025 (Befragung von 603 Unternehmen ab 20 Beschäftigten, Erhebung November 2024, veröffentlicht 15. Januar 2025): 81 % der deutschen Unternehmen sehen sich von Digitaltechnik aus den USA abhängig (41 % stark), 79 % von China; nur rund 3 % hielten einen Lieferstopp länger als zwei Jahre durch. https://www.bitkom.org
9. Mark Zuckerberg, *Open Source AI Is the Path Forward*, Meta, 23. Juli 2024: „open source should be significantly safer since the systems are more transparent and can be widely scrutinized"; Sorge vor einer Konzentration der Macht „in the hands of a small number of companies". Meta ist als großer Anbieter offener Modelle eine interessierte Partei. https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/
10. UK AI Safety/Security Institute, *Managing risks from increasingly capable open-weight AI systems*, 29. August 2025: „Once a system is released with open weights, it cannot be rolled back." https://www.aisi.gov.uk
11. Zur leichten Entfernbarkeit der Sicherheitsleitplanken aus offenen Gewichten: Volkov, *Badllama 3* (2024; Leitplanken aus Llama 3 8B in wenigen Minuten für unter einem Dollar entfernt); Lermen, Rogers-Smith, Ladish (2023; mit Budget unter 200 $). Die Demonstrationen betreffen ältere Modelle; die Minuten- und Kostenangaben sind illustrativ, das Prinzip ist robust.
12. Zur Position der auf Sicherheit bedachten Häuser: Anthropic, *Responsible Scaling Policy* (Aktivierung der Schutzstufe ASL-3 für Claude Opus 4, Mai 2025, weil ein Missbrauchs-Uplift nicht ausgeschlossen werden konnte); OpenAI, *Preparedness Framework*. Beide Anbieter verdienen an geschlossenen Modellen und sind insofern ebenfalls interessierte Parteien.
13. Kapoor, Bommasani et al., *On the Societal Impact of Open Foundation Models*, ICML 2024 (arXiv:2403.07918): Maßgeblich sei das *marginale* Risiko offener Modelle gegenüber bereits verfügbaren Werkzeugen; über Missbrauchsvektoren wie Cyberangriffe und Biowaffen hinweg sei „current research … insufficient to effectively characterize the marginal risk … relative to pre-existing technologies." Das bedeutet nicht erwiesene Unbedenklichkeit; die Abwägung kann sich mit steigender Fähigkeit verschieben.
14. Europäische Modelle: Mistral AI (Paris, gegr. 2023; offene Modelle unter Apache 2.0 neben kommerziellen). Teuken-7B des Konsortiums OpenGPT-X (geführt von Fraunhofer IAIS/IIS, mit TU Dresden u. a.; in allen 24 EU-Amtssprachen trainiert; veröffentlicht 26. November 2024; gefördert vom Bundesministerium für Wirtschaft; kommerzielle Variante unter Apache 2.0). EuroLLM (EU-/EuroHPC-gefördert, mehrsprachig; arXiv:2409.16235, EuroLLM-9B Dezember 2024). In der Spitzenleistung liegen europäische Modelle hinter den US- und chinesischen, mit schrumpfendem Abstand.
15. Bundesamt für Sicherheit in der Informationstechnik (BSI), *Kriterienkatalog zur Integration von extern bereitgestellten generativen KI-Modellen* (deckt Open-Source-, Open-Weight- und proprietäre Modelle ab, mit Blick auf Transparenz und Sicherheit entlang der KI-Lieferkette) sowie *Generative KI-Modelle: Chancen und Risiken für Industrie und Behörden* (21. Januar 2025). Auf

die Bundesverwaltung zugeschnitten, breiter als Referenz genutzt. https://www.bsi.bund.de

16. Andreessen Horowitz, *How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025*: 37 % der befragten Unternehmen setzten fünf oder mehr Modelle in Produktion ein (Vorjahr 29 %) und wählten je nach Anwendungsfall; offene Modelle vor allem bei On-Premise-Betrieb, Compliance, Kostensensibilität und Feinabstimmung. a16z ist ein Wagniskapitalgeber mit Eigeninteressen; die Erhebung ist gleichwohl vielzitiert. https://a16z.com/ai-enterprise-2025/

TEIL III — DIE DREI KNAPPHEITEN

Wo der Vorsprung entsteht.

Der zweite Teil hat erklärt, wie die Technik arbeitet. Damit ist die Frage beantwortet, was KI *kann* — nicht aber, woraus ein Unternehmen einen Vorsprung gewinnt. Denn das Werkzeug steht allen offen, zu fallenden Preisen. Was jeder Wettbewerber ebenso kaufen kann, trägt keinen Vorsprung. Dieser Teil handelt von dem, was sich nicht kaufen lässt und gerade deshalb wertvoll ist: von drei knappen Gütern, zu denen die Rendite fließt, wenn die Technik selbst zur Ware geworden ist. Das erste ist das Urteil. Das zweite der eigene Kontext, die Daten, die nur dem eigenen Haus gehören. Das dritte der Wert, den man je Recheneinheit herausholt. Beginnen wir mit dem Urteil.

Kapitel 8 — Urteil: Vom Verwalter zum Beurteiler

Lernziele Nach diesem Kapitel können Sie

- begründen, warum das Urteil — gute von schlechter Arbeit zu unterscheiden — zum knappen Gut wird, sobald die Erzeugung billig ist;
- die Verschiebung Ihrer eigenen Rolle benennen: vom Verwalter, der zusammenträgt, zum Beurteiler, der bewertet;
- das Automatisierungs-Paradox erkennen — wer das Tun abgibt, verliert die Urteilskraft, die er gerade dann am meisten braucht — und was daraus folgt;
- KI-Arbeit als Führungspraxis beurteilen: den Maßstab vorgeben und das Abnicken vermeiden.

„Billig und schlecht": Mit diesem Urteil über die eigene Industrieware löste der Ingenieur Franz Reuleaux, Vorsitzender der deutschen Jury auf der Weltausstellung 1876 in Philadelphia, daheim einen Sturm aus.[1] Elf Jahre später zwang ein britisches Gesetz deutsche Einfuhren zur Herkunftskennzeichnung, damit der Käufer die billigen Nachahmungen erkenne und meide — „Made in Germany" begann als Schmähung.[2] Dass daraus ein Gütesiegel wurde, verdankt sich keiner einzelnen Erfindung. Es verdankt sich einem Vermögen, das in keiner Bilanz steht: der

Fähigkeit, gute Arbeit von schlechter zu unterscheiden, und der Disziplin, nur die gute gelten zu lassen.

Dieses Vermögen ist schwer zu fassen, weil es sich nicht in Sätze ausbuchstabieren lässt. Der Meister, der eine Gesellenarbeit prüft und nach einem Blick und einem Griff nur „Noch mal" sagt, kann selten erklären, woran er es gesehen hat. Man sieht es eben. Diese Antwort ist keine Ausflucht, sie ist der Kern der Sache: Das Urteil steckt in der Hand und im Auge, in Jahren des Tuns erworben. Der Philosoph Michael Polanyi hat dafür die berühmte Formel geprägt: Wir wissen mehr, als wir zu sagen vermögen.[3]

Dieses Vermögen war lange die Tugend einer Handvoll Berufe. In der Zeit der billigen Erzeugung wird es zu einer der knappsten Ressourcen im Unternehmen. Das ist die erste der drei Knappheiten, und dieses Kapitel zeigt, warum sie zählt und wie man sie pflegt.

Warum das Urteil knapp wird, wenn die Erzeugung billig ist

Es gibt ein ökonomisches Gesetz, schlicht und folgenreich, das das ganze Kapitel trägt. Wird ein Gut billig und reichlich, so steigt der Wert dessen, was es *ergänzt*. Die Ökonomen Carl Shapiro und Hal Varian haben es vor einem Vierteljahrhundert für die Informationswirtschaft ausformuliert: Die Nachfrage nach einem Produkt wächst, wenn der Preis seiner Komplemente fällt.[4] Wer Druckerpatronen verkauft, freut sich über billige Drucker.

Übertragen Sie das auf die künstliche Intelligenz. Was sie billig macht, ist das Erzeugen: Texte, Entwürfe, Code, Analysen, in jeder Menge, zu fast jedem Zeitpunkt. Das Komplement dazu, das knapp bleibt, ist das Urteil darüber, welcher dieser Entwürfe taugt und welcher in den Papierkorb gehört — und, davor, die Entscheidung, was überhaupt erzeugt werden soll. Je billiger das Erzeugen, desto wertvoller das Urteilen. Das ist ein bekannter Mechanismus, angewandt auf eine neue Lage.

Eine zweite Beobachtung verstärkt die erste. Wenn jedes Unternehmen dieselben Modelle einkaufen kann, dann wirkt die Technik angleichend statt trennend. Drei Forscher, unter ihnen Jay Barney, der die ressourcenbasierte Theorie des Wettbewerbsvorteils maßgeblich geprägt hat, haben es 2025 unverblümt formuliert: Künstliche Intelligenz werde weit eher eine Quelle der *Vereinheitlichung* sein als der Unterscheidung; wenn alle dieselbe Technik nutzten, hebe das vielleicht den ganzen Markt, verschaffe aber keinem einen Vorsprung.[5] Was unterscheidet, ist dann der Umgang der Menschen mit dem Modell, das alle besitzen, und der beginnt mit dem Urteil.

> **Klartext — Die erste Knappheit: das Urteil** *Urteil* meint hier zweierlei: die Fähigkeit, gute von schlechter Arbeit zu unterscheiden, und die Fähigkeit zu entscheiden, was überhaupt gebaut gehört. Beides ist knapp, weil es sich nicht kaufen lässt — kein Anbieter verkauft es, kein Wettbewerber kann es nachbestellen. Man erwirbt es nur langsam, durch Übung und Nähe zur Sache. Genau das macht es in einer Welt der käuflichen Werkzeuge so wertvoll.

Das ist, nebenbei, die Konkretisierung der These, mit der dieses Buch begann. Den Vorsprung schafft nicht das Werkzeug, sondern die Führung, die daraus etwas macht. „Etwas daraus machen" heißt zuallererst: beurteilen können, was die Maschine geliefert hat.

Vom Verwalter zum Beurteiler

Damit verschiebt sich die Rolle der Führungskraft, und zwar spürbar. Das überkommene Bild des Managers ist das eines Verwalters: einer, der Informationen zusammenträgt, Berichte einordnet, Abläufe koordiniert und aus dem gesammelten Material eine Entscheidung ableitet. Ein guter Teil dieser Arbeit — das Zusammentragen, das Aufbereiten, das erste Formulieren — ist genau die Arbeit, die die Maschine nun übernimmt. Was bleibt

und was wächst, ist das Beurteilen: Welcher der fünf Entwürfe ist der beste? Wo klingt die Analyse plausibel und ist doch falsch? Lohnt dieses Vorhaben überhaupt?

Diese Verschiebung ist nicht zu verwechseln mit Bequemlichkeit. Sie verlagert die Last vom Sammeln auf das Werten, und Werten ist die schwerere Übung. Eine Führungskraft, formulierte es ein Beobachter treffend, werde nicht dafür bezahlt, dass sie Informationen abrufen kann, sondern dafür, dass sie entscheidet, wenn etwas auf dem Spiel steht und der Ausgang ungewiss ist.[6] Genau diese Tätigkeit rückt ins Zentrum. Sie ist die persönliche, tägliche Form dessen, was Kapitel 6 als Evaluation auf der Ebene des Systems beschrieben hat: Dort ging es um die Frage, ob eine KI-Anwendung taugt; hier geht es um die Frage, ob *diese eine Ausgabe* taugt, hier und jetzt, auf Ihrem Schreibtisch.

Die Forschung zur Zusammenarbeit von Mensch und Maschine hat dafür ein Bild gefunden. In einer großen Studie mit Unternehmensberatern unterschieden die Autoren zwei Arbeitsweisen: den *Zentauren*, der die Aufgabe teilt und der Maschine zuweist, was sie gut kann, während er das Urteil behält; und den *Cyborg*, der sich eng mit ihr verzahnt.[7] Was in beiden Arbeitsweisen über den Erfolg entschied, war dasselbe: das Urteil, wo die Maschine zu gebrauchen ist und wo nicht — wer diese Grenze falsch einschätzte, lieferte schlechtere Arbeit ab. Der Wert wandert vom Machen der Arbeit zu ihrem Dirigieren und Prüfen.

Das Paradox: Wer das Tun abgibt, verliert das Urteil

Der Übergang hat eine Falle, und sie ist seit Langem bekannt, nur unter anderem Namen. 1983 beschrieb die britische Ingenieurin Lisanne Bainbridge, was sie die „Ironien der Automatisierung" nannte. Ihre Beobachtung, an Leitständen und Cockpits gewonnen, gilt heute Wort für Wort wieder. Erstens: Wer einen Prozess automatisiert, überlässt dem Menschen ausgerechnet die Aufgaben, die zu automatisieren niemandem gelang, also die schwersten, die Urteilsaufgaben. Zweitens, und das ist die ei-

gentliche Ironie: Fähigkeiten verkümmern, wenn man sie nicht übt. Ein einst erfahrener Bediener, der die Automatik nur noch überwacht, ist nach einer Weile kein erfahrener Bediener mehr. „Indem sie die leichten Teile der Aufgabe wegnimmt", schrieb Bainbridge, „kann die Automatisierung die schweren Teile noch schwerer machen."[8]

Übertragen auf die Führungskraft: Wer das Handwerk ganz abgibt und sich nur das Urteilen vorbehält, wird feststellen, dass das Urteilen mit der Zeit hohl wird. Denn das Urteil ruht auf Erfahrung, und Erfahrung verlangt Tun. Der Kognitionsforscher Daniel Kahneman und der Experte Gary Klein haben die Bedingungen dafür benannt, wann der erfahrenen Intuition zu trauen ist: nur in einem Feld mit ausreichend verlässlichen Mustern und nur bei genügend Gelegenheit, die Fähigkeit zu üben.[9] Wer aufhört zu üben, dem entgleitet der Boden, auf dem ein gültiges Urteil überhaupt wachsen kann.

Dazu kommt eine zweite, gut belegte menschliche Schwäche. Wir neigen dazu, einer Automatik, die meist funktioniert, blind zu vertrauen und sie immer weniger zu prüfen; die Forschung nennt es Automatisierungs-Selbstgefälligkeit, und sie tritt bei Laien wie bei Fachleuten auf und lässt sich durch Training nicht einfach abstellen.[10] Genau das ist die Gefahr im Umgang mit einer KI, die in neun von zehn Fällen Brauchbares liefert: Man gewöhnt sich das Hinsehen ab. Und dann kommt der zehnte Fall, die plausible, flüssige, falsche Antwort an der zerklüfteten Grenze aus Kapitel 3, und niemand bemerkt sie.

Stolperfalle — Das ausgehöhlte Urteil Die verführerischste Fehlentscheidung des Übergangs lautet: „Die Routine übernimmt die Maschine, das Urteil behalte ich." Das klingt nach kluger Arbeitsteilung und höhlt doch genau die Fähigkeit aus, die man behalten wollte. Urteilskraft ist kein Besitz, den man hat, sondern eine Übung, die man verliert, wenn man sie aussetzt. Wer nie mehr selbst einen Entwurf schreibt, einen Code liest, eine Rechnung nachvollzieht, verlernt mit der Zeit, gute von schlechter Arbeit zu unterscheiden — und merkt es nicht, weil die Maschine ihn in Sicherheit wiegt.

Daraus folgt nicht, die Technik zu meiden, wohl aber: am Handwerk zu bleiben. Eine Führungskraft, die beurteilen können will, was die Maschine produziert, muss nah genug an der Sache bleiben, um das geschulte Auge zu behalten; sie muss gelegentlich selbst tun, was sie sonst bewertet. Nur so trägt das Urteil. Dieselbe Frage stellt sich eine Ebene tiefer: Wenn die Maschine die Übungsarbeit der Jüngeren übernimmt, muss ein Haus bewusst Gelegenheiten schaffen, an denen der Nachwuchs sein Urteil noch erwerben kann — sonst fehlen in zehn Jahren die Beurteiler. (Die größere, das ganze Unternehmen betreffende Frage, was geschieht, wenn eine ganze Belegschaft sich von der Maschine abhängig macht, behandelt Kapitel 13.)

KI-Arbeit beurteilen: zwei Disziplinen

Wie sieht gutes Urteilen über KI-Arbeit konkret aus? Zwei Disziplinen tragen es.

Die erste: Sie müssen sagen können, was „gut" ist. Eine Ausgabe lässt sich nur beurteilen, wenn man einen Maßstab hat, und diesen Maßstab zu formulieren, ist selbst eine Urteilsleistung. Wer einen Auftrag nicht präzise fassen kann, kann das Ergebnis nicht prüfen; er kann nur sagen, ob es ihm gefällt. Das geschulte „Noch mal" des Meisters setzt voraus, dass er weiß, wie das gute Stück auszusehen hat. Die Fähigkeit, dieses Ziel scharf zu benennen, ist im Zeitalter der Maschine kein Beiwerk, sondern Kernkompetenz.

Die zweite: das Abnicken vermeiden. Die größte Gefahr ist nicht, dass die Maschine Fehler macht — das tut sie —, sondern dass der Mensch sie durchwinkt. Wie real das ist, zeigt ein kontrolliertes Experiment der Universität Stanford: Versuchsteilnehmer, die mit einem KI-Assistenten programmierten, schrieben unsichereren Code als die Vergleichsgruppe ohne, und hielten ihn zugleich für *sicherer*.[11] Das ist die Automatisierungs-Selbstgefälligkeit in Reinform: schlechtere Arbeit bei größerem Vertrauen. Die anwaltlichen Schriftsätze mit den erfundenen Urteilen aus Kapitel 3 sind dieselbe Geschichte aus der Welt der Justiz. Der Schutz dagegen ist eine Haltung: an den Stellen, an denen es zählt, prüft man, statt zu vertrauen.

> **Profi-Tipp — Drei Fragen an jede KI-Ausgabe, bevor Sie sie übernehmen**
>
> 1. Habe ich vorher festgelegt, woran ich erkenne, dass diese Ausgabe gut ist — oder beurteile ich sie nach Gefühl?
> 2. Liegt diese Aufgabe in der verlässlichen Zone (begrenzt, prüfbar) oder jenseits der zerklüfteten Grenze?
> 3. Habe ich die eine Stelle geprüft, an der ein Fehler teuer würde — oder habe ich nur den flüssigen Ton für Richtigkeit genommen?

Dass dies kein akademisches Problem ist, zeigen die deutschen Unternehmen selbst. In einer repräsentativen Befragung des Digitalverbands Bitkom nennen sie die mangelhafte Ergebnisqualität und die fehlende Nachvollziehbarkeit der KI-Ausgaben unter den größten Hürden ihres Einsatzes, und ein gutes Drittel fürchtet schlicht falsche Ergebnisse.[12] Das ist, von der anderen Seite gelesen, die Nachfrage nach Urteil: Wo die Maschine an Qualität und Verlässlichkeit zweifeln lässt, wird der Mensch gebraucht, der das prüft.

Eine Tradition, die zum knappen Gut passt

Hier liegt eine Chance, die der deutschsprachige Raum nicht übersehen sollte. Wenn Urteilskraft zum knappsten Gut wird, dann besitzt eine Wirtschaft einen Vorteil, die das Beurteilen kultiviert hat wie kaum eine andere. Der Meisterbrief, die Handwerkskammern, die ganze Architektur, die das geschulte Auge über Generationen weitergibt — all das ist ein Reservoir genau jener Fähigkeit, die jetzt an Wert gewinnt. Übertragbar ist dabei nicht das geschulte Auge selbst — das bleibt an seine Domäne gebunden —, wohl aber die Kultur der Abnahme: dass Arbeit einen expliziten Maßstab hat und erst gilt, wenn sie geprüft ist. Die deutsche Berufsbildungsforschung hat dem impliziten Erfahrungswissen, der Könnerschaft, die in keiner Anleitung steht, eine eigene Disziplin gewidmet; Institute wie das Fraunhofer IFF arbeiten daran, das Erfahrungswissen ausscheidender Fachleute zu bewahren, bevor es mit ihnen in den Ruhestand geht.[13]

Doch Vorsicht vor der Verklärung. Erstens ist das Meisterprinzip umstritten: Die Monopolkommission, der wettbewerbspolitische Berater der Bundesregierung, sieht in der Meisterpflicht eine Marktzutrittsschranke und hat 2020 vor ihrer teilweisen Wiedereinführung für zwölf Gewerke gewarnt.[14] Zweitens, und wichtiger: Das geschulte Auge nützt nur, wenn es sich auf das neue Medium richtet. Die Urteilskraft, die jahrzehntelang Schubladen und Schweißnähte prüfte, muss nun lernen, auch die Ausgaben einer Maschine zu prüfen: Texte, Analysen, Code. Die Tradition ist ein Startkapital, kein Ruhekissen. Sie verschafft einen Vorsprung nur dem, der sie in die neue Aufgabe überträgt.

Was dieses Kapitel für Sie ändert

Sie haben die erste der drei Knappheiten kennengelernt, und sie ist die persönlichste. Das Urteil — gute von schlechter Arbeit zu unterscheiden und zu entscheiden, was zählt — wird umso wertvoller, je billiger das Erzeugen wird. Es lässt sich nicht kaufen, nur erwerben, und es erodiert, wenn man das Tun ganz aus der

Hand gibt. Ihre Rolle verschiebt sich vom Verwalter zum Beurteiler, und diese Rolle ist anspruchsvoller, nicht bequemer. Wer sie ausfüllen will, bleibt nah genug am Handwerk, um das geschulte Auge zu behalten, sagt klar, was gut ist, und gewöhnt sich das Abnicken nicht an.

Das Urteil entscheidet, was gute Arbeit ist. Aber woran *arbeitet* die Maschine im eigenen Haus besser als bei jedem anderen? Das führt zur zweiten Knappheit: dem eigenen Kontext, den Daten und Rückkopplungen, die nur Ihnen gehören und die kein Wettbewerber nachkaufen kann. Ihr gehört das nächste Kapitel.

Das Wichtigste in 60 Sekunden

- Wird das Erzeugen billig, steigt der Wert des **Urteils** — des knappen Komplements. Wenn alle dieselben Modelle haben, wirkt die Technik angleichend; was unterscheidet, ist, was Menschen damit anfangen.
- Die Rolle der Führung verschiebt sich **vom Verwalter zum Beurteiler**: weg vom Zusammentragen, hin zum Werten und Entscheiden. Das ist die schwerere Übung, nicht die bequemere.
- **Das Paradox:** Wer das Tun ganz abgibt, höhlt das Urteil aus, das er behalten wollte. Urteilskraft ruht auf Übung; sie verkümmert ohne Nähe zur Sache. Deshalb: **am Handwerk bleiben.**
- In der Praxis heißt gutes Urteilen: **vorher sagen, was „gut" ist**, und **das Abnicken vermeiden** (mit KI-Assistent schrieben Versuchsteilnehmer unsichereren Code — und hielten ihn für sicherer).
- Der deutschsprachige Raum hat mit seiner **Handwerks- und Qualitätstradition** ein Reservoir genau dieser Fähigkeit — ein Startkapital, das nur nützt, wer es auf die KI-Arbeit überträgt.

Merksatz Die Maschine erzeugt die Arbeit; ob sie taugt, entscheidet ein Urteil, das sich nicht kaufen, nur erarbeiten lässt.

Drei Fragen an Ihr Unternehmen

1. An welchen Stellen verlassen sich Ihre Führungskräfte schon so auf KI-Ausgaben, dass sie das eigene Urteil nicht mehr üben — und woran würden Sie das merken?
2. Wer in Ihrem Haus kann für die wichtigen KI-gestützten Aufgaben präzise sagen, woran sich „gut genug" erkennen lässt — und wo fehlt dieser Maßstab?
3. Wie sorgen Sie dafür, dass Ihre besten Beurteiler nah genug am Handwerk bleiben, um ihr geschultes Auge nicht zu verlieren?

Quellen

1. Franz Reuleaux, *Briefe aus Philadelphia* (1876): Der Vorsitzende der deutschen Jury auf der Weltausstellung in Philadelphia urteilte über die deutsche Industrieware „billig und schlecht" und löste damit eine Qualitätsdebatte aus. Die Wendung stammt aus Reuleaux' Briefen, nicht aus dem britischen Gesetz.
2. Britischer Merchandise Marks Act 1887 (in Kraft 23. August 1887): verlangte die Herkunftskennzeichnung eingeführter Waren, ursprünglich, um deutsche Nachahmungen britischer Produkte als minderwertig zu kennzeichnen. Die als Schmähung gedachte Kennzeichnung kehrte sich später in ein Gütesiegel um. Quellen: German History Intersections (Leibniz/GHI Washington); Gesetzestext. (Den genauen Zeitpunkt der Umkehrung nicht auf ein einzelnes Jahr festlegen.)
3. Michael Polanyi, *The Tacit Dimension* (1966), S. 4: „we can know more than we can tell." Polanyi begründet den Begriff des impliziten (taziten) Wissens: Könnerschaft und Urteil sind durch Übung und Beispiel erworben und lassen sich nicht vollständig in Regeln fassen. Die spätere Kurzform „Polanyi's Paradox" (David Autor, 2014) bezeichnet die Folge für die Automatisierung.
4. Carl Shapiro & Hal R. Varian, *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business School Press, 1999): Die Nachfrage nach einem Gut steigt, wenn der Preis seiner Komplemente fällt; daraus die Strategie, Komplemente zu verbilligen. Die Anwendung „billige Erzeugung → knappes Urteil" ist eine eigene Schlussfolgerung aus diesem Prinzip, kein Zitat.
5. David Wingate, Barclay L. Burns & Jay B. Barney, *Why AI Will Not Provide Sustainable Competitive Advantage*, MIT Sloan Management Review, 8. Mai 2025: „Far from being a source of differentiation, artificial intelligence will be a source of homogenization"; wenn alle dieselbe Technik nutzten, werde das den Markt bewegen, aber niemandem einen Vorsprung verschaffen. Barney hat den ressourcenbasierten Ansatz (resource-based view) maßgeblich geprägt (begründet von Wernerfelt 1984, aufbauend auf Penrose). Begründetes Argument, kein empirischer Befund.
6. Ravikiran Kalluri, *What's Your Edge? Rethinking Expertise in the Age of AI*, MIT Sloan Management Review, 20. Oktober 2025: „Leaders aren't paid because they can access information; they're paid to make decisions when the stakes are real and the outcomes are uncertain." Reputabler Meinungsessay, als Kommentar zu lesen. Zur Urteilsqualität als realer, gestaltbarer Größe vgl. Kahneman, Sibony & Sunstein, *Noise* (2021).
7. Dell'Acqua et al., *Navigating the Jagged Technological Frontier*, HBS Working Paper 24-013 (2023; publiziert in *Organization Science* 2025): Feldexperiment mit 758 Unternehmensberatern; Unterscheidung der Arbeitsweisen „Centaur" (Aufgabe teilen, Urteil behalten) und „Cyborg" (enge Verzahnung). Innerhalb der Modellfähigkeiten deutliche Gewinne; bei einer Aufgabe jenseits der Fähigkeiten rund 19 Prozentpunkte häufiger falsche Ergebnisse (bereits in Kap. 3 eingeführt).

8. Lisanne Bainbridge, *Ironies of Automation*, Automatica 19(6), 1983, S. 775–779: Dem Bediener bleiben „the tasks which the designer cannot think how to automate"; „physical skills deteriorate when they are not used … a formerly experienced operator … may now be an inexperienced one"; Kern-Paradox: „by taking away the easy parts of his task, automation can make the difficult parts of the human operator's task more difficult." Konzeptionelle Arbeit aus der Prozess- und Luftfahrtautomatisierung, kein Experiment.
9. Daniel Kahneman & Gary Klein, *Conditions for Intuitive Expertise: A Failure to Disagree*, American Psychologist 64(6), 2009, S. 515–526: „Two conditions must be satisfied for skilled intuition to develop: an environment of sufficiently high validity and adequate opportunity to practice the skill." Gültige Experten-Intuition entsteht nur in hinreichend regelhaften Umgebungen mit genügend geübter Praxis.
10. Raja Parasuraman & Dietrich Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, Human Factors 52(3), 2010, S. 381–410: Unter Mehrfachbelastung sinkt die Überwachung einer gut funktionierenden Automatik, ihre Fehler werden übersehen; die Selbstgefälligkeit tritt bei Laien wie Experten auf und lässt sich durch Training oder bloße Übung nicht zuverlässig beseitigen. Befunde überwiegend aus Labor- und Simulatorstudien (vor dem LLM-Zeitalter); die Übertragung auf das Abnicken von KI-Ausgaben ist eine Analogie.
11. Neil Perry, Megha Srivastava, Deepak Kumar & Dan Boneh, *Do Users Write More Insecure Code with AI Assistants?*, ACM CCS 2023 (arXiv:2211.03622): In einem randomisierten Experiment (n = 47) schrieben Teilnehmer mit KI-Assistent unsichereren Code und hielten ihn zugleich eher für sicher. Kleine Stichprobe, Modellstand 2022; der belastbare Punkt ist das Verhalten (Über-Vertrauen), nicht die exakte Quote. Ergänzend Sandoval et al., *Lost at C*, USENIX Security 2023 (arXiv:2208.09727).
12. Bitkom, *Künstliche Intelligenz in Deutschland 2025* (repräsentative Unternehmensbefragung): mangelhafte Ergebnisqualität und fehlende Nachvollziehbarkeit der Ergebnisse zählen zu den meistgenannten Hürden des KI-Einsatzes; rund 38 % fürchten falsche Ergebnisse. *(Vor Druck: exakte Werte und Erhebungsjahr gegen den Bitkom-Studienbericht prüfen.)*
13. Zum impliziten Erfahrungswissen in der deutschen Berufsbildung: Georg Hans Neuweg u. a. (Hrsg.), *Implizites Wissen* (wbv, 2020), aufbauend auf Polanyi; Fraunhofer IFF, Programme zur Identifikation und Übertragung des Erfahrungswissens ausscheidender Fachkräfte (iff.fraunhofer.de). Zum institutionellen Rahmen (Meisterbrief, Handwerkskammern, duale Ausbildung): Zentralverband des Deutschen Handwerks (ZDH); Handwerksordnung von 1953.
14. Monopolkommission, Sondergutachten 31, *Reform der Handwerksordnung*: Die Meisterpflicht wirkt als Marktzutritts- und Berufszugangsschranke. Die Kommission sprach sich gegen die zum 14. Februar 2020 erfolgte Wiedereinführung der Meisterpflicht für zwölf zuvor (2004) freigegebene Gewerke aus. Der Hinweis dient dazu, die Handwerkstradition nicht zu verklären; das Argument für die Urteilskraft hängt nicht am Meisterzwang.

Kapitel 9 — Kontext: der Vorsprung, den niemand nachkaufen kann

Lernziele Nach diesem Kapitel können Sie

- begründen, warum Ihr eigener Kontext — proprietäre Daten und nachprüfbare Rückkopplungen — ein Vorsprung ist, den Sie besitzen und nicht bloß mieten, anders als das Modell selbst;
- die verbreitete Übertreibung, schon das Anhäufen von Daten verschaffe einen Vorsprung, einordnen und sagen, was Daten wirklich verteidigbar macht;
- das Daten-Schwungrad und die Rolle der Rückkopplung erklären;
- Souveränität — wo Ihre Daten und Ihre Rechenleistung liegen — als Vorstandsfrage behandeln, nicht als IT-Detail.

In den Kellern und Rechenzentren der deutschen Industrie liegt ein Vermögen, das kein Wettbewerber kaufen kann. Eine starke Fertigung erzeugt es von selbst: jahrzehntelange Aufzeichnungen aus Produktion, Wartung und Qualitätssicherung, von Temperaturen und Drücken über Taktzeiten bis zu Ausschussquoten und Wartungsprotokollen. Auf diesen Daten hat kein öffentliches KI-Modell je trainiert, und sie gehören jeweils nur einem Haus. Es ist genau die Art von Vermögen, das übrig bleibt, wenn das Modell selbst zur Ware geworden ist.

Und doch bleibt dieses Vermögen meist unberührt. Nach einer Erhebung des Branchenverbands Bitkom aus dem Jahr 2024 schöpfen nur sechs Prozent der deutschen Unternehmen das Potenzial ihrer Daten nach eigener Einschätzung voll aus.[8] Man hat die Daten, aber man hebt sie nicht. In dieser Lücke zwischen Haben und Heben liegt das Thema dieses Kapitels. Es handelt von der zweiten der drei Knappheiten, dem Kontext — und von der Ernüchterung, die dazugehört. Denn so wertvoll der eigene Kontext ist, so verbreitet ist der Irrtum, schon das Anhäufen von Daten verschaffe einen Vorsprung. Beides gehört auseinandergehalten.

Warum der Kontext ein Vermögenswert ist, den Sie besitzen

Erinnern wir uns an die Lage. Das Modell ist gemietet und für alle gleich; jeder Wettbewerber kann dasselbe einkaufen (Kapitel 7). Das Urteil aus dem vorigen Kapitel ist eine knappe *Fähigkeit*, die man sich erarbeitet. Der Kontext dagegen — die eigenen Daten und die nachprüfbaren Rückkopplungen aus dem eigenen Geschäft — ist ein knapper *Vermögenswert*: das eine der drei knappen Güter, das man im strengen Sinn *besitzt*. Ein Konkurrent kann das Modell kaufen; die Anlagendaten zweier Jahrzehnte kann er nicht kaufen. Freilich sitzt auch mancher Wettbewerber auf einem Archiv derselben Art. Gegen ihn entscheidet nicht der Besitz, den beide haben, sondern wer sein Archiv zuerst ordnet und in eine Rückkopplung bringt — davon handelt der Rest dieses Kapitels. Hier liegt die Bedeutung des Kapiteltitels. Der eigene Kontext ist einer von mehreren Wettbewerbsvorteilen; Marke, Vertrieb, Größe, Standort und Patente zählen ebenso. Unter ihnen ist er derjenige, den die KI am unmittelbarsten schafft und den ein gemietetes, für alle gleiches Modell niemandem geben kann: der Vorsprung, den niemand nachkaufen kann.

Diese Bedeutung wächst, und zwar aus einem Grund, der die ganze Branche umtreibt. Die öffentlich verfügbaren Texte, aus

denen die großen Modelle bisher lernten, gehen zur Neige. Das Forschungsinstitut Epoch AI schätzt den nutzbaren Bestand an öffentlichem, von Menschen geschriebenem Text auf eine Größenordnung, die die Modelle bei anhaltendem Tempo bis Anfang der 2030er-Jahre ausgeschöpft haben werden.[1] Je knapper das öffentliche Material, desto wertvoller das, was nicht öffentlich ist: die Daten, die nur in Ihrem Haus liegen. Ein bestimmtes Jahr ist damit nicht gesagt, wohl aber eine Richtung, und sie zeigt auf den proprietären Kontext.

> **Klartext — Die zweite Knappheit: der Kontext** *Kontext* meint hier die eigenen, vertrauenswürdigen Daten und die nachprüfbaren Rückkopplungen aus dem eigenen Geschäft: Maschinendaten, Kundenhistorien, Prüfprotokolle, die Rückmeldung, ob eine Empfehlung am Ende stimmte. Aus ihm wird die Maschine in Ihrem Geschäft besser, als sie es bei irgendjemand anderem sein könnte. Er ist die zweite der drei Knappheiten und die einzige unter ihnen, die ein Vermögenswert ist: Man kann ihn besitzen, und ein Wettbewerber kann ihn nicht nachkaufen.

Doch Vorsicht: Daten allein sind kein Vorsprung

Hier ist eine Korrektur nötig, denn um kaum eine Behauptung ranken sich so viele Übertreibungen. „Daten sind das neue Öl", „wer die Daten hat, gewinnt" — solche Sätze sind zu Allgemeinplätzen geworden, und sie sind, so pauschal, falsch. Aufschlussreich ist, dass ausgerechnet jene, die man als Apostel des Datenvorsprungs zitiert, am deutlichsten widersprechen. Die Risikokapitalgesellschaft Andreessen Horowitz, oft als Kronzeugin angeführt, nannte den Datenvorsprung schon 2019 ein „leeres Versprechen": Aus dem bloßen Mehr an Daten entstehe in der Regel kein selbstverstärkender Effekt.[2] Die Ökonomen Andrei Hagiu und Julian Wright fassten es im *Harvard Business Review* noch schärfer: In den meisten Fällen überschätzten Führungskräfte den Vorteil, den Daten verschaffen, gewaltig.[3]

Sie haben recht, und der Grund ist technischer Natur. Der Nutzen zusätzlicher Daten nimmt ab: Jede Verdopplung der Datenmenge bringt einen kleineren Zuwachs an Genauigkeit als die vorige, ein Zusammenhang, der sich über viele Größenordnungen als Potenzgesetz zeigt.[4] Hinzu kommt, dass moderne Modelle aus wenigen Beispielen lernen (Kapitel 4) und dass sorgfältig ausgewählte, saubere Daten oft mehr ausrichten als riesige, ungeordnete Mengen; in einer vielbeachteten Untersuchung genügten tausend gut kuratierte Beispiele, um ein Modell zuverlässig auf eine Aufgabe einzustellen.[5] Wer also glaubt, ein großer, ungeordneter Datenbestand — ein „Data Lake", wie die Fachwelt ihn nennt — verschaffe ihm schon einen Vorsprung, irrt.

Was macht Daten dann wirklich verteidigbar? Hagiu und Wright geben die brauchbarste Antwort, und es ist bezeichnend, dass sie von den Skeptikern kommt. Ein Vorsprung entsteht nur, wenn mehrere Bedingungen zusammenkommen: Die Daten müssen einzigartig und nicht anderswo zu beschaffen sein; sie müssen für eine wertvolle Aufgabe relevant sein; ihr Wert darf nicht zu schnell veralten; und die Verbesserungen, die man aus ihnen zieht, müssen für andere schwer zu imitieren sein.[3] Es geht um diese Eigenschaften, nicht um die Menge. Die Anlagendaten eines Fertigers aus zwei Jahrzehnten erfüllen sie womöglich alle; ein gekaufter Adressdatensatz erfüllt keine.

> **Stolperfalle — Die Data-Lake-Illusion** Der teuerste Irrtum in diesem Feld ist der Glaube, das Sammeln und Speichern großer Datenmengen sei für sich genommen ein Vorsprung. Ein Becken voller Rohdaten, die niemand ordnet, prüft und in eine Rückkopplung einspeist, ist kein Vorsprung, sondern ein Kostenposten. Verteidigbar wird Kontext erst durch vier Eigenschaften: einzigartig, aufgabenrelevant, ständig erneuert, schwer kopierbar. Fragen Sie bei jedem „Daten-Asset", das man Ihnen vorrechnet, welche dieser vier Eigenschaften es wirklich erfüllt.

Das Schwungrad und die Rückkopplung

Wie aus Daten ein Vorsprung wird, der sich von selbst vergrößert, beschreibt ein Bild, das der Managementautor Jim Collins geprägt hat: das Schwungrad.[6] Ein schweres Rad lässt sich nur mühsam in Bewegung setzen, doch mit jeder Umdrehung läuft es leichter, bis es aus eigener Kraft trägt. Auf Daten übertragen: Ein Produkt gewinnt Nutzer; deren Nutzung erzeugt Daten und Rückmeldungen; daraus wird das Produkt besser; das bessere Produkt gewinnt weitere Nutzer. Wer dieses Rad früh in Schwung bringt, zieht davon.

Entscheidend ist dabei, was das Rad eigentlich antreibt. Der Treibstoff ist selten das rohe Datenvolumen. Es ist die *Rückkopplung*: die Information, ob der Vorschlag angenommen, korrigiert oder verworfen wurde, ob die Vorhersage am Ende eintraf, ob die Reparatur hielt. Diese nachprüfbaren Rückmeldungen sind das eigentlich Knappe, denn sie sagen dem System, was *richtig* war, und sie fallen nur dort an, wo ein Geschäft wirklich läuft. Hierin liegt der zweite Grund, warum der Kontext eines arbeitenden Unternehmens so schwer nachzubauen ist: Man kann Daten kopieren, aber nicht die gelebte Rückkopplungsschleife eines fremden Geschäfts.

Zwei Einschränkungen gehören dazu, sonst wird aus dem Bild ein Werbeversprechen. Erstens garantiert ein Schwungrad kein Ziel. Der Automobilhersteller Tesla etwa sammelt seit Jahren Fahrdaten aus einer riesigen Flotte und hat damit ein beeindruckendes Schwungrad gebaut; dennoch kam das unbeaufsichtigte autonome Fahren um Jahre langsamer voran, als das Unternehmen Jahr für Jahr versprach.[7] Ein Datenvorsprung beschleunigt; ein Ergebnis garantiert er nicht. Zweitens kann sich ein Schwungrad auch in die falsche Richtung drehen: Wer ein Modell auf den eigenen, verzerrten Rückmeldungen trainiert, verstärkt womöglich seine Fehler. Und dem, was man mit Rückkopplungsdaten tun darf, setzt die Datenschutz-Grundverordnung enge Grenzen, auf die Kapitel 12 zurückkommt.

Zahlen, die zählen

- Öffentlicher Trainingstext-Bestand: rund **300 Billionen Tokens**, voraussichtlich **bis Anfang der 2030er-Jahre ausgeschöpft** (Schätzung mit großer Unsicherheit).[1]
- Nur **6 %** der deutschen Unternehmen schöpfen ihr Datenpotenzial nach eigener Einschätzung voll aus; nur **17 %** geben Daten an andere weiter.[8]
- US-Anbieter halten rund **70 %** des europäischen Cloud-Marktes; europäische Anbieter rund **15 %** (Stand Mitte 2025).[9]

Souveränität: wo Ihre Daten und Ihre Rechenleistung liegen

Wenn der Kontext der wertvollste Vermögenswert ist, dann wird eine Frage zur Vorstandsfrage, die viele noch für ein technisches Detail halten: Wo liegen diese Daten physisch, und unter wessen Hoheit? In Kapitel 7 ging es um die Wahl zwischen offenen und geschlossenen Modellen; hier geht es um die Ebene darunter, die Infrastruktur, auf der alles läuft.

Die Ausgangslage ist eine Abhängigkeit. Rund siebzig Prozent des europäischen Cloud-Marktes entfallen auf drei amerikanische Anbieter; die europäischen zusammen kommen auf etwa fünfzehn Prozent, mit fallender Tendenz über die Jahre.[9] Das wäre eine bloße Marktbeobachtung, hätte es nicht eine rechtliche Kehrseite. Der amerikanische CLOUD Act von 2018 verpflichtet Anbieter mit Sitz in den USA, Daten herauszugeben, über die sie Verfügungsgewalt haben, gleich, wo auf der Welt diese Daten gespeichert sind.[10] Das heißt im Klartext: Ein Rechenzentrum in Frankfurt macht Ihre Daten nicht souverän, solange der Betreiber amerikanischem Recht untersteht. Souveränität nach dem Standort des Servers ist nicht dasselbe wie Souveränität nach der Hoheit über ihn. Die rechtliche Grundlage für den transatlantischen Datenverkehr wiederum, der Angemes-

senheitsbeschluss von 2023, ist zwar in Kraft, aber angefochten; die Berufung liegt beim Europäischen Gerichtshof — und die beiden Vorgängerfundamente sind dort bereits weggebrochen.[11]

Daraus erwächst die Frage, die in jeden Vorstand gehört: Wo liegen unsere Daten und unsere Rechenleistung, unter welcher Hoheit, und was geschähe, wenn dieser Zugang eines Tages verteuert, beschränkt oder politisch gekappt würde? Der Branchenverband Bitkom hat gemessen, wie verwundbar man sich fühlt: 81 Prozent der deutschen Unternehmen sehen sich von US-Digitaltechnik abhängig, und über die Hälfte traute sich nicht zu, einen Importstopp auch nur ein Jahr zu überstehen.[12]

Europa versucht zu antworten. Die Initiative GAIA-X soll Regeln und Standards für eine vertrauenswürdige, europäische Dateninfrastruktur schaffen; daneben sind „souveräne" Cloud-Angebote entstanden. Man sollte beides nüchtern betrachten. GAIA-X gilt vielen als langsam und bürokratisch, ein Projekt, das mehr Papier als Wirkung erzeugt hat, zumal es ausgerechnet jene amerikanischen Konzerne als Mitglieder aufnahm, gegen deren Übermacht es gedacht war.[13] Und „souverän" ist nicht gleich souverän: Manche dieser Angebote laufen ihrerseits auf amerikanischer Technik und mildern die Hoheitsfrage nur, während andere, in deutscher oder europäischer Hand, sie wirklich auflösen.[14] Die ehrliche Vorstandsfrage lautet deshalb nicht „Steht der Server in Europa?", sondern „Wer hat die rechtliche Verfügungsgewalt?"

Auch hier gilt es, das Maß zu wahren. Vollständige Abschottung hat ihren Preis an Fähigkeit, Kosten und Anschluss, und die Grenze zwischen berechtigter Souveränität und bloßem Protektionismus ist fließend.[15] Souveränität ist eine Risikoabwägung, kein Selbstzweck: Für die einen Daten und Anwendungen ist die Hoheit entscheidend, für die anderen zählt die Leistungsfähigkeit mehr. Die Aufgabe des Vorstands ist, diese Abwägung bewusst zu treffen, statt sie dem Zufall der bequemsten Lösung zu überlassen.

Der deutsche Sonderfall: ein Schatz, der kaum gehoben wird

Damit zurück zur deutschen Industrie, mit der dieses Kapitel begann, denn hier liegen für den deutschsprachigen Raum die eigentliche Chance und die eigentliche Mahnung in einem. Auf dem Datenschatz, um den andere sie beneiden, sitzt sie längst. Forschungseinrichtungen wie die Plattform Industrie 4.0 und die Akademie acatech betonen seit Jahren, welches Geschäftspotenzial in diesen industriellen Daten steckt.[16]

Doch zwischen Haben und Heben klafft eine Lücke. Die sechs Prozent vom Anfang dieses Kapitels sind nur die Spitze: Nach derselben Bitkom-Erhebung nutzt gut die Hälfte der Unternehmen ihr Datenpotenzial nach eigenem Bekunden eher wenig oder gar nicht, und nur 17 Prozent geben Daten überhaupt an andere weiter, gebremst von Datenschutzsorgen, Rechtsunsicherheit und der Furcht vor Missbrauch.[8] Der Schatz ist da; gehoben hat ihn kaum jemand. Genau deshalb sind kollektive Anläufe entstanden: Datenräume wie Catena-X in der Automobilindustrie oder das staatlich mit rund 150 Millionen Euro geförderte Manufacturing-X für die Industrie insgesamt, die das Teilen von Daten geordnet und souverän ermöglichen sollen.[17]

> **Aus der Praxis — Wie ein Maschinenbauer seinen Datenschatz hob**
> Der Heidelberger Druckmaschinen-Konzern hat aus seiner vernetzten Maschinenflotte ein datengetriebenes Geschäft gemacht: Die laufenden Betriebsdaten der angeschlossenen Druckereien fließen in eine Cloud, die den Betrieben Vergleichswerte und KI-gestützte Optimierungsvorschläge zurückspielt — zeitweise bis hin zu Verträgen, die nicht die Maschine, sondern den erzielten Ausstoß abrechneten.[18] Das ist das Schwungrad in Reinform: Je mehr Maschinen angeschlossen sind, desto besser die Vergleichsbasis, desto wertvoller der Dienst. Der Rohstoff dafür — die Flottendaten — lag vorher schon vor. Den Unterschied machte die Entscheidung, ihn zu heben.

Die Lehre ist unbequem und ermutigend zugleich. Der Vorsprung liegt nicht im Anschaffen einer KI, sondern im Heben eines Schatzes, den viele längst besitzen. Das ist Arbeit: Daten ordnen, verknüpfen, ihre Qualität sichern, eine Rückkopplung aufbauen — organisatorische Arbeit, die an das Organisationskapital aus Kapitel 2 und an das Urteil aus Kapitel 8 anschließt. Die technischen Kanäle dafür kennen Sie aus Kapitel 4: das Nachschlagen, das der Maschine die eigenen Unterlagen vorlegt, und, seltener, die Feinabstimmung. Der Rohstoff ist im deutschsprachigen Raum reichlich vorhanden; es hapert am Heben.

Was dieses Kapitel für Sie ändert

Sie haben die zweite Knappheit kennengelernt, und sie ist die einzige der drei, die ein Vermögenswert ist. Das Modell mieten alle; Ihren Kontext besitzen nur Sie. Aber zum Vorsprung wird er erst, wenn die Daten einzigartig, aufgabenrelevant, erneuert und schwer kopierbar sind, wenn eine Rückkopplungsschleife sie in Schwung hält, und wenn Sie wissen und entscheiden, wo sie liegen und unter wessen Hoheit. Das bloße Anhäufen von Daten ist nichts davon. Für den deutschsprachigen Raum ist die Botschaft doppelt: Der Rohstoff ist da, in den Kellern der Industrie; die Aufgabe ist, ihn zu heben und seine Hoheit zu wahren.

Damit bleibt die dritte und letzte Knappheit. Sie haben das Urteil, gute Arbeit zu erkennen, und den Kontext, an dem die Maschine besser wird als bei jedem anderen. Es bleibt die Frage, wie viel das alles kostet und was es einbringt — die Ökonomie der Rechenleistung, der Wert, den Sie je Recheneinheit herausholen. Ihr gehört das nächste Kapitel, und mit ihm die Kennzahl, die diesem Buch ihren Namen leiht.

Das Wichtigste in 60 Sekunden

- Der **Kontext** — eigene Daten und nachprüfbare Rückkopplungen — ist die einzige der drei Knappheiten, die ein **Vermögenswert** ist: Das Modell mieten alle, den Kontext besitzen nur Sie. Mit dem Versiegen öffentlicher Trainingsdaten gewinnt er an Wert.
- **Daten allein sind kein Vorsprung.** Volumen nützt wenig (abnehmender Grenznutzen). Verteidigbar wird Kontext nur, wenn er einzigartig, aufgabenrelevant, erneuert und schwer kopierbar ist.
- Das **Schwungrad** treibt nicht das rohe Datenvolumen an, sondern die **Rückkopplung** — die Information, was richtig war. Sie lässt sich nicht aus einem fremden Geschäft kopieren.
- **Souveränität ist Vorstandssache:** Ein Server in Frankfurt ist nicht souverän, wenn der Betreiber US-Recht untersteht (CLOUD Act). Die Frage lautet: Wer hat die Verfügungsgewalt?
- **Der deutsche Sonderfall:** Die Industrie besitzt einen außergewöhnlichen Datenschatz — und hebt ihn kaum (nur 6 % schöpfen ihr Potenzial voll aus). Der Rohstoff ist da; es hapert am Heben.

Merksatz Das Modell ist gemietet, Ihre Daten sind Ihr Eigen — doch zum Vorsprung werden sie erst, wenn Sie den Schatz heben.

Drei Fragen an Ihr Unternehmen

1. Welche Ihrer Daten sind wirklich einzigartig, aufgabenrelevant und schwer kopierbar — und wo verwechseln Sie schiere Menge mit Vorsprung?
2. Wo schließt sich in Ihrem Geschäft eine Rückkopplungsschleife, wo erfahren Sie nachprüfbar, ob eine KI-Ausgabe richtig war, und wo versickert diese wertvollste Information ungenutzt?
3. Wissen Sie für Ihre wichtigsten Daten, unter wessen rechtlicher Hoheit sie liegen, und haben Sie diese Wahl bewusst getroffen?

Quellen

1. Villalobos, Ho, Sevilla, Besiroglu, Heim, Hobbhahn (Epoch AI), *Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data*, ICML 2024 (arXiv:2211.04325): nutzbarer Bestand öffentlichen, von Menschen erzeugten Texts in der Größenordnung von rund 300 Billionen Tokens (breites Konfidenzintervall); voraussichtliche Ausschöpfung zwischen 2026 und 2032. Eine Projektion mit großer Unsicherheit; die Schlussfolgerung, proprietäre Daten gewönnen dadurch an relativem Wert, ist eine begründete Ableitung, kein gemessener Befund.
2. Martin Casado & Peter Lauten, *The Empty Promise of Data Moats*, Andreessen Horowitz, 9. Mai 2019: aus dem bloßen Mehr an Daten entstehe in der Regel kein inhärenter Netzwerkeffekt; der Grenznutzen zusätzlicher Daten falle, die Beschaffungskosten stiegen. (Aus dem Jahr 2019, vor der jüngeren, datenfreundlicheren Lesart desselben Hauses.)
3. Andrei Hagiu & Julian Wright, *When Data Creates Competitive Advantage*, Harvard Business Review, Januar–Februar 2020: „In most instances people grossly overestimate the advantage that data confers." Die Autoren nennen die Bedingungen, unter denen Daten doch einen dauerhaften Vorteil schaffen (u. a. einzigartig/nicht anderswo beschaffbar, aufgabenrelevant, nicht zu schnell veraltend, schwer imitierbar).
4. Hestness et al., *Deep Learning Scaling Is Predictable, Empirically*, arXiv:1712.00409 (2017): Der Generalisierungsfehler sinkt mit wachsender Trainingsdatenmenge nach einem Potenzgesetz, d. h. jede Verdopplung bringt einen unterproportionalen Zuwachs. Das belegt „Volumen allein ist kein Vorsprung", nicht „Daten sind wertlos".
5. Zhou et al. (Meta), *LIMA: Less Is More for Alignment*, NeurIPS 2023 (arXiv:2305.11206): rund 1.000 sorgfältig kuratierte Beispiele genügten, um ein großes Modell zuverlässig auf Instruktionsbefolgung einzustellen. Betrifft die stilistische Anpassung, nicht das Einspeisen neuen Domänen-Faktenwissens.
6. Jim Collins, *Good to Great* (2001), Kap. 8: das Schwungrad als Bild für eine sich selbst verstärkende, kumulative Dynamik. Es ist eine Metapher, kein Mechanismus; auf Daten übertragen beschreibt sie die Form des Vorsprungs, nicht seinen Beweis.
7. Zur Tesla-Flottendatenstrategie („shadow mode"): Mark Harris, *Tesla's Autopilot Depends on a Deluge of Data*, IEEE Spectrum, 4. August 2022. Trotz eines erheblichen Datenvorsprungs blieb das unbeaufsichtigte autonome Fahren um Jahre hinter den wiederholten Ankündigungen zurück — ein Schwungrad beschleunigt, garantiert aber kein Ergebnis. (*Vor Druck: aktuellen Stand des Robotaxi-Betriebs prüfen; Formulierung robust halten.*)
8. Bitkom Research, Unternehmensbefragung 2024 (603 Unternehmen ab 20 Beschäftigten): nur 6 % schöpfen das Potenzial ihrer Daten nach eigener Einschätzung voll aus (31 % „eher stark", 42 % „eher wenig", 18 % „gar nicht"); 17

% geben Daten an andere weiter (36 % erhalten welche); meistgenannte Hürden des Datenteilens: Datenschutz (58 %), Rechtsunsicherheit (44 %), Furcht vor Missbrauch (41 %). Bitkom-Pressemitteilungen „Deutsche Unternehmen nutzen ihre Daten kaum" und „Unternehmen wollen Daten nutzen, aber nicht teilen" (2024).

9. Synergy Research Group, Juli 2025: US-Hyperscaler (AWS, Microsoft, Google) halten rund 70 % des europäischen Marktes für Cloud-Infrastrukturdienste; der Anteil europäischer Anbieter sank von rund 29 % (2017) auf etwa 15 % (seit 2022). Marktschätzung eines Branchendienstes, quartalsweise beweglich; die europäische Zahl nicht mit dem globalen Markt verwechseln.
10. Clarifying Lawful Overseas Use of Data (CLOUD) Act, USA, 2018: verpflichtet dem US-Recht unterstehende Anbieter, Daten in ihrer Verfügungsgewalt auf rechtmäßige Anordnung herauszugeben, unabhängig vom physischen Speicherort. Quelle: Congressional Research Service R45173. Der Act gewährt keinen pauschalen Zugriff und ist anfechtbar; entscheidend ist, dass der Speicherort allein keine Hoheit begründet, wenn der Betreiber US-Recht untersteht.
11. EU-US Data Privacy Framework: Angemessenheitsbeschluss der EU-Kommission vom 10. Juli 2023; das Gericht der Europäischen Union wies die Nichtigkeitsklage (Latombe) am 3. September 2025 ab, eine Berufung beim Europäischen Gerichtshof ist seit Ende Oktober 2025 anhängig. Die beiden Vorgängerrahmen (Safe Harbor, Privacy Shield) wurden vom EuGH gekippt. Status zeitkritisch — vor Drucklegung prüfen; derzeit gültig, aber angefochten.
12. Bitkom, *Digitale Souveränität 2025* (Befragung von rund 600 Unternehmen, Februar 2025): 81 % der deutschen Unternehmen sehen sich von Digitaltechnik aus den USA abhängig; 96 % importieren digitale Technik; rund 57 % könnten einen Importstopp nicht länger als ein Jahr überstehen. (Bereits in Kapitel 7 angeführt.)
13. Zur Kritik an GAIA-X (langsam, bürokratisch, Aufnahme der US-Hyperscaler als Mitglieder): u. a. The Register, „Gaia-X project doesn't have a future, claims Nextcloud boss" (Januar 2024); Forrester, „Gaia-X Is Back, But Only Adoption Will Signal Success". GAIA-X ist kein Cloud-Anbieter, sondern eine Initiative für Standards und Regeln einer europäischen Dateninfrastruktur; ihre praktische Wirkung ist umstritten.
14. „Souveräne" Cloud-Angebote unterscheiden sich erheblich in ihrer rechtlichen Unabhängigkeit: Manche (etwa die Delos Cloud) laufen auf amerikanischer Technik unter deutschem Betrieb und mildern die Hoheitsfrage; andere (etwa STACKIT oder Ionos) stehen in europäischer Hand ohne US-Bindung. SAP kündigte Investitionen von über 20 Mrd. € in souveräne Cloud-Angebote an (eine Absichtserklärung, kein verbuchter Betrag). Quellen: heise online; SAP-Mitteilungen 2025.
15. Zur Abwägung und zur Gefahr, dass Souveränität in Protektionismus kippt: Center for European Policy Analysis (CEPA), „A Tech Recipe for Europe: Digital Pragmatism"; EU Institute for Security Studies (EUISS), „Technical is po-

litical: when a cloud certification scheme divides Europe". Vollständige Datenlokalisierung hat Kosten an Fähigkeit, Preis und Anschluss.

16. Zum Geschäftspotenzial industrieller Daten: acatech / Forschungsbeirat Industrie 4.0, Impulspapier zu datengetriebenen Geschäftsmodellen im Maschinen- und Anlagenbau („Vom Datenpotenzial zum Geschäftserfolg"); Plattform Industrie 4.0 (BMWK). Diese Einrichtungen rahmen den Datenschatz ausdrücklich als *Potenzial*, nicht als bereits realisierten Vorteil.
17. Catena-X: föderierter Datenraum der Automobilindustrie (u. a. BMW, Mercedes-Benz, SAP, Bosch, ZF), aufgebaut auf GAIA-X-Prinzipien, BMWK-gefördert. Manufacturing-X: Initiative des BMWK, dieses Modell auf die gesamte Industrie auszuweiten, gefördert mit rund 150 Mio. € (Projektstart 2024). Quellen: BMWK-Förderkonzept; Plattform Industrie 4.0; BMW Group.
18. Heidelberger Druckmaschinen AG: datengetriebenes Subscription-Modell mit der Prinect-Cloud und „Print Shop Analytics" (KI-gestütztes Benchmarking aus der vernetzten Maschinenflotte). Die Existenz des Modells, der Cloud und des Analytics-Dienstes ist belegt (heidelberg.com); konkrete Leistungsangaben sind Herstellerangaben und nicht unabhängig geprüft. (*Vor Druck: aktuellen Umfang des Subscription-/Outcome-Modells auf heidelberg.com prüfen; das Unternehmen hat es seit ca. 2020 zurückgefahren.*)

Kapitel 10 — Wert je Recheneinheit (Return on Tokens)

Lernziele Nach diesem Kapitel können Sie

- *Return on Tokens* — den Wert, den Sie je Recheneinheit herausholen — als Kennzahl erklären und sagen, warum sie und nicht der Stückpreis über den Erfolg entscheidet;
- begründen, warum fallende Token-Preise Ihre KI-Rechnung nicht senken, sondern oft erhöhen (das Jevons-Paradox);
- die wichtigsten Hebel im Spannungsfeld von Kosten, Tempo und Qualität benennen — das richtige Modell, Zwischenspeichern, Stapeln;
- erkennen, warum jede Eigenbau-Rechnung ein Verfallsdatum hat, und eine einfache Heuristik für das eigene Budget anwenden.

Der Preis der Maschinenintelligenz fällt rasant. Eine Abfrage auf jenem Leistungsniveau, das Ende 2022 als Stand der Technik galt, kostete im November jenes Jahres rund zwanzig Dollar je Million Tokens. In nicht einmal zwei Jahren war derselbe Leistungsstand für sieben Cent zu haben, ein Rückgang auf weniger als ein Zweihundertachtzigstel.[1] Branchenanalysten und Anbieter beziffern den Verfall grob auf das Zehnfache pro Jahr.[2] Was 2023 ein Vermögen kostete, kostet heute Kleingeld.

Die naheliegende Folgerung lautet: Dann löst sich das Kostenproblem von selbst, man muss nur warten. Sie ist falsch. Während der Preis je Einheit fällt, steigen die KI-Rechnungen vieler Unternehmen, und der Geschäftswert bleibt aus. Die Lücke haben wir in Kapitel 2 gesehen: KI ist fast überall im Einsatz und zahlt sich fast nirgends spürbar aus; nur ein kleiner einstelliger Anteil der Unternehmen zieht daraus einen spürbaren Ergebnisbeitrag.[3] Das knappe Gut in dieser dritten und letzten Knappheit ist also nicht die Rechenleistung. Die ist im Überfluss vorhanden und wird täglich billiger. Knapp ist die Disziplin, aus ihr messbaren Wert zu ziehen. Diese Disziplin trägt in diesem Buch einen Namen, und ihr gehört das letzte Kapitel von Teil III.

Was Return on Tokens misst

Erinnern wir uns an zwei Begriffe aus Teil II. Der Token aus Kapitel 3 ist die kleinste Einheit, in der ein Modell liest, schreibt und abgerechnet wird. Die Inferenz aus Kapitel 4 ist das Ausführen des fertigen Modells auf eine Anfrage, und sie ist, anders als das einmalige Training, Ihre laufende, mit der Nutzung wachsende Rechnung. Tokens sind die Einheit, Inferenz ist der Posten. *Return on Tokens* setzt beides ins Verhältnis: den Geschäftswert, den eine KI-Anwendung schafft, geteilt durch die Recheneinheiten, die sie dafür verbraucht.

> **Klartext — Return on Tokens (Wert je Recheneinheit)** *Return on Tokens* ist der Geschäftswert, den Sie je verbrauchter Recheneinheit herausholen — der erzielte Nutzen geteilt durch die dafür aufgewendeten Tokens. In der Praxis messen Sie den Nenner über Ihre Inferenzrechnung: verbrauchte Tokens mal Preis. Die Kennzahl lenkt den Blick weg vom Stückpreis eines Tokens hin zu dem, was am Ende dabei herauskommt: Wert je Vorgang, nicht Kosten je Anfrage. Sie ist die dritte der drei Knappheiten und die ökonomische Disziplin, die aus billiger Rechenleistung erst Rendite macht.

Ein Bruch hat einen Zähler und einen Nenner, und die meiste Aufmerksamkeit richtet sich auf den falschen. Wer KI-Kosten senken will, feilscht am Nenner: an den Token-Kosten. Doch die fallen ohnehin von allein, Jahr für Jahr. Der Hebel, der einen Unterschied macht, liegt fast immer im Zähler: im Wert. Und der Wert entsteht nicht in der Rechenleistung, sondern in den beiden anderen Knappheiten: im Urteil aus Kapitel 8, das erkennt, welches Problem überhaupt zu lösen sich lohnt, und im Kontext aus Kapitel 9, an dem die Maschine im eigenen Geschäft besser wird als bei jedem anderen. Return on Tokens ist die Kennzahl, die beide zusammenbindet und in eine Zahl zwingt, für die jemand geradestehen muss.

Warum die Rechnung steigt, während der Preis fällt

Dass billigere Stückpreise nicht zu kleineren Rechnungen führen, ist ein alter ökonomischer Befund, kein Sondereffekt der KI. Der englische Ökonom William Stanley Jevons beschrieb ihn 1865 am Beispiel der Kohle: Je effizienter die Dampfmaschinen wurden, desto mehr Kohle verbrauchte England, nicht weniger. „Es ist ein Denkfehler anzunehmen, die sparsame Verwendung eines Brennstoffs bedeute geringeren Verbrauch", schrieb er. „Das Gegenteil ist wahr."[4] Was billiger wird, wird mehr genutzt, und der Mehrverbrauch übertrifft die Ersparnis.

Für die KI hat diesen Zusammenhang im Januar 2025, als ein günstiges Modell aus China die Märkte erschütterte, ausgerechnet der Vorstandschef von Microsoft in Erinnerung gerufen: Das Jevons-Paradox schlage wieder zu; je effizienter und zugänglicher KI werde, desto mehr werde sie genutzt, bis sie zur Ware werde, von der man nicht genug bekommen könne.[5] Man muss die strategische Absicht hinter solchen Worten nicht teilen, um den Mechanismus zu sehen. Genau er ist der Grund, warum man sich nicht zu einem guten Return on Tokens hinunterwarten kann. Die Preise fallen, die Nutzung wächst schneller, und am Ende des Quartals steht eine größere Rechnung, obwohl jeder einzelne Token billiger war. Wer keine Kennzahl auf den Wert

führt, merkt nicht einmal, ob sich der Mehrverbrauch gelohnt hat.

Das Spannungsfeld: Kosten, Tempo, Qualität

Bevor es um den Wert geht, ein klarer Blick auf die Kosten, denn an ihnen lässt sich mehr drehen, als die meisten glauben. Drei Größen ziehen gegeneinander: Kosten, Tempo und Qualität. Das stärkste Modell ist meist das teuerste und oft das langsamste; das billige Modell ist schnell, aber schwächer. Der teure Reflex besteht darin, für jede Aufgabe das stärkste Modell zu bemühen — als ließe man jeden Brief vom Vorstandsvorsitzenden austragen. Die Disziplin besteht darin, jede Aufgabe dem billigsten Modell zuzuweisen, das sie noch besteht. Was „noch besteht" heißt, beantwortet nicht das Bauchgefühl, sondern die Evaluation aus Kapitel 6: die eigene Beispielsammlung, an der sich prüfen lässt, ob das günstigere Modell für diese eine Aufgabe ausreicht.

Drei Hebel senken die Kosten, ohne den Wert anzutasten. Erstens das richtige Modell: Eine gestaffelte Anordnung, die einfache Fälle vom kleinen Modell erledigen lässt und nur die schweren an das große weiterreicht, kann die Qualität des großen Modells zu einem Bruchteil der Kosten erreichen; eine vielzitierte Untersuchung zeigte Einsparungen von bis zu achtundneunzig Prozent.[6] Zweitens das Zwischenspeichern: Wiederkehrende Teile einer Anfrage, etwa eine lange, immer gleiche Anweisung, lassen sich vorhalten, statt sie jedes Mal neu zu bezahlen, was die Kosten für diesen Teil um bis zu neunzig Prozent senkt. Drittens das Stapeln: Wer auf eine Antwort nicht sofort warten muss, reicht die Aufgaben gebündelt ein und zahlt dafür bei den großen Anbietern etwa die Hälfte.[7] Keiner dieser Hebel verlangt ein eigenes Modell. Sie verlangen, dass jemand die Rechnung versteht. Und groß wird die Rechnung dort, wo die Nutzung sich vervielfacht: bei Agenten, die je Auftrag ein Vielfaches der Tokens eines einfachen Dialogs verbrauchen (Kapitel 5), bei langen Kontexten, bei Millionen von Vorgängen. Genau dort zahlen sich die drei Hebel aus.

> **Profi-Tipp — Das billigste Modell, das die Prüfung besteht** Beginnen Sie jede Anwendung mit dem kleinsten, günstigsten Modell und prüfen Sie an Ihrer eigenen Beispielsammlung (Kapitel 6), ob es ausreicht. Steigen Sie erst zur nächsten Stufe auf, wenn die Prüfung das verlangt, nicht, weil das größere Modell beeindruckender klingt. Dieselbe Reihenfolge gilt für die Anpassung: erst der Prompt, dann das Nachschlagen, zuletzt die Feinabstimmung (Kapitel 4). Die teuerste Stufe zuerst zu wählen ist der häufigste Weg, Geld ohne Gegenwert auszugeben.

Das Verfallsdatum jeder Eigenbau-Rechnung

Aus den fallenden Preisen folgt eine unbequeme Wahrheit für jede größere Investitionsentscheidung. Wer ausrechnet, ab welcher Nutzungsmenge sich eigene Hardware oder ein selbst betriebenes Modell gegenüber der gemieteten Inferenz lohnt, rechnet mit den Preisen von heute. Diese Preise aber fallen, und mit ihnen wandert der Punkt, an dem sich der Eigenbau lohnt, immer weiter nach hinten. Jede solche Rechnung hat ein Verfallsdatum. Das ist kein Argument gegen den Eigenbetrieb; für die Souveränität kann er entscheidend sein, wie Kapitel 7 und 9 gezeigt haben. Es ist aber eines gegen die Begründung „es ist auf Dauer billiger". Auf Dauer ist Mieten oft billiger, weil der Vermieter den Preisverfall an Sie weitergibt — allerdings nur an den, der wechselt: Der Preis fällt je Leistungsniveau, nicht je Vertrag. Wer sein Modell nie überprüft und nie umzieht, zahlt die Preise von gestern.

Im deutschsprachigen Raum kommt ein handfester Standortfaktor hinzu. Rechenleistung im eigenen Haus zu betreiben heißt, sie mit eigenem Strom zu betreiben, und der ist hier teuer: In der zweiten Hälfte des Jahres 2024 zahlten deutsche Nicht-Haushaltskunden mit rund neununddreißig Cent je Kilowattstunde die höchsten Strompreise der Europäischen Union, weit über dem Durchschnitt von etwa neunzehn Cent.[8] Wer also den Eigenbetrieb gegen die Miete rechnet, rechnet im deutschspra-

chigen Raum gegen einen doppelten Gegenwind: fallende Mietpreise und hohe heimische Energiekosten.

> **Stolperfalle — Den Nenner optimieren und den Zähler vergessen** Der teuerste Fehler in diesem Feld ist, alle Energie auf die Senkung der Token-Kosten zu richten und nie zu fragen, welchen Wert die Anwendung überhaupt schafft. Ein perfekt optimierter Prozess, der das falsche Problem löst, ist Geldverschwendung mit Nachkommastellen. Bevor Sie an den Kosten feilen, klären Sie den Nutzen: Was ist dieser Vorgang dem Geschäft wert, und woran messen wir das?

Der Hebel sitzt im Zähler: vom Token-Preis zum Wert je Vorgang

Damit zurück zum Wert, denn dort entscheidet sich alles. Die Disziplin, die im Cloud-Betrieb als Kostenverantwortung gewachsen ist, hat dafür ein brauchbares Vorgehen entwickelt: Sie beginnt bei den Kosten je Token und bewegt sich von dort zu ergebnisbezogenen Maßstäben, etwa den Kosten je erledigtem Vorgang, je beantworteter Anfrage, je abgewendetem Servicefall.[9] Das ist der entscheidende Schritt. Eine Zahl wie „Kosten je gelöstem Servicefall" verbindet beide Seiten des Bruchs: Sie sinkt, wenn die Tokens billiger oder sparsamer werden, und sie sinkt ebenso, wenn das System mehr Fälle wirklich löst. Sie zwingt zu der Frage, die nach Kapitel 2 in zu vielen Häusern niemand beantwortet: Was bringt das, und wer steht dafür gerade?

Mach es heute — Ihre erste Zahl je Vorgang Nehmen Sie die KI-Monatsrechnung Ihres wichtigsten Anwendungsfalls und teilen Sie sie durch die Zahl der Vorgänge, die das System im selben Monat wirklich erledigt hat — was „erledigt" heißt, definiert Ihre Beispielsammlung aus Kapitel 6. Das ist der Nenner. Für den Zähler schätzen Sie, was ein erledigter Vorgang dem Geschäft wert ist, auf einem von drei Wegen: ersparte Bearbeitungszeit mal Vollkostensatz, vermiedener Folgeaufwand (der Rückruf, der Fehlerfall) oder zusätzlicher Ertrag. Zwei Zahlen, eine halbe Stunde — und Sie haben Ihre erste eigene Return-on-Tokens-Rechnung. Grob, aber aussagekräftiger als jede Gesamtrechnung am Monatsende.

Die Return-on-Tokens-Rechnung Eine Anwendung im Kundendienst beantwortet eine typische Anfrage mit etwa 1.000 Tokens Eingabe und 500 Tokens Ausgabe. Auf einem kleinen Modell kostet das je Anfrage Bruchteile eines Cents, auf dem stärksten Frontier-Modell leicht das Zwanzig- bis Dreißigfache.[10] Rechnen wir es durch, mit offengelegten Annahmen: Von hundert Anfragen löst das kleine Modell 60 abschließend, das große 70; jeder ungelöste Fall kostet einen Rückruf, angesetzt mit acht Euro. Das kleine Modell kostet je hundert Anfragen rund 10 Cent an Tokens plus 320 Euro an Rückrufen — etwa 5,30 Euro je gelöstem Fall. Das große kostet rund 3 Euro an Tokens plus 240 Euro an Rückrufen — etwa 3,50 Euro je gelöstem Fall. Das dreißigfach teurere Modell ist je gelöstem Fall das deutlich günstigere. Erst die Zahl je gelöstem Fall, nicht der Token-Preis, sagt Ihnen, welches Modell Sie wählen sollten. (Zahlen illustrativ; Token-Preise Stand Mitte 2026, sie ändern sich rasch.)

Dieser Perspektivwechsel ist auch der Grund, warum Deutsch in dieser Rechnung eine Rolle spielt, aber nicht die Hauptrolle. Derselbe Inhalt braucht auf Deutsch rund dreißig bis fünfundsechzig Prozent mehr Tokens als auf Englisch (Kapitel 3), und die Ausgabe kostet ohnehin ein Mehrfaches der Eingabe. Das verteuert den Nenner spürbar und ist ein realer Standortnachteil. Aber es ändert nichts an der Logik: Wer den deutschen Aufschlag

durch ein sparsameres Modell oder Zwischenspeichern auffängt, gewinnt am Nenner; wer das richtige Problem mit dem eigenen Kontext löst, gewinnt am Zähler. Der zweite Hebel ist der größere.

Der deutsche Befund: Kosten als Sorge, nicht als Ausrede

Die Sorge vor den Kosten ist im deutschsprachigen Raum real und berechtigt. Nach einer Erhebung des Branchenverbands Bitkom von 2023 nennen rund vierzig Prozent der Unternehmen, die sich mit generativer KI befassen, die Kosten als Hürde, und ein Drittel der KI-Nutzer berichtet von deutlich höheren Kosten als erwartet.[11] Diese Sorge verdient eine ehrliche Antwort, und die lautet nicht „es wird schon billig". Sie lautet: Kosten ohne Wertmaßstab sind unkontrollierbar, Kosten mit Wertmaßstab sind steuerbar. Ihre KI-Ausgaben im Griff hat, wer für jede Ausgabe einen Gegenwert benennen kann — nicht, wer am wenigsten ausgibt.

Das fügt sich in das ökonomische Bild, das dieses Buch trägt. Die Forschung zur Ökonomie der KI beschreibt Intelligenz als einen Eingangsfaktor mit sinkendem Stückpreis.[12] Wenn die Kosten der Kognition so fallen, verschiebt sich der Wettbewerb von der Frage, wer sich die Technik leisten kann, zu der Frage, wer je Einheit den meisten Wert herausholt. Das ist Return on Tokens, und es ist die ökonomische Kehrseite der Kernthese dieses Buches: Die billige Rechenleistung allein verschafft keinen Vorsprung. Den Vorsprung schafft die Führung, die aus ihr messbaren Wert macht.

Was dieses Kapitel für Sie ändert

Sie haben die dritte Knappheit kennengelernt und damit die drei beisammen: das Urteil, gute Arbeit zu erkennen; den Kontext, an dem die Maschine besser wird als bei jedem anderen; und den Wert je Recheneinheit, die Disziplin, aus beidem Rendite zu ziehen. Return on Tokens ist die Kennzahl, die verhindert, dass die ersten beiden Knappheiten in einem Budget versickern, das nie-

mand verantwortet. Die Heuristik für Ihr Budget hat damit drei Teile: den Stückpreis ignorieren, eine Zahl je Vorgang bilden, für jede Ausgabe einen Gegenwert verlangen. Die Preise werden weiter fallen; das nimmt Ihnen die Arbeit nicht ab, es vergrößert sie, weil die Versuchung zur sorglosen Nutzung wächst.

Damit ist Teil III abgeschlossen. Die drei Knappheiten benennen, *wo* der Vorsprung entsteht. Sie entstehen aber nicht von selbst. Ein Urteil, das niemand zur Geltung kommen lässt, ein Kontext, den keine Abteilung hebt, eine Kennzahl, für die keine Rolle geradesteht: Sie bleiben Potenzial. Was aus dem Potenzial Rendite macht, ist der Umbau des Unternehmens: seine Organisation, seine Steuerung, sein menschlicher Übergang. Ihm gehört der letzte Teil dieses Buches.

Das Wichtigste in 60 Sekunden

- **Return on Tokens (Wert je Recheneinheit)** ist die dritte Knappheit: der Geschäftswert je verbrauchter Recheneinheit. Ein Bruch — der Hebel liegt fast immer im Zähler (Wert), nicht im Nenner (Token-Preis).
- **Fallende Preise senken Ihre Rechnung nicht.** Billiger je Einheit heißt mehr Nutzung; die Gesamtrechnung steigt oft trotzdem (Jevons-Paradox). Man kann sich nicht zu einem guten Return on Tokens hinunterwarten.
- **Kosten, Tempo, Qualität** ziehen gegeneinander. Drei Hebel sparen ohne Wertverlust: das billigste Modell, das die Prüfung besteht; Zwischenspeichern; Stapeln.
- **Jede Eigenbau-Rechnung hat ein Verfallsdatum:** fallende Mietpreise und hohe deutsche Energiekosten verschieben den Break-even. Mieten ist oft auf Dauer billiger; Eigenbetrieb begründet man mit Souveränität, nicht mit „billiger".
- **Messen Sie Wert je Vorgang**, nicht Kosten je Anfrage (etwa: Kosten je gelöstem Fall). Das zwingt zu der Frage, die zu oft niemand beantwortet: Was bringt es, und wer steht dafür gerade?

Merksatz Die Recheneinheit wird täglich billiger; der Vorsprung gehört dem, der je Einheit den meisten Wert herausholt.

Drei Fragen an Ihr Unternehmen

1. Kennen Sie für Ihre wichtigste KI-Anwendung eine Zahl je Vorgang — etwa die Kosten je gelöstem Fall — oder nur die Gesamtrechnung am Monatsende?
2. Wo setzen Sie das stärkste Modell ein, wo ein günstigeres genügte — und haben Sie das an Ihrer eigenen Beispielsammlung je geprüft?
3. Welche Ihrer Eigenbau- oder Self-Hosting-Rechnungen beruhen auf den Preisen von heute, und was geschieht mit ihnen, wenn die Mietpreise sich erneut halbieren?

Quellen

1. Stanford HAI, *Artificial Intelligence Index Report* 2025, Kapitel 1 (April 2025): Die Abfrage eines Modells auf GPT-3.5-Niveau (MMLU 64,8) fiel von 20,00 $ je Mio. Tokens (November 2022) auf 0,07 $ je Mio. Tokens (Oktober 2024, Gemini-1.5-Flash-8B) — auf weniger als ein 280stel in rund zwei Jahren. Methodik: 3:1-gewichteter Eingabe-/Ausgabepreis, gestützt auf Daten von Artificial Analysis und Epoch AI. Es handelt sich um den Vergleich eines Frontier-Listenpreises mit dem günstigsten Budget-Modell desselben Niveaus, nicht um denselben Dienst. Eine verwandte Analyse von Epoch AI beziffert den Preisverfall je nach Aufgabe auf das 9- bis 900-Fache pro Jahr (Cottier et al., 12.3.2025) — dieselbe Datenbasis, daher keine unabhängige Bestätigung.
2. Guido Appenzeller, *Welcome to LLMflation*, Andreessen Horowitz, 12. November 2024: Inferenzkosten für ein festes Fähigkeitsniveau fielen rund um das Zehnfache pro Jahr, etwa das Tausendfache in drei Jahren. Illustrative Größenordnung eines Wagniskapitalgebers, über ein Fähigkeitsniveau und über Anbieter hinweg gerechnet. Ähnlich, als Führungsaussage einer interessierten Partei, Sam Altman, *Three Observations*, 9. Februar 2025: „the cost to use a given level of AI falls about 10x every 12 months."
3. Rückruf auf Kapitel 2. McKinsey & Company, *The State of AI* (Ausgabe 2025): rund 88 % der Unternehmen nutzen KI in mindestens einem Bereich, etwa 39 % melden einen messbaren EBIT-Effekt, nur rund 6 % („AI high performers") führen mehr als 5 % ihres EBIT auf KI zurück. Vor Druck mit den in Kapitel 2 zitierten Werten abgleichen.
4. William Stanley Jevons, *The Coal Question* (1865), Kapitel 7: „It is a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption. The very contrary is the truth." Der Effizienzgewinn senkt den Stückverbrauch, erhöht aber durch vermehrte Nutzung den Gesamtverbrauch (Jevons-Paradox / Rebound-Effekt).
5. Satya Nadella (Vorstandsvorsitzender von Microsoft) auf der Plattform X, 27. Januar 2025, in Reaktion auf die Marktreaktion nach dem Erscheinen von DeepSeek: „Jevons paradox strikes again! As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of." Wortlaut über Fortune, NPR Planet Money und Business Today gesichert; der Originalbeitrag ist nur eingeschränkt abrufbar. Aussage einer interessierten Partei; der zugrunde liegende Mechanismus ist davon unberührt.
6. Chen, Zaharia, Zou (Stanford), *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*, arXiv:2305.05176 (2023): Eine Modell-Kaskade (zunächst ein günstiges, nur bei Bedarf ein teures Modell) kann die Qualität des stärksten Einzelmodells „mit bis zu 98 % Kostensenkung" erreichen. Datensatzspezifischer Wert aus einem Preprint von 2023 (vor der jüngsten Modellgeneration); als Beleg für das Prinzip der Staffelung zu lesen, nicht als garantierte Einsparung.
7. Zwischenspeichern und Stapeln: Anbieterdokumentation (OpenAI, Anthropic, Google), Stand Juni 2026. Zwischengespeicherte Eingabe-Tokens sind

bei den großen Anbietern bis zu rund 90 % günstiger; die Stapelverarbeitung (asynchron, in der Regel binnen 24 Stunden) gewährt rund 50 % Rabatt. Bedingungen ändern sich rasch und sind vor verbindlicher Kalkulation am aktuellen Anbieterstand zu prüfen.

8. Eurostat, *Electricity price statistics* (Pressemitteilung vom 6. Mai 2025, Daten zweite Jahreshälfte 2024): Deutschland wies mit rund 0,3943 €/kWh den höchsten Strompreis für Nicht-Haushaltskunden in der EU aus (EU-Durchschnitt rund 0,19 €/kWh). Standard-Verbrauchsband; energieintensive Betriebe zahlen mit Entlastungen real deutlich weniger. Periode H2 2024; die Rangfolge verschiebt sich (in H2 2025 lag Irland vorn). Exakter Wert vor Druck aus Eurostat-Tabelle nrg_pc_205 zu ziehen.
9. FinOps Foundation, *Unit Economics* sowie *FinOps for AI* (framework-Dokumentation, Stand Juni 2026): „early unit economics often starts with cost per token and expands toward outcome oriented measures, for example cost per assist, cost per agent action, or cost per case deflected." Laut *State of FinOps 2025* steuern inzwischen rund 63 % der Befragten ihre KI-Ausgaben aktiv (2024: 31 %); „Getting to unit economics" zählt zu den am stärksten steigenden Prioritäten. Die Formulierung „Kosten je gelöstem Fall" ist eine Paraphrase dieser ergebnisbezogenen Maßstäbe.
10. Beispielrechnung auf Basis veröffentlichter Token-Preise der großen Anbieter, Stand Mitte 2026: kleine Modelle rund 0,10–1 $ je Mio. Eingabe- und 0,40–5 $ je Mio. Ausgabe-Tokens; Frontier-Modelle rund 2–5 $ bzw. 12–30 $ — ein Preisunterschied in der Größenordnung des Zehn- bis Dreißigfachen. Die Ausgabe kostet je nach Anbieter grob das Fünf- bis Achtfache der Eingabe (vgl. Kapitel 3). Alle Werte sind schnelllebig und generisch zu lesen.
11. Bitkom, *Künstliche Intelligenz in deutschen Unternehmen* (Befragung von 605 Unternehmen ab 20 Beschäftigten, Welle 2023): rund 40 % der mit generativer KI befassten Unternehmen nennen die Kosten als Hürde; rund 33 % der KI-Nutzer berichten von deutlich höheren Kosten als erwartet. Kosten rangieren als Hürde unter Datenschutz und Fachkräftemangel; sie sind eine bedeutende, nicht die größte Hürde. Wortlaut der jeweils aktuellen Welle vor Druck prüfen.
12. Agrawal, Brynjolfsson, Korinek (Hg.), *The Economics of Transformative AI*, NBER / University of Chicago Press (2025): KI als Eingangsfaktor („intelligence-as-input") mit über die Zeit fallenden Stückkosten der Kognition; die algorithmische Effizienz steigt um eine Größenordnung von rund dem 2,5-Fachen pro Jahr. Begutachteter Sammelband; die ökonomische Schlussfolgerung, dass dann die Produktivität je Recheneinheit über den Vorsprung entscheidet, ist die eigene Position dieses Buches.

Kapitel 11 — Die Organisation neu denken

Lernziele Nach diesem Kapitel können Sie

- erklären, was „Umbau" konkret heißt — wie Information und Entscheidungen durch ein Unternehmen fließen und warum die KI das verschiebt;
- begründen, warum KI Hierarchien flacher macht, und erkennen, wo das Flachmachen in bloße Zentralisierung und Überlastung umschlägt;
- den Weg vom kleinen, verantworteten Anfang zur Breite beschreiben, statt in der Pilotitis zu enden;
- benennen, welche „Bürokratie" Sie behalten oder stärken müssen, wenn die KI in die Fläche geht.

In vielen Unternehmen wird gerade eine Ebene dünner: die mittlere. In den großen amerikanischen Entlassungswellen der Jahre 2023 und 2024 entfiel rund ein Drittel der gestrichenen Stellen auf das mittlere Management, deutlich mehr als der langjährige Anteil von etwa einem Fünftel.[1] Häufig fällt dabei das Wort künstliche Intelligenz: Die Maschine erledige die Berichte, die Auswertungen, die Weiterleitung, also brauche es weniger Menschen, die zwischen oben und unten vermitteln.

Der Schluss ist verführerisch und in dieser Form falsch. Schichten zu streichen ist nicht dasselbe, wie eine Organisation neu zu denken. Eine vielbeachtete Untersuchung über das „Flachwerden" amerikanischer Konzerne fand, dass die Zahl der direkt an den Vorstandschef Berichtenden sich über zwei Jahrzehnte mehr als verdoppelte, von knapp fünf auf fast zehn, und dass die Entscheidungen dabei nach oben wanderten statt nach unten, näher an die Spitze.[2] Heraus kam ein zentralisierteres Haus mit überlasteten Verbliebenen, kein flacheres und agileres. Wer nur die Axt anlegt, bekommt ein billigeres altes Unternehmen, kein neues.

Die richtige Frage ist deshalb nicht, wie viele Ebenen man streichen kann, sondern was die KI daran ändert, wo Wissen und Entscheidungen in einem Unternehmen hingehören. Kapitel 1 hat am Beispiel der Elektrifizierung gezeigt, dass der Ertrag einer Basistechnologie erst aus der Reorganisation kommt, nicht aus der Technik selbst. Kapitel 2 hat daraus die Frage abgeleitet, die jedes Vorhaben zu stellen hat: Welchen Arbeitsablauf bauen wir um, wer verantwortet das Ergebnis, und woran messen wir es? Dieses Kapitel beantwortet den strukturellen Teil. Es handelt davon, wie ein Unternehmen aussehen muss, in dem die drei Knappheiten zur Rendite werden.

Wozu es Hierarchie überhaupt gibt

Um zu verstehen, was die KI mit einer Struktur macht, muss man zuerst verstehen, wozu die Struktur da ist. Der Ökonom Luis Garicano hat dafür ein klares Bild geliefert: Ein Unternehmen ist eine *Wissenshierarchie*.[3] Die häufigen, einfachen Fälle werden unten erledigt, von denen, die nah an der Sache sind. Was darüber hinausgeht, eine Ausnahme, ein schwieriger Fall, wird nach oben gereicht, zu den selteneren, teureren Fachleuten. So spart eine Organisation ihr knappstes Gut: Sie muss nicht jeden mit allem Wissen ausstatten; es genügt, die Ausnahme zu dem zu leiten, der sie lösen kann. Wie viele Ebenen es gibt und wie breit jede ist, hängt nach Garicano an zwei

Kosten: was es kostet, eine Frage weiterzureichen, und was es kostet, Wissen selbst zu erwerben.

Dahinter steht eine ältere Einsicht. Der Ökonom Friedrich August von Hayek hat schon 1945 darauf bestanden, dass das entscheidende Wissen stets verstreut vorliegt, als das Wissen „der besonderen Umstände von Zeit und Ort", über das jeder an seinem Platz verfügt.[4] Daraus folgt eine Faustregel guter Organisation: Entscheidungen gehören dorthin, wo das Wissen sitzt. Die Managementforschung hat das später zugespitzt: Wo Wissen teuer zu übertragen ist, delegiert man besser die Entscheidung nach unten, als die Information nach oben zu schaffen.[5]

> **Klartext — Die Wissenshierarchie** Eine *Wissenshierarchie* ist die Organisation als Apparat zur Bewirtschaftung knappen Fachwissens: Routine wird dort gelöst, wo sie anfällt; Ausnahmen steigen zu selteneren Fachleuten auf. Ihre Form, also wie viele Ebenen und wie große Leitungsspannen, ergibt sich aus zwei Kosten: dem Aufwand, eine Frage weiterzureichen, und dem Aufwand, Wissen selbst vorzuhalten. Wer eine dieser Kosten senkt, verändert die Form.

Was die KI verschiebt

Genau hier greift die KI ein, und zwar bei der zweiten dieser beiden Kosten. Sie senkt die Kosten, Wissen an der Front verfügbar zu haben. Der Sachbearbeiter, die Pflegekraft, der Monteur hat nun, im besten Fall, einen kompetenten Berater unmittelbar am Platz, einen, der die Vorschrift kennt, den ähnlichen Fall erinnert, den Entwurf vorlegt. Damit lassen sich mehr der Fälle, die früher nach oben wandern mussten, dort lösen, wo sie entstehen. Die Folge für die Form der Organisation liegt nahe: Wenn die Ausnahme seltener eskaliert, verliert die Ebene, deren Hauptzweck das Weiterleiten und das Bündeln von Routine-Expertise war, ihren ökonomischen Grund.[6]

Das erklärt, warum es ausgerechnet die Mitte trifft. Sie war über weite Strecken eine Informationsschicht: Sie sammelte ein,

verdichtete, gab nach oben weiter und Anweisungen nach unten zurück. Vieles davon kann die Maschine. Was sie nicht kann, bleibt und gewinnt an Gewicht: das Urteil unter Unklarheit, die Rechenschaft für ein Ergebnis, die Führung von Menschen, die Abstimmung über Abteilungsgrenzen hinweg. Eine mittlere Führungskraft, die heute rund ein Fünftel ihrer Zeit — einen Arbeitstag pro Woche — mit Verwaltung verbringt, so eine McKinsey-Befragung, wird nicht überflüssig; aber der Teil ihrer Arbeit, der wertvoll ist, verschiebt sich von der Verwaltung zur Führung.[7]

Flacher ist nicht von selbst besser

Daraus folgt nicht, dass man nur Ebenen herausschneiden müsste. Der Gegenbefund aus ebenjener Untersuchung über das „Flachwerden" ist die wichtigste Warnung dieses Kapitels: Wer Schichten streicht, ohne die Entscheidungsrechte zu verschieben, zentralisiert nur. Die Ausnahme wandert dann zur überlasteten Spitze statt zur informierten Front; dort verwaltet man nun noch mehr Direktberichte und hat noch weniger Zeit zum Entscheiden. Flach ist eine Struktur erst dann ein Vorteil, wenn mit den Ebenen auch die Befugnisse nach unten gehen.

Die Technik selbst gibt diese Richtung nicht vor. Dieselbe KI, die der Front Wissen zuträgt, liefert der Spitze eine Übersicht in Echtzeit — und legt manchem Geschäftsführer nahe, nun erst recht alles selbst zu entscheiden. Was den Ausschlag gibt, ist die Art des Wissens: Das kodifizierte — die Vorschrift, der Präzedenzfall, der Vergleichswert — lässt sich billig in beide Richtungen tragen. Das situative Wissen der Front — der Kunde am Telefon, das Werkstück in der Hand — bleibt teuer nach oben zu übertragen. Deshalb gehört die Entscheidung zur Front, nicht noch mehr Information zur Spitze.

Stolperfalle — Schichten streichen, Entscheidungen oben behalten
Der teuerste Umbaufehler ist, die mittlere Ebene zu verschlanken, die Entscheidungen aber weiter oben zu treffen. Man spart Gehälter und verliert Tempo: Die Front weiß jetzt mehr als je zuvor, darf aber nichts entscheiden, und die Spitze entscheidet über Dinge, von denen sie weniger versteht als die Front. Prüfen Sie vor jedem Abbau einer Ebene, welche Entscheidung dadurch *nach unten* wandert. Wandert keine, ist es bloßer Stellenabbau und kein Umbau.

Wohin also mit den Entscheidungen? Dorthin, wo die KI das Wissen jetzt hinträgt: an die Front, zu dem, der den Fall in der Hand hat. Das verlangt zweierlei. Erstens klare Entscheidungsrechte: Für jede wichtige Frage muss benannt sein, wer sie entscheiden darf, und das sollte möglichst nah an der Sache sein. Zweitens, und das ist das Gegenstück dazu, für jedes Ergebnis genau ein verantwortlicher Kopf. Kapitel 2 hat die häufigste organisatorische Ursache des Scheiterns benannt: Niemand besitzt das Ergebnis; die Fachleute besitzen das Experiment, den Wirkbetrieb besitzt niemand. Die Abhilfe ist eine alte Regel guter Führung, keine neue Stabsstelle: ein Vorhaben, ein Verantwortlicher, der für das betriebliche Ergebnis geradesteht, nicht für die Vorführung.[8]

Dabei hilft eine Beobachtung, die der Informatiker Melvin Conway schon 1968 festhielt: Jede Organisation, die ein System entwirft, bringt einen Entwurf hervor, dessen Struktur ihre eigene Kommunikationsstruktur abbildet.[9] Wer also einen durchgängigen, KI-gestützten Ablauf will, ihn aber von vier Abteilungen bauen lässt, die nur über Schnittstellen reden, bekommt einen Ablauf, der an jeder Abteilungsgrenze bricht. Integrierte Abläufe brauchen integrierte, abteilungsübergreifende Teams. Die Struktur, in der Sie bauen, schreibt sich in das ein, was Sie bauen.

Klein anfangen, dann verbreitern

Wie beginnt man? Nicht mit einem großen Reorganisationsprogramm, das das ganze Haus auf einmal umstellt, und nicht mit dem Pilotprojekt, das beeindruckt und nie den Wirkbetrieb erreicht. Tragfähig ist der Weg dazwischen: ein einzelner Arbeitsablauf, mit einem klaren Verantwortlichen und einer Zahl, an der sich der Erfolg in drei Monaten messen lässt. Diesen einen Ablauf baut man nicht an, sondern um — von Grund auf um die Frage herum, wo das Wissen sitzt und wer entscheidet. Funktioniert er, wird er zur Vorlage: Erst dann verbreitert man, Ablauf für Ablauf.

Für dieses Vorgehen spricht zweierlei. Die Deutsche Akademie der Technikwissenschaften rät dem Mittelstand ausdrücklich, mit kleinen „Leuchtturm"-Projekten zu beginnen, statt das ganze Unternehmen umzukrempeln, und an ihnen zu lernen.[10] Und die Beratungshäuser, die das Skalieren beobachten, finden mit einiger Regelmäßigkeit dasselbe: Was die wenigen Unternehmen, die aus KI einen Ergebnisbeitrag ziehen, von der Mehrheit trennt, ist nicht das bessere Modell, sondern der grundlegend neu gebaute Arbeitsablauf, samt einer Führung, die sich darum kümmert.[11] Das deckt sich mit der in Kapitel 2 genannten Faustregel, dass der Löwenanteil der Mühe in Menschen, Prozesse und Organisation gehört, nicht in die Technik.

> **Mach es heute — Einen Ablauf kartieren** Nehmen Sie einen wichtigen Arbeitsablauf und zeichnen Sie ihn in dreißig Minuten auf, ehrlich: An welcher Stelle wird ein Fall heute nach oben gereicht, und warum? Wer entscheidet ihn? Welches Wissen fehlt der Front, das die KI ihr geben könnte? Und wenn die Front es hätte: welche Entscheidung könnte dann dort fallen, wo der Fall entsteht? Wo Sie eine solche Entscheidung finden, haben Sie den ersten Ablauf, der sich zu lohnen verspricht.

Welche Bürokratie bleiben muss

Es wäre ein Missverständnis, den Umbau als Befreiung von aller Bürokratie zu lesen. Manche Regelhaftigkeit ist kein Hemmschuh, sondern der Grund, warum man einer Organisation trauen kann. Max Weber hat die Tugenden der regelgebundenen Verwaltung vor hundert Jahren nüchtern benannt: Berechenbarkeit, Gleichbehandlung, Nachvollziehbarkeit, die Aktenführung, die festhält, wer was aus welchem Grund entschieden hat.[12] Diese Güter werden wertvoller, nicht geringer, wenn Entscheidungen an die Front und an die Maschine wandern. Wenn ein KI-gestützter Ablauf einem Kunden einen Kredit verweigert oder einer Patientin eine Behandlung vorschlägt, muss nachvollziehbar bleiben, auf welcher Grundlage das geschah, wer es hätte prüfen können und wer dafür einsteht. Entscheidungs-Protokolle, klare Eskalationswege und eine menschliche Aufsicht sind dann die Bürokratie, die der Umbau braucht; ein Teil davon wird vom europäischen KI-Recht ohnehin verlangt, worauf das nächste Kapitel eingeht.

Die Kunst liegt in der Unterscheidung. Behalten und stärken sollte man die Bürokratie, die Rechenschaft schafft. Streichen kann man die, die nur Information weiterreichte, welche heute eine Maschine weiterreicht. Und man sollte einer Versuchung widerstehen: Dieselbe KI, die Bürokratie abbauen könnte, kann sie auch vermehren. Wo das Verfassen von Berichten nichts mehr kostet, entstehen mehr Berichte, die niemand liest: gefällige Hülsen, die nach Arbeit aussehen und keine sind, und die anderen erst recht Arbeit machen.[13] Mehr Output ist nicht mehr Wert. Auch hier gilt die Kennzahl aus dem vorigen Kapitel: Was zählt, ist der Wert je Vorgang, nicht die Menge des Erzeugten.

Der deutsche Befund: ein flacher Anfang, aber Abteilungsmauern

Für den deutschsprachigen Raum hält dieser Umbau eine gute und eine schlechte Nachricht bereit. Die gute: Ein großer Teil seiner Stärke, der industrielle Mittelstand, ist strukturell schon

nah an dem, was hier verlangt wird. Die „Hidden Champions", von denen allein in Deutschland rund anderthalbtausend sitzen, sind bekannt für flache Hierarchien und kurze, direkte Wege zur Spitze.[14] Wo Wege kurz sind, lässt sich eine Entscheidung leichter dorthin verschieben, wo das Wissen sitzt. Doch dieselbe Nähe hat eine Kehrseite: Diese Häuser sind meist eigentümer- oder familiengeführt, und die kurze Linie endet häufig bei einer Person, die vieles selbst entscheidet. Die flache Struktur erleichtert den Umbau, sie ersetzt ihn nicht. Zum Vorsprung wird sie erst, wenn die Spitze die Entscheidungen, die nun an der informierten Front fallen könnten, auch tatsächlich dorthin abgibt.

Die schlechte Nachricht ist die alte: das Denken in Abteilungen. In einer Erhebung aus dem Jahr 2015 nannten Führungskräfte das „Silodenken der Fachabteilungen" als größte Hürde auf dem Weg zur digitalen Organisation, mit Abstand vor allen technischen Fragen.[15] Wir sind dem schon in Kapitel 2 begegnet, wo die Daten in getrennten Silos lagen; hier kehrt es als Strukturproblem wieder. Conways Beobachtung sagt voraus, was daraus folgt: Wer in Silos organisiert ist, baut Abläufe mit Brüchen an den Silogrenzen. Der Umbau verlangt deshalb gerade hier den Mut, Teams quer zu den Abteilungen zu bilden.

Eine letzte Eigenheit verdient ein faires Wort, weil sie oft missverstanden wird. Die deutsche Organisation gilt als hierarchisch; genauer ist sie regel-, prozess- und fachautoritätsgebunden. Sie misstraut der Improvisation und vertraut dem geordneten Verfahren; die Prokura, die gesetzlich geregelte, im Handelsregister eingetragene Vertretungsmacht, ist ein Sinnbild dieser Kultur der verbrieften Zuständigkeit.[16] Das ist zweischneidig. Es ist ein Aktivposten für eine nachvollziehbare, prüffeste KI-Steuerung, wie das nächste Kapitel sie verlangt, und zugleich eine Bremse, wenn Entscheidungen rasch an die Front sollen. Beides bewusst zu führen, die Stärke zu nutzen und die Bremse zu lösen, ist die eigentliche Führungsaufgabe dieses Umbaus. Wie man die Belegschaft dabei zur Mitgestalterin macht, statt sie nur mitzunehmen, und welche Rolle die Mitbestimmung dabei spielt, ist Sache des übernächsten Kapitels.

Was dieses Kapitel für Sie ändert

Der Umbau beginnt mit der Struktur, und die Struktur beginnt mit einer einzigen Verschiebung: Entscheidungen wandern dorthin, wo die KI das Wissen nun hinträgt, an die Front, und jedes Ergebnis bekommt einen verantwortlichen Kopf. Daraus folgt, dass die mittlere Schicht dünner wird, weil ihre alte Aufgabe, Information weiterzureichen, entfällt, nicht weil man spart; was bleibt, ist Führung. Sie fangen mit einem Ablauf an, bauen ihn um statt an, und verbreitern erst, wenn er trägt. Und Sie behalten die Bürokratie, die Rechenschaft schafft, während Sie die streichen, die nur Information verschob.

Damit steht das erste der drei Stücke des Umbaus. Eine Struktur, die Entscheidungen an die Front verlegt und Maschinen mitentscheiden lässt, wirft sofort die nächste Frage auf: Wie hält man das sicher und verantwortbar? Wer beaufsichtigt die Agenten, wer haftet, was verlangt das Gesetz? Das ist die Governance, und ihr gehört das nächste Kapitel.

Das Wichtigste in 60 Sekunden

- Ein Unternehmen ist eine **Wissenshierarchie**: Routine unten, Ausnahmen steigen auf. KI senkt die Kosten, Fachwissen an der Front zu haben — also werden mehr Fälle unten gelöst, und die mittlere **Informationsschicht** dünnt aus.
- **Flacher ist nicht von selbst besser.** Schichten zu streichen, ohne Entscheidungsrechte nach unten zu verschieben, zentralisiert nur und überlastet die Spitze (CEO-Leitungsspanne historisch von ~5 auf ~10 verdoppelt).
- **Entscheidungen an die Front, ein Kopf je Ergebnis.** Integrierte KI-Abläufe brauchen abteilungsübergreifende Teams (Conway: das System bildet die Organisation ab, die es baut).
- **Klein anfangen, dann verbreitern:** ein Ablauf, ein Verantwortlicher, eine Zahl; umbauen statt anbauen; bei Erfolg zur Vorlage machen. Nicht das große Programm, nicht die Pilotitis.
- **Behalten Sie die Bürokratie, die Rechenschaft schafft** (Protokolle, Aufsicht, Prüfspur); streichen Sie die, die nur Information weiterreichte. KI kann Bürokratie auch *vermehren*.

Merksatz Den Vorsprung schafft nicht die flachere Struktur, sondern die Entscheidung, die endlich dort fällt, wo das Wissen sitzt.

Drei Fragen an Ihr Unternehmen

1. An welcher Stelle wird in Ihrem wichtigsten Ablauf ein Fall nach oben gereicht — und könnte ihn die mit KI ausgestattete Front nicht selbst entscheiden, wenn sie dürfte?
2. Wenn Sie zuletzt eine Führungsebene verschlankt haben: Welche Entscheidung ist dadurch nach unten gewandert — oder nur nach oben?
3. Welche Ihrer Verfahren schaffen echte Rechenschaft (und müssen bleiben), und welche reichen nur Information weiter, die heute eine Maschine bewegt?

Quellen

1. Anteil des mittleren Managements an Entlassungen: rund 32 % (2023) bzw. 29 % (2024) gegenüber einer Basislinie von etwa 20 % (2018–2022). Quelle: Live Data Technologies, berichtet von Bloomberg/Business Insider, 2024. Es handelt sich um einen privaten Stellenwechsel-Datensatz, nicht um amtliche Statistik; die Abgrenzung des „mittleren Managements" variiert. Als Indiz für einen realen Trend zu lesen, nicht als exakte Quote.
2. Guadalupe, Li & Wulf, *The Flattened Firm: Not as Advertised*, California Management Review / Harvard Business School Working Paper 12-087 (2012). Die Zahl der direkt an den CEO Berichtenden stieg über rund zwei Jahrzehnte von etwa 4,7 auf 9,8; Entscheidungsrechte verlagerten sich zur Spitze. Das Flachwerden glich eher einer Zentralisierung als einer Dezentralisierung.
3. Luis Garicano, *Hierarchies and the Organization of Knowledge in Production*, Journal of Political Economy 108(5), 2000. Organisationen als Wissenshierarchien; Leitungsspanne und Ebenenzahl ergeben sich aus Kommunikations- und Wissenserwerbskosten. Stilisiertes Gleichgewichtsmodell, kein empirisch geschätztes Ergebnis.
4. Friedrich A. von Hayek, *The Use of Knowledge in Society*, American Economic Review 35(4), 1945: das entscheidende Wissen liegt verstreut vor, als Wissen „der besonderen Umstände von Zeit und Ort". Die pro-dezentrale Schlussfolgerung ist Hayeks (gegen die zentrale Planung gerichtetes) Argument, kein gemessener Befund.
5. Jensen & Meckling, *Specific and General Knowledge, and Organizational Structure*, in Werin & Wijkander (Hg.), *Contract Economics*, 1992 (Nachdruck Journal of Applied Corporate Finance 8(2), 1995): „spezifisches" Wissen ist teuer zu übertragen, weshalb Entscheidungsrechte besser dorthin delegiert werden, wo dieses Wissen sitzt — gekoppelt an Messung und Anreize.
6. Garicano & Rossi-Hansberg, *Knowledge-Based Hierarchies: Using Organizations to Understand the Economy*, Annual Review of Economics 7, 2015 (Modellkern: Quarterly Journal of Economics 121(4), 2006): Sinkende Kommunikations- und Lernkosten verändern die optimale Hierarchieform. Die Anwendung speziell auf KI ist Gegenstand jüngerer, noch nicht begutachteter Arbeiten (u. a. Ide & Talamàs, arXiv:2312.05481, 2025: „co-pilot"-KI hilft den weniger Wissenden, autonome KI eher den Wissendsten) — als Hypothese, nicht als gesicherter Befund zu lesen.
7. McKinsey, *Stop wasting your most precious resource: middle managers*, 2023 (Befragung 2022): mittlere Führungskräfte verbringen unter 25 % ihrer Zeit mit Mitarbeiterführung und rund 20 % mit Verwaltung. Selbstauskunft. Die Aussage, KI verschiebe diese Arbeit von der Verwaltung zur Führung, ist eine begründete Folgerung; eine Copilot-Feldstudie (referiert in Harvard Business Review, 2025) zeigte bei Entwicklern einen Rückgang der Koordinationstätigkeit um rund 10 %.
8. Zur Abhilfe „ein Vorhaben, ein Verantwortlicher": Rogers & Blenko, *Who Has the D? How Clear Decision Roles Enhance Organizational Performance*, Harvard Business Review 2006 (RAPID-Modell mit genau einer „Decide"-Rolle); als ge-

lebte Praxis der „single-threaded owner" bei Amazon dokumentiert in Bryar & Carr, *Working Backwards*, 2021. Beratungs- bzw. Firmenpraxis, illustrativ, keine kausale Studie.

9. Melvin E. Conway, *How Do Committees Invent?*, Datamation 14(4), 1968: „Any organization that designs a system … will produce a design whose structure is a copy of the organization's communication structure." Empirisch gestützt (für Software) durch MacCormack, Rusnak & Baldwin, *Exploring the Duality between Product and Organizational Architectures*, Research Policy 41(8), 2012 („mirroring hypothesis").
10. acatech — Deutsche Akademie der Technikwissenschaften, *Industrie 4.0 Maturity Index* (Update 2020): der Erfolg hängt wesentlich von Organisationsstruktur und Kultur ab; ein vollständiger Umbau ist nicht nötig, empfohlen wird der Beginn mit kleinen Leuchtturm-Projekten. (Bereits in Kapitel 2 für das Silo-Problem angeführt.)
11. McKinsey, *The State of AI* (Ausgabe 2025): das grundlegende Neugestalten von Arbeitsabläufen hat den größten Effekt auf den EBIT-Beitrag der KI; „High Performer" haben ihre Abläufe weit häufiger neu gebaut, und nur bei einer Minderheit beaufsichtigt die Unternehmensspitze die KI-Governance — jener Faktor, der am stärksten mit der Ergebniswirkung zusammenhängt. Rund zwei Drittel haben mit dem Skalieren noch nicht begonnen. Selbstauskunft; die Werte (88 %/39 %/6 %, BCG-Faustregel 10-20-70) sind die in Kapitel 2 zitierten und werden hier nur zurückgerufen.
12. Max Weber, *Wirtschaft und Gesellschaft* (1922), zur rational-legalen Bürokratie: regelgebundene Zuständigkeit, Amtshierarchie, Aktenführung, unpersönliche Regelanwendung; Tugenden sind Präzision, Berechenbarkeit, Gleichbehandlung, Nachvollziehbarkeit. Idealtypus, nicht Checkliste; Weber selbst sah die Ambivalenz („stahlhartes Gehäuse"). Zur ausdrücklichen Verteidigung dieser Güter: Paul du Gay, *In Praise of Bureaucracy*, 2000.
13. Zum Risiko, dass KI Bürokratie vermehrt: BetterUp Labs & Stanford Social Media Lab, referiert in Harvard Business Review, *AI-Generated „Workslop" Is Destroying Productivity*, September 2025: 41 % der Befragten begegneten KI-Ausgaben, die „nach guter Arbeit aussehen, aber ohne Substanz" sind, mit im Schnitt rund zwei Stunden Nacharbeit. Einzelne Selbstauskunfts-Umfrage, nicht begutachtet; als Größenordnung zu lesen.
14. Hermann Simon, *Hidden Champions* (deutsche Ausgaben seit 2007; *Hidden Champions of the 21st Century*, 2009): in Deutschland rund 1.500–1.600 weltmarktführende Mittelständler, etwa die Hälfte aller weltweit. Typische Merkmale: flache Hierarchien und kurze Wege zur Spitze, zugleich überwiegend eigentümer- bzw. familiengeführte, zentrale Führung. Die genaue Zahl variiert je nach Ausgabe und Abgrenzung.
15. Hays / Pierre Audoin Consultants, *Von starren Prozessen zu agilen Projekten — Unternehmen in der digitalen Transformation*, 2015 (Befragung von 225 Führungskräften): „Wettbewerbs- und Silodenken der Fachabteilungen" wurde als größte Hürde einer digitalen Organisation genannt (72 %, in großen Unternehmen höher). Beratungsstudie aus dem Jahr 2015; als gut belegtes, aber

datiertes Stimmungsbild zu lesen, gestützt durch die acatech-Befunde zu Datensilos (Anm. 10).

16. Zur deutschen Führungskultur: in den Kulturdimensionen nach Hofstede weist Deutschland eine hohe Unsicherheitsvermeidung (rund 65) bei zugleich niedriger Machtdistanz (rund 35) auf — die Autorität ist eher regel- und fachgebunden als status-steil. Die Prokura (§§ 48 ff. Handelsgesetzbuch) — eine gesetzlich umrissene, im Handelsregister eingetragene Vertretungsmacht — illustriert diese Kultur der verbrieften, formalen Zuständigkeit.

Kapitel 12 — Governance und Sicherheit

Lernziele Nach diesem Kapitel können Sie

- KI und Agenten im Betrieb sicher beaufsichtigen — mit abgestufter Autonomie, geringsten Rechten und Freigaben an der Stelle, wo aus Lesen Handeln wird;
- den EU AI Act praktisch einordnen: welche Pflichten Sie als Betreiber wirklich treffen und welche nicht;
- ein Risikoregister führen, das diffuse Sorge in gesteuertes Risiko verwandelt;
- begründen, warum Governance Chefsache ist und sich nicht an die IT-Abteilung wegdelegieren lässt.

Die Werkzeuge sind im Haus, die Abläufe werden umgebaut, und an immer mehr Stellen handeln Maschinen mit. Die Regeln dafür fehlen meist. Nach einer Erhebung des Branchenverbands Bitkom haben nur dreiundzwanzig Prozent der deutschen Unternehmen feste Regeln für den Einsatz generativer KI; gut ein Drittel hat sich mit der Frage noch gar nicht befasst.[1] So klafft eine Lücke zwischen dem, was die Systeme schon dürfen, und dem, was das Unternehmen über sie weiß und an ihnen steuert.

Wie schnell diese Lücke teuer wird, hat Kapitel 5 am Fall „EchoLeak" gezeigt, bei dem ein Assistent, der fremde Inhalte

liest und zugleich handeln darf, zur Datenschleuse wurde, ohne dass ein Nutzer etwas falsch machte.[2] Governance ist die Antwort darauf, und sie ist nicht die Bremse, als die sie oft gilt. Sie ist das, was den Umbau überlebbar macht: Erst wer seine Systeme beaufsichtigen, prüfen und im Ernstfall anhalten kann, kann ihnen überhaupt etwas anvertrauen. Dieses Kapitel nimmt die Fäden auf, die frühere Kapitel liegen gelassen haben: die Aufsicht über Agenten, die Angriffsfläche im Betrieb, das Risikoregister, den EU AI Act und das Datenrecht in der Anwendung.

Abgestufte Autonomie: Vertrauen, das man sich verdient

Die wichtigste operative Regel folgt unmittelbar aus Kapitel 5. Dort lag die entscheidende Linie zwischen Lesen und Handeln: Eine lesende KI, die einen Entwurf vorschlägt, ist harmlos; eine handelnde, die bestellt, überweist, versendet oder Daten ändert, greift in die Welt ein. Daraus wird im Betrieb ein einfaches Prinzip: Ein System verdient sich seine Rechte, es bekommt sie nicht geschenkt. Man beginnt mit der geringsten Stufe, dem reinen Lesen und Vorschlagen, und erweitert die Befugnisse erst, wenn sich das System an der eigenen Aufgabe bewährt hat. Je unumkehrbarer eine Aktion, desto eher gehört ein Mensch dazwischen, der sie freigibt.

> **Klartext — Abgestufte Autonomie und geringste Rechte** *Abgestufte Autonomie* heißt: Ein KI-System erhält nur so viel Handlungsspielraum, wie es sich an der eigenen Aufgabe nachweislich verdient hat, von „darf nur lesen und vorschlagen" über „darf mit Freigabe handeln" bis „darf in engen Grenzen selbst handeln". *Geringste Rechte* (in der IT-Sicherheit das Prinzip der minimalen Berechtigung) heißt: Jedes Werkzeug, auf das ein Agent zugreifen darf, ist so eng zugeschnitten wie möglich. Was ein Agent nicht darf, kann er auch nicht missbrauchen, und was er nicht erreicht, kann kein Angreifer über ihn erreichen.

Dieselbe Logik beantwortet die offene Frage aus Kapitel 5, wie man mit der Angriffsfläche umgeht. Ein Agent, der fremde Inhalte verarbeitet und zugleich handeln darf, ist durch eingeschleuste Anweisungen verführbar: die Prompt-Injection. Vollständig verhindern lässt sie sich nicht, weil Anweisung und Daten denselben Kanal teilen. Aber man kann den Schaden begrenzen: enge Rechte, Freigaben vor heiklen Aktionen, Filter, abgeschottete Ausführungsumgebungen, Protokollierung jeder Aktion und regelmäßige Angriffsversuche aus den eigenen Reihen. Die internationale Sicherheits-Initiative OWASP führt die typischen Schwachstellen von KI-Anwendungen und Agenten in zwei vielbeachteten Listen, die im Kern genau auf diese Maßnahmen hinauslaufen, allen voran die Warnung vor zu weit gefassten Handlungsrechten.[3]

Was im Betrieb bleibt: Prüfen hört nicht beim Start auf

Kapitel 6 hat gezeigt, dass man ein KI-System an einer eigenen, aufgabennahen Beispielsammlung prüft, ehe man ihm traut. Im Betrieb hört dieses Prüfen nicht auf, es verstetigt sich. Modelle ändern sich, Anbieter tauschen Versionen aus, und die Wirklichkeit, auf die das System trifft, wandert von der ab, an der es einmal gut war. Wer den Erfolg nicht fortlaufend misst, merkt nicht, wenn die Qualität schleichend abrutscht. Zur Governance gehört deshalb die laufende Beobachtung: Stichproben, eine Auswahl an Fällen, die ein Mensch nachprüft, eine Kennzahl, die Alarm schlägt, wenn sie kippt. Die anerkannten Rahmenwerke benennen das als eigene Daueraufgabe; das amerikanische Normungsinstitut NIST etwa fasst KI-Risikoarbeit in den vier Schritten Steuern, Erkennen, Messen, Behandeln, von denen keiner je „fertig" ist.[4]

Das Risikoregister

Diffuse Sorge ist ein schlechter Ratgeber; sie führt entweder zur Lähmung oder zum Wegschauen. Das Mittel dagegen ist alt und unspektakulär: ein Register. Man trägt zusammen, wo im Haus

KI eingesetzt wird, und für jeden Einsatz vier Dinge: was schiefgehen kann, wie wahrscheinlich und wie schwer das wäre, was dagegen getan ist, und wer dafür verantwortlich ist. So wird aus einem Nebel von Möglichkeiten eine überschaubare Liste von Posten, die man steuern kann. Die internationale Norm ISO 42001 beschreibt ein solches KI-Managementsystem; Unternehmen können sich danach zertifizieren lassen, und in Ausschreibungen wird das zunehmend verlangt.[5] Entscheidend ist die Gewohnheit, die ein solches Siegel erzwingt: hinzusehen, bevor etwas passiert.

Die erste Zeile eines solchen Registers ist in den meisten Häusern übrigens keine Agenten-Anwendung, sondern die ungeregelte Alltagsnutzung: die privat beschafften Chatbots, in die Beschäftigte längst Kundendaten tippen. Eine Seite Nutzungsregeln — welche Werkzeuge erlaubt sind, welche Datenkategorien hineindürfen, wann der Einsatz offenzulegen ist — schließt die größte Lücke zuerst und kostet fast nichts.

> **Mach es heute — Das Register auf einer Seite** Nehmen Sie Ihren autonomsten KI-Einsatz, also den, der am meisten selbst tut, und füllen Sie für ihn fünf Spalten aus: Was tut das System? Was ist das Schlimmste, das es anrichten könnte? Wie wahrscheinlich ist das? Was hält es davon ab? Und wer im Haus steht dafür gerade? Wenn Sie die letzte Spalte nicht ausfüllen können, haben Sie das dringendste Governance-Problem schon gefunden.

Die letzte Spalte ist die wichtigste, und sie schließt an Kapitel 11 an. Ein Risiko, das niemandem gehört, wird nicht behandelt. Für jedes KI-System gehört ein verantwortlicher Kopf benannt: nicht die Datenabteilung im Allgemeinen, sondern ein Mensch, der für das betriebliche Ergebnis und seine Folgen geradesteht.

Der EU AI Act, praktisch

Kaum ein Gesetz beschäftigt deutsche Vorstände derzeit so sehr wie der EU AI Act, und kaum eines wird so oft missverstanden.

Drei Klarstellungen helfen. Erstens ist er risikobasiert: Er teilt KI-Anwendungen in wenige verbotene, einige hochriskante, einige mit bloßer Kennzeichnungspflicht und einen großen Rest mit minimalem Risiko, für den keine besonderen Pflichten gelten.[6] Die meisten betrieblichen Anwendungen fallen in diesen Rest. Eine Pflicht bleibt freilich auch dort: Wer KI einsetzt, muss seit Februar 2025 dafür sorgen, dass seine Leute genug davon verstehen — das Gesetz nennt es KI-Kompetenz (Artikel 4), und es verlangt damit nichts anderes, als was dieses Buch ohnehin für unverzichtbar hält.[6] Wer einen Assistenten für Entwürfe oder eine Auswertung einsetzt, betreibt in aller Regel kein Hochrisiko-System. Panik ist also fehl am Platz; das Erste, was ein Unternehmen braucht, ist eine nüchterne Einordnung seiner Anwendungen.

Zweitens: Selbst wo ein System hochriskant ist, treffen die schweren Pflichten zunächst den Anbieter, der es baut, nicht den Betreiber, der es nutzt. Als Betreiber haben Sie eine kürzere, handhabbare Liste: das System gemäß Anleitung einsetzen, eine kompetente menschliche Aufsicht bestellen, den Betrieb überwachen und schwere Vorfälle melden, die Protokolle mindestens ein halbes Jahr aufbewahren und die Betroffenen informieren.[7] Eine Falle ist dabei zu kennen: Wer ein fremdes System unter eigenem Namen vertreibt, es wesentlich verändert oder für einen anderen Zweck einsetzt, gilt rechtlich selbst als Anbieter und erbt dessen volle Pflichten. Die Geldbußen sind ernst: bis zu fünfunddreißig Millionen Euro oder sieben Prozent des weltweiten Jahresumsatzes für verbotene Praktiken, bis zu fünfzehn Millionen oder drei Prozent für andere Verstöße.[7]

Drittens, und das ist beim Schreiben dieses Buches der wichtigste Vorbehalt: Der Zeitplan ist in Bewegung. Stand Mitte 2026 hatte die im November 2025 vorgeschlagene Initiative „Digital Omnibus" eine vorläufige Einigung von Parlament und Rat erreicht, die Pflichten für Hochrisiko-Systeme deutlich nach hinten zu verschieben, auf Dezember 2027 und August 2028; förmlich verabschiedet war sie noch nicht, und bis zur Veröffentli-

chung galten die ursprünglichen Fristen weiter.[8] Für die Praxis heißt das zweierlei: Verlassen Sie sich auf kein einzelnes Datum, prüfen Sie den Stand, bevor Sie planen. Und warten Sie nicht auf die letzte Frist, denn was das Gesetz für Hochrisiko-Systeme verlangt, ist im Kern das, was Kapitel 6 ohnehin als gute Praxis beschrieben hat: belegen können, dass das System funktioniert. Wer das schon kann, hat die Regulierung nicht zu fürchten.

Zahlen, die zählen

- Nur **23 %** der deutschen Unternehmen haben feste Regeln für generative KI; **36 %** haben sich damit noch nicht befasst.[1]
- EU-AI-Act-Bußgelder: bis **35 Mio. €** / **7 %** Weltumsatz (verbotene Praktiken), bis **15 Mio. € / 3 %** (andere Verstöße).[7]
- Hochrisiko-Pflichten: ursprünglich ab **August 2026/2027**, nach dem „Digital Omnibus" voraussichtlich auf **Dezember 2027 / August 2028** verschoben, Stand Mitte 2026 noch nicht endgültig.[8]

Auch im eigenen Land bewegt sich etwas, spät und bezeichnend. Deutschland verfehlte die europäische Frist, seine Aufsichtsbehörden zu benennen, deutlich: Erst im Juni 2026, zehn Monate nach der Frist, beschloss der Bundestag das nationale Durchführungsgesetz, das die Bundesnetzagentur zur zentralen Aufsicht macht, neben den Fachbehörden für Finanzen, Medizinprodukte und IT-Sicherheit.[9] Wer in Deutschland mit KI Ernst macht, wird es also künftig mit der Bundesnetzagentur als erster Adresse zu tun haben.

Datenrecht im Betrieb

Kapitel 9 hat die Souveränität zur Vorstandsfrage gemacht: wo die Daten liegen und unter wessen Hoheit. Im Betrieb kommt die zweite Frage hinzu: Was dürfen Sie mit ihnen tun? Hier setzt die Datenschutz-Grundverordnung enge Grenzen, gerade dort, wo

das Daten-Schwungrad aus Kapitel 9 am verlockendsten ist. Daten, die zu einem Zweck erhoben wurden, dürfen nicht ohne Weiteres zu einem anderen, etwa dem Training eines Modells, verwendet werden; die Zweckbindung verlangt eine eigene Rechtsgrundlage.[10] Besondere Vorsicht gilt, wo eine KI über Menschen entscheidet. Eine Entscheidung mit erheblicher Wirkung, etwa eine Kreditabsage oder eine Bewerbervorauswahl, darf nicht allein der Maschine überlassen bleiben; es braucht eine echte menschliche Prüfung, und der Europäische Gerichtshof hat klargestellt, dass schon das bloße Erzeugen eines Score darunterfällt, wenn andere maßgeblich darauf bauen.[10] Für solche Anwendungen ist in der Regel eine Datenschutz-Folgenabschätzung Pflicht, und wer ein KI-Werkzeug aus der Cloud nutzt, braucht einen Auftragsverarbeitungsvertrag mit dem Anbieter — und muss prüfen, ob der die Daten nicht seinerseits verwendet, etwa für das Training; dann trägt kein Auftragsverarbeitungsvertrag. Die deutschen Aufsichtsbehörden haben das in einer gemeinsamen Orientierungshilfe zusammengefasst, die jedem Einstieg vorausgehen sollte.[10]

Wer geradesteht, wenn es schiefgeht

Eine Frage entscheidet, ob Governance ernst genommen wird: die Haftung. Hier verschärft sich die Rechtslage gerade. Die neue europäische Produkthaftungs-Richtlinie — bis Ende 2026 in deutsches Recht umzusetzen — zählt Software und KI ausdrücklich zu den Produkten, für deren Fehler verschuldensunabhängig gehaftet wird, und sie erleichtert Geschädigten die Beweisführung in technisch undurchsichtigen Fällen.[11] Eine eigene KI-Haftungs-Richtlinie, die darüber hinausgehen sollte, hat die Kommission dagegen Ende 2025 zurückgezogen; es bleibt beim allgemeinen Recht.[11] Für die Unternehmensführung ist der Kern einfacher, als das Klein-Klein vermuten lässt. Das Aktien- und das GmbH-Recht verlangen von Vorstand und Geschäftsführung ein wirksames System zur Früherkennung von Risiken und

eine Entscheidung auf angemessener Informationsgrundlage.[12] Eine KI, die unbeaufsichtigt Schaden anrichtet, ist genau das Risiko, das ein solches System erfassen muss. Governance lässt sich an die IT-Abteilung delegieren, die Verantwortung dafür nicht. Sie bleibt, wo sie immer war: an der Spitze.

Der deutsche Befund: die Aktenkultur als Aktivposten

Für den deutschsprachigen Raum trägt dieses Kapitel eine ungewohnt gute Nachricht. Vieles, was die KI-Governance verlangt, beherrschen deutsche Unternehmen seit jeher: das geordnete Verfahren, die Aktenführung, die nachvollziehbare Zuständigkeit, das Prüfen und Zertifizieren, das eine ganze Industrie von Sachverständigen trägt. Was in Kapitel 11 als Bremse erschien, die Vorliebe für das verbriefte, dokumentierte Vorgehen, wird hier zum Aktivposten. Wer ohnehin alles protokolliert und Zuständigkeiten verbrieft, dem fällt die prüffeste KI-Steuerung leichter als einer Kultur, die auf Tempo und Improvisation setzt. Eine Zertifizierung nach ISO 42001 oder ein Siegel eines Prüflabors kann so vom Kostenposten zum Vertrauenssignal werden, das im Markt etwas wert ist.

Zwei Dinge gehören zur Ehrlichkeit dazu. Erstens ist dieselbe Gründlichkeit, die hilft, auch eine Gefahr: die der Übererfüllung. Wer jede Anwendung behandelt, als wäre sie hochriskant, erstickt den Umbau in Formularen, ehe er begonnen hat. Die Kunst ist, die Strenge dort einzusetzen, wo das Risiko sitzt, und es anderswo leicht zu halten. Zweitens ist Gründlichkeit nicht dasselbe wie Geschwindigkeit, und Deutschland war beim eigenen KI-Gesetz das langsamste Glied. Der Vorsprung entsteht dort, wo ein Haus seine vorhandene Ordnungsstärke in zügiges, verhältnismäßiges Handeln übersetzt.

Stolperfalle — Die Aufsicht, die keine ist Der häufigste Selbstbetrug der Governance ist die Aufsicht, die nur auf dem Papier steht. Ein Mensch, der einen KI-Vorschlag abnickt, ohne ihn wirklich prüfen zu können oder zu dürfen, ist keine Aufsicht, sondern ein Feigenblatt, rechtlich wie praktisch. Echte Aufsicht verlangt, dass die prüfende Person die Sache versteht, die Zeit hat und die Befugnis, Nein zu sagen. Fragen Sie bei jeder „menschlichen Kontrolle" in Ihren Abläufen: Könnte dieser Mensch die Maschine im Zweifel wirklich stoppen — und tut er es je?

Was dieses Kapitel für Sie ändert

Governance ist die Bedingung des Umbaus. Eine Organisation, die Entscheidungen an die Front und an Maschinen verlagert (Kapitel 11), kann das nur verantworten, wenn sie zugleich die Aufsicht aufbaut, die hier beschrieben ist: abgestufte Autonomie und geringste Rechte, laufendes Prüfen, ein Risikoregister mit einem Verantwortlichen je Eintrag, eine nüchterne Einordnung unter den EU AI Act, ein sauberer Umgang mit den Daten, und das Wissen, dass die Haftung an der Spitze bleibt. Nichts davon verlangt, auf das perfekte Regelwerk zu warten. Das meiste davon ist schlicht die Sorgfalt, die ein ernsthaftes Unternehmen ohnehin kennt, übertragen auf eine neue Art von Werkzeug.

Damit stehen zwei der drei Stücke des Umbaus: die Struktur und ihre Steuerung. Es fehlt das dritte und schwerste, weil es nicht von Kästchen und Regeln handelt, sondern von Menschen: von Rollen, die sich verschieben, von Können, das neu erworben sein will, von der Mitbestimmung und von der Frage, was mit den Fähigkeiten geschieht, die wir an die Maschine abgeben. Ihm gehört das nächste Kapitel.

Das Wichtigste in 60 Sekunden

- **Governance ist die Bedingung des Umbaus, nicht seine Bremse:** Nur wer beaufsichtigen, prüfen und stoppen kann, darf einer KI etwas anvertrauen. Die Lücke klafft zwischen dem, was Systeme dürfen, und dem, was Sie über sie steuern (nur 23 % haben Regeln).
- **Abgestufte Autonomie und geringste Rechte:** Ein System verdient sich Rechte an der eigenen Aufgabe; vor unumkehrbaren Aktionen gehört eine Freigabe. So begrenzt man auch die Angriffsfläche (Prompt-Injection aus Kapitel 5).
- **Risikoregister mit einem Verantwortlichen je Eintrag.** Was niemandem gehört, wird nicht behandelt.
- **EU AI Act praktisch:** Die meisten Anwendungen sind nicht hochriskant. Als Betreiber haben Sie eine kürzere Pflichtenliste als der Anbieter. Der Zeitplan ist in Bewegung (Digital Omnibus) — Stand prüfen, aber nicht auf die letzte Frist warten.
- **Haftung bleibt Chefsache.** Software und KI fallen jetzt unter die Produkthaftung; das Aktienrecht verlangt ein Risikofrüherkennungssystem. Delegieren lässt sich die Arbeit, nicht die Verantwortung.

Merksatz Vertrauen in eine Maschine ist kein Glaube, sondern eine Befugnis, die man stuft, prüft und jederzeit zurücknehmen kann.

Drei Fragen an Ihr Unternehmen

1. Für Ihren autonomsten KI-Einsatz: Wer dürfte und könnte ihn im Zweifel wirklich stoppen — und woran würde diese Person überhaupt merken, dass etwas schiefläuft?
2. Welche Ihrer KI-Anwendungen sind nach dem EU AI Act tatsächlich hochriskant — und haben Sie diese Einordnung je nüchtern vorgenommen, statt zu raten?
3. Wo verwenden Sie Daten, die zu einem anderen Zweck erhoben wurden, für KI — und auf welcher Rechtsgrundlage?

Quellen

1. Bitkom, *Künstliche Intelligenz in Deutschland* (Studienbericht Februar 2026, Daten 2025; repräsentative Befragung von 604 Unternehmen ab 20 Beschäftigten): 23 % haben feste Regeln für den Einsatz generativer KI, 31 % planen solche, 36 % haben sich damit noch nicht befasst, 16 % schließen Regeln aus. Selbstauskunft; „Regeln" ist nicht einheitlich definiert. Werte vor Druck gegen die jeweils aktuelle Welle prüfen.
2. „EchoLeak" (CVE-2025-32711), in Kapitel 5 ausführlich behandelt: eine Zero-Click-Lücke in Microsoft 365 Copilot, durch die ein Angreifer per präparierter E-Mail Daten abfließen lassen konnte, ohne Zutun des Nutzers; entdeckt von Aim Labs, von Microsoft serverseitig behoben (Mai 2025), kein bekannter Missbrauch in freier Wildbahn. Microsoft stufte die Lücke als kritisch ein (CVSS 9,3); die NVD bewertete sie mit 7,5.
3. OWASP, *Top 10 for LLM Applications* (Ausgabe 2025) und *Agentic AI — Threats and Mitigations* (v1.0, Februar 2025): systematische Auflistung der typischen Schwachstellen von KI-Anwendungen und Agenten, darunter Prompt-Injection und „Excessive Agency" (zu weit gefasste Handlungsrechte). OWASP ist eine gemeinnützige Sicherheits-Initiative; die Listen sind Branchenkonsens, kein Gesetz. (Eine spätere, separate „Top 10 for Agentic Applications" von Ende 2025 nicht mit der hier genannten Agenten-Liste verwechseln.)
4. NIST (US-amerikanisches National Institute of Standards and Technology), *AI Risk Management Framework* 1.0 (2023) mit den vier Funktionen Govern, Map, Measure, Manage, ergänzt um das *Generative AI Profile* (NIST AI 600-1, 2024). Freiwilliges Rahmenwerk, keine Zertifizierung; als Strukturierungshilfe, nicht als EU-Rechtspflicht zu verstehen.
5. ISO/IEC 42001:2023 — internationale Norm für ein KI-Managementsystem; nach ihr ist eine akkreditierte Zertifizierung möglich. Ergänzend ISO/IEC 23894:2023 (Leitfaden KI-Risikomanagement, nicht zertifizierbar). Der hier skizzierte „Risikoregister"-Aufbau ist eine redaktionelle Synthese aus diesen Normen und dem NIST-Rahmenwerk, kein eigenständiger Standard.
6. Verordnung (EU) 2024/1689 (EU AI Act / KI-VO), in Kraft seit 1. August 2024: risikobasierter Ansatz mit verbotenen Praktiken (Art. 5), Hochrisiko-Systemen (Anhang III sowie Anhang I), Transparenzpflichten für bestimmte Systeme (Art. 50) und einem nicht regulierten Rest mit minimalem Risiko. Artikel 4 (KI-Kompetenz) gilt seit dem 2. Februar 2025 für Anbieter wie Betreiber aller KI-Systeme: Beschäftigte, die mit KI arbeiten, müssen über hinreichende KI-Kompetenz verfügen; eine Sanktionsnorm eigens hierfür fehlt, die Pflicht gilt gleichwohl. *(Vor Druck: prüfen, ob der Digital Omnibus Art. 4 unverändert lässt.)* Quelle: eur-lex.europa.eu/eli/reg/2024/1689; digital-strategy.ec.europa.eu.
7. EU AI Act: Betreiberpflichten für Hochrisiko-Systeme in Artikel 26 (Einsatz nach Anleitung, kompetente menschliche Aufsicht, Überwachung und Meldung schwerer Vorfälle, Aufbewahrung der Protokolle für mindestens sechs Monate, Information der Betroffenen); menschliche Aufsicht in Artikel 14; die „Anbieter-Falle" in Artikel 25 (wer umbenennt, wesentlich ändert oder

den Zweck ändert, wird selbst Anbieter); Bußgeldrahmen in Artikel 99 (bis 35 Mio. € / 7 % bzw. 15 Mio. € / 3 %; für KMU jeweils der niedrigere Betrag). Die schweren technischen Pflichten (Art. 9–15) treffen primär den Anbieter.

8. „Digital Omnibus": Vorschlag der Europäischen Kommission vom 19. November 2025, der die Anwendung der Hochrisiko-Pflichten an die Verfügbarkeit von Normen koppelt und mit Rückfalldaten versieht (Anhang III voraussichtlich 2. Dezember 2027, Anhang I 2. August 2028). Stand Mitte 2026 bestand eine vorläufige Einigung (7. Mai 2026), aber noch keine förmliche Verabschiedung; bis zur Veröffentlichung im Amtsblatt galten die ursprünglichen Fristen (2. August 2026 / 2027). **Zeitkritisch — vor Drucklegung den aktuellen Stand prüfen.** Quellen: Europäisches Parlament, Legislative Train; Kanzlei-Analysen (White & Case, Gibson Dunn).
9. Deutsches Durchführungsgesetz zum EU AI Act (KI-Marktüberwachungs- und Innovationsförderungsgesetz): vom Bundestag im Juni 2026 beschlossen; die Bundesnetzagentur wird zentrale Marktüberwachungsbehörde, neben Fachbehörden (u. a. BaFin für Finanzen, BfArM für Medizinprodukte, BSI für IT-Sicherheit); der Bundesdatenschutzbeauftragte behält die Datenschutzaufsicht. Deutschland verfehlte die EU-Frist zur Behördenbenennung (2. August 2025) um rund zehn Monate (bezogen auf den Bundestagsbeschluss; das Inkrafttreten stand Mitte 2026 noch aus). **Stand zeitkritisch — Bundesratspassage und Inkrafttreten vor Druck prüfen.** Quelle: heise online (Juni 2026).
10. Datenschutz-Grundverordnung in der KI-Anwendung: Zweckbindung (Art. 5 Abs. 1 lit. b) begrenzt die Weiterverwendung erhobener Daten, etwa fürs Training; das Verbot rein automatisierter Einzelentscheidungen mit erheblicher Wirkung (Art. 22), vom Europäischen Gerichtshof im Urteil SCHUFA (C-634/21, 7. Dezember 2023) auch auf das bloße Erzeugen eines Score erstreckt; die Pflicht zur Datenschutz-Folgenabschätzung (Art. 35) bei Profiling als Entscheidungsgrundlage; der Auftragsverarbeitungsvertrag (Art. 28) bei Nutzung von Cloud-Diensten. Zusammengefasst in: Datenschutzkonferenz (DSK), *Orientierungshilfe — KI und Datenschutz*, Version 1.0, 6. Mai 2024.
11. Richtlinie (EU) 2024/2853 über die Haftung für fehlerhafte Produkte (in Kraft seit Dezember 2024, Umsetzung bis 9. Dezember 2026): bezieht Software und KI ausdrücklich in die verschuldensunabhängige Produkthaftung ein und erleichtert die Beweisführung. Die vorgeschlagene KI-Haftungs-Richtlinie (AILD) wurde von der Europäischen Kommission zurückgezogen (Rücknahme im Amtsblatt, 6. Oktober 2025).
12. §§ 91, 93 Aktiengesetz: Pflicht des Vorstands zu einem Überwachungssystem für bestandsgefährdende Risiken (§ 91 Abs. 2), zu angemessenem internen Kontroll- und Risikomanagementsystem bei börsennotierten Gesellschaften (§ 91 Abs. 3, seit dem FISG 2021) sowie die Sorgfaltspflicht und die Business Judgment Rule, die eine Entscheidung „auf der Grundlage angemessener Information" verlangt (§ 93 Abs. 1); für die GmbH entsprechend § 43 GmbHG. Eine höchstrichterliche Entscheidung speziell zum Versagen der KI-Governance liegt, Stand Mitte 2026, nicht vor; die Pflicht ist aus den bestehenden Normen hergeleitet.

Kapitel 13 — Der menschliche Übergang

Lernziele Nach diesem Kapitel können Sie

- erklären, warum KI das Ausführen billig macht und den Wert auf das Urteil verschiebt, und was das für die Menschen im Unternehmen heißt;
- begründen, warum dieser Wandel Stellen kostet, auch erfahrene und verdiente, und warum Dienstjahre dabei nicht schützen;
- die Entfertigung benennen: die Gefahr, das Urteil zu verlieren, das man behalten wollte;
- den Übergang ehrlich führen: umschulen, wo es trägt, fair trennen, wo nicht, und die Mitbestimmung als legitimen Weg nutzen, eine harte Last zu teilen.

Jede große Technik hat Berufe beendet und neue hervorgebracht. Rund sechzig Prozent der heute in den USA Beschäftigten arbeiten in Tätigkeiten, die es 1940 nicht gab; im selben Zeitraum sind viele alte verschwunden.[1] In der Summe ist das beruhigend: Unterm Strich ging die Arbeit nicht aus. Für den Einzelnen war der Übergang selten sanft, denn wer im alten Beruf stand, wechselte nicht von allein in den neuen. Die vom Weltwirtschaftsforum befragten Arbeitgeber erwarten für die

kommenden Jahre dasselbe Muster im Großen, Millionen neue Stellen und Millionen wegfallende zugleich.[2] Eine Führungskraft, die ihrer Belegschaft sagt, es werde sich nichts wirklich ändern, sagt die Unwahrheit.

Worin besteht das Muster bei der KI? Sie macht das Ausführen billig. Was sie an Arbeit übernimmt, ist die Ausführung: der Entwurf, die Auswertung, der erste Text, die Standardlösung. Was sie nicht mitliefert, ist das Urteil: zu erkennen, was gut ist, welches Problem zu lösen sich lohnt, was dem Kunden wirklich hilft (Kapitel 8). Damit verschiebt sich der Wert im Unternehmen: weg von denen, deren Beitrag die Ausführung war, hin zu denen, die das Urteil haben.

Wer gewinnt, wer verliert

Wie zweischneidig das ist, zeigen die Befunde. Im standardisierten Kundendienst, wo es darauf ankommt, viele Fälle auf ein bekanntes Niveau zu bringen, hob ein KI-Assistent vor allem die Unerfahrenen: Sie wurden deutlich besser, die Routiniers kaum.[3] Dort lieferte die Maschine den Anfängern die Ausführung, die ihnen noch fehlte. In der Facharbeit dürfte sich dieselbe Logik umkehren: Wer das Urteil hat (die Fachkenntnis, den Geschmack, die Kundenbeziehung) und sich der KI bedient, erledigt die Arbeit mehrerer, deren Wert in der Ausführung lag. Ob die KI also hebt oder verdrängt, hängt weder am Alter noch an den Dienstjahren, sondern daran, ob der Beitrag eines Menschen seine Ausführung war oder sein Urteil.

Der ehrliche Teil: Stellen fallen weg

Daraus folgt, was viele Bücher über KI aussparen: Der Übergang kostet Stellen, auch gute. Wo die Ausführung zur Ware wird, wird der Mensch entbehrlich, dessen Beitrag die Ausführung war: der erfahrene Spezialist, dessen Handwerk die Maschine inzwischen erreicht, ebenso wie der Berufsanfänger, dessen Einstiegsfertigkeiten nun überall zu haben sind. Dreißig Jahre Erfahrung schützen nicht, wenn diese Erfahrung in einer Ausfüh-

rung steckt, die jetzt jeder mit einem guten Modell abruft. Eine Fachkraft mit Urteil, die sich der Werkzeuge bedient, kann mehrere ersetzen, die das nicht tun.

Es ist unredlich, das als bloße Entlastung zu verkaufen. Eine Belegschaft spürt den Unterschied zwischen Entlastung und Abbau genau, und der Schaden einer beschönigten Erzählung ist größer als der einer harten, aber ehrlichen. Wer Vertrauen für den ganzen Umbau braucht, kann es sich an dieser Stelle am leichtesten verspielen.

> **Stolperfalle — Die beschönigte Erzählung** Der häufigste Fehler im menschlichen Übergang ist, Stellenabbau als reine Entlastung auszugeben: „Die KI nimmt euch nur die lästige Arbeit ab." Wo das zutrifft, gut. Wo in Wahrheit Rollen wegfallen, durchschaut die Belegschaft die Beschönigung sofort — und mit ihr ist das Vertrauen verspielt, das Sie für jeden weiteren Schritt brauchen. Sagen Sie, was gilt: welche Rollen wachsen, welche sich wandeln, welche enden.

Was der Führung zu tun bleibt

Wenn der Wandel Stellen kostet, verschiebt sich die Führungsaufgabe vom Beschönigen zum ehrlichen Gestalten. Dreierlei gehört dazu. Erstens, früh benennen, welche Rollen wachsen, welche sich wandeln und welche enden; anhaltende Ungewissheit zermürbt eine Belegschaft mehr als eine unbequeme Klarheit. Zweitens, dort umschulen, wo es trägt. Umschulung heißt nicht, jemandem ein neues Werkzeug zu zeigen, sondern ihn in eine andere Rolle zu bringen, meist eine, in der sein Urteil zählt und nicht seine alte Ausführung. Das gelingt, wo Eignung und betrieblicher Bedarf zusammenkommen, und es gelingt nicht bei jedem; das auszusprechen, gehört zur Ehrlichkeit. Drittens, fair trennen, wo Umschulung nicht trägt: mit Vorlauf, mit Anstand, mit den Mitteln, die Sozialstaat und Tarifvertrag dafür bereithalten. Die Würde des Übergangs liegt nicht darin, dass niemand geht, sondern darin, wie die gehen, die gehen müssen.

Ein vierter Punkt betrifft die, die bleiben, und er überrascht viele: Gerade die besten Leute nehmen die Werkzeuge oft am zögerlichsten an. Das ist zunächst rational, nicht störrisch — wer die Ausführung selbst beherrscht, gewinnt durch die Maschine am wenigsten; im Kundendienst-Befund oben profitierten die Routiniers ja kaum. Die Antwort darauf ist keine Werkzeugpflicht von oben, sondern eine neue Rolle: Die Erfahrensten werden zu denen, die die Ausgaben der Maschine beurteilen und die anderen darin anleiten. So wird ihr Urteil zu dem, was es sein soll — das knappe Gut aus Kapitel 8, kein Relikt.

Die Entfertigung

Für die, die bleiben, gibt es eine zweite Gefahr, und Kapitel 8 hat sie für die einzelne Führungskraft schon benannt: Wer das Tun ganz an die Maschine abgibt, verliert mit der Zeit das Urteil, das ihn wertvoll macht, weil Urteil auf Übung beruht. Was dort für den Einzelnen galt, gilt für die ganze Belegschaft. Behält ein Unternehmen seine Leute gerade wegen ihres Urteils und lässt dieses Urteil dann verkümmern, hat es am Ende weder die Ausführung, die es abgegeben hat, noch das Urteil, dessentwegen es die Menschen behalten hat.

> **Klartext — Die Entfertigung** *Entfertigung* meint den schleichenden Verlust von Können, wenn eine Fertigkeit dauerhaft an eine Maschine abgegeben wird. Die Luftfahrt kennt das Muster: Die amerikanische Luftaufsicht warnt ausdrücklich vor dem „Verfall der manuellen Flugfertigkeiten" bei Piloten, die fast nur noch den Autopiloten überwachen.[4] Eine Belegschaft, die KI-Vorschläge nur noch bestätigt, verlernt ebenso, sie zu prüfen.

Das Gegenmittel ist keine Technikabkehr, sondern bewusste Übung: Fälle von Hand durchdenken, auch wenn der Vorschlag schon dasteht; den menschlichen Prüfer wirklich prüfen lassen und nicht abnicken (Kapitel 12); den Ausfall des Systems gele-

gentlich proben. Resilienz heißt, die Fähigkeit zu bewahren, ohne die Maschine zu arbeiten, gerade weil man es im Alltag selten braucht.

Mitbestimmung: eine harte Last teilen

Gerade weil der Übergang hart ist, lohnt ein nüchterner Blick auf ein Instrument, das im deutschsprachigen Raum bereitsteht: die Mitbestimmung. In Kapitel 2 stand sie als Bremse, und das bleibt sie: Sie macht den Umbau langsamer. Ihr eigentlicher Wert liegt woanders. Sie ist der geregelte Weg, eine unvermeidliche Last zu verteilen. Das Betriebsverfassungsgesetz verlangt Interessenausgleich und Sozialplan, wenn ein Umbau Stellen kostet, und es hat die Beteiligung 2021 ausdrücklich auf die KI erstreckt; der europäische KI-Act verlangt zudem, die Beschäftigten vor dem Einsatz eines Hochrisiko-Systems zu informieren.[5] Wer den Betriebsrat erst vor die fertige Entscheidung setzt, erntet Widerstand; wer ihn früh einbezieht, teilt eine schwere Entscheidung mit denen, die sie tragen. Das macht den Abbau nicht angenehm, aber legitim. Und an fehlender Legitimität zerbrechen Umbauten am ehesten.

Der deutsche Sonderfall, nüchtern betrachtet

Zwei Trostgründe, die man dem deutschsprachigen Raum gern zuspricht, verdienen eine ehrliche Prüfung. Der erste ist die Demografie. Dem Land gehen die Arbeitskräfte aus; 2024 blieben rund eine halbe Million Fachkraftstellen unbesetzt, und das Erwerbspersonenpotenzial schrumpft über die kommenden Jahre weiter.[6] Das mildert die gesamtwirtschaftliche Härte: Wer fähig ist und gehen muss, findet auf einem leergefegten Arbeitsmarkt eher etwas Neues. Es nimmt der einzelnen Firma aber nicht die Entscheidung ab, eine überflüssig gewordene Rolle zu beenden, und es erspart dem Einzelnen den harten Wechsel nicht. Die Demografie ist ein Trost für die Volkswirtschaft, kein Freibrief für das Unternehmen.

Der zweite vermeintliche Vorteil ist das duale Ausbildungssystem. Anders als die in Kapitel 8 gewürdigte Meister-Tradition, die das knappe Urteil schult, bildet die breite duale Ausbildung vor allem für standardisierte fachliche Ausführung aus, und genau die macht die KI billig. Eine Wirtschaft, die stark auf solche Ausführung baut, steht dem Wandel deshalb eher exponierter gegenüber als geschützter.[7] Zum Aktivposten wird die duale Ausbildung erst, wenn sie das mitvermittelt, was knapp bleibt: Urteil, Verantwortung, die geübte Arbeit mit der Maschine. Darauf ist sie heute kaum eingerichtet, und darin liegt eine stille Standortaufgabe, die größer ist als jedes einzelne KI-Projekt.

Was dieses Kapitel für Sie ändert

Der menschliche Übergang ist das Stück des Umbaus, das sich weder anordnen noch automatisieren lässt. Ihn gut zu führen heißt, die Ehrlichkeit der Beschönigung vorzuziehen: zu benennen, was endet; zu stärken, was knapp ist, nämlich das Urteil; umzuschulen, wo es trägt, und fair zu trennen, wo nicht; und das Urteil, das Sie behalten, vor dem eigenen Verfall zu schützen. Das ist kein bequemer Weg, aber ein gangbarer, und er verlangt genau die Führung, von der dieses Buch handelt. Wer die Menschen ehrlich durch den Übergang führt, gewinnt am Ende mehr Vertrauen, als der gewinnt, der ihnen einen schmerzlosen Wandel verspricht, den es nicht gibt.

Damit stehen die drei Stücke des Umbaus beisammen: die Organisation, ihre Steuerung und die Menschen. Was bleibt, ist, das Ganze in eine Reihenfolge zu bringen — in einen Fahrplan, der sagt, was am Montag zu tun ist. Ihm gehört das letzte Kapitel.

Das Wichtigste in 60 Sekunden

- KI macht das **Ausführen** billig und verschiebt den Wert auf das **Urteil**. Wer gebraucht wird, entscheidet sich daran, ob sein Beitrag Ausführung oder Urteil war — nicht an den Dienstjahren.
- **Der Übergang kostet Stellen, auch gute.** Erfahrene Spezialisten, deren Handwerk die Maschine erreicht, und Anfänger, deren Einstiegsfertigkeiten überall zu haben sind, sind gleichermaßen gefährdet. Das als bloße Entlastung zu verkaufen, kostet das Vertrauen der Belegschaft.
- **Führen statt beschönigen:** früh benennen, was endet; umschulen (in eine neue Rolle, nicht in ein neues Werkzeug), wo Eignung und Bedarf zusammenkommen; fair trennen, wo nicht.
- **Entfertigung:** Wer das Tun ganz an die Maschine abgibt, verliert mit der Zeit das Urteil, das ihn wertvoll macht. Resilienz heißt, das Können zu üben, auch wenn die Maschine es übernehmen könnte.
- **Mitbestimmung** ist der legitime Weg, eine harte Last zu teilen — früh genutzt, nicht erst zur Abnahme. Die Demografie mildert die Härte für die Volkswirtschaft, nicht für die einzelne Firma; das duale System schützt nicht, solange es Ausführung lehrt statt Urteil.

Merksatz Die KI macht das Ausführen billig. Gebraucht wird, wer das Urteil hat — nicht, wer am längsten dabei ist.

Drei Fragen an Ihr Unternehmen

1. Bei welchen Rollen steckt der Wert noch in der Ausführung, die die KI gerade billig macht — und haben Sie das den Betroffenen gegenüber ehrlich benannt?
2. Wen können Sie ernsthaft in eine Rolle umschulen, in der sein Urteil zählt — und wo wäre ein fairer Abschied redlicher als ein Scheinangebot?
3. Welche Ihrer besten Köpfe drohen ihr Urteil zu verlieren, weil sie das Tun vollständig an die Maschine abgegeben haben?

Quellen

1. David Autor, Caroline Chin, Anna Salomons & Bryan Seegmiller, *New Frontiers: The Origins and Content of New Work, 1940–2018*, Quarterly Journal of Economics 139(3), 2024: rund 60 % der Beschäftigung von 2018 entfielen auf Tätigkeiten, die es 1940 nicht gab. US-Daten. Der Befund zeigt zugleich Schaffung und Verschwinden von Tätigkeiten; die historische Gesamtbeschäftigung blieb erhalten, der Übergang traf die Betroffenen aber sehr ungleich. Zur Aufgabenlogik (Verdrängungs- und Wiedereinsetzungseffekt): Acemoglu & Restrepo, *Automation and New Tasks*, Journal of Economic Perspectives 33(2), 2019.
2. World Economic Forum, *The Future of Jobs Report 2025* (Januar 2025): Arbeitgeber erwarten bis 2030 weltweit rund 170 Millionen neue und 92 Millionen wegfallende Stellen, netto rund 78 Millionen zusätzliche. Es handelt sich um eine Erwartungsumfrage unter Arbeitgebern, nicht um gemessene Ergebnisse; die Bruttozahl der wegfallenden Stellen ist hier so wichtig wie die Nettozahl.
3. Brynjolfsson, Li & Raymond, *Generative AI at Work*, Quarterly Journal of Economics 140(2), 2025 (Arbeitspapier NBER 31161, 2023): in einem Kundendienst mit gut 5.000 Beschäftigten stieg die Produktivität durch einen KI-Assistenten im Schnitt um rund 15 %, mit den größten Zuwächsen bei den weniger erfahrenen Beschäftigten und kaum Wirkung bei den erfahrensten. Feldstudie an standardisierter Dienstleistungsarbeit; sie belegt die eine Richtung der Asymmetrie (die Maschine liefert den Unerfahrenen die fehlende Ausführung). Die Gegenrichtung in der Facharbeit — Urteilsträger plus KI ersetzen reine Ausführende — ist die begründete Kehrseite desselben Mechanismus, nicht ein gemessenes Ergebnis dieser Studie.
4. Zur Entfertigung in der Luftfahrt: US-amerikanische Luftfahrtbehörde (FAA), *Safety Alert for Operators* 13002 (2013) und der Bericht der PARC/CAST-Arbeitsgruppe *Operational Use of Flight Path Management Systems* (2013): dokumentierter „Verfall der manuellen Flugfertigkeiten" durch Überabhängigkeit von der Automatik. Das Grundmuster — wer das Tun abgibt, verliert das Können — geht auf Lisanne Bainbridge, *Ironies of Automation* (1983, in Kapitel 8 eingeführt), zurück; vgl. Nicholas Carr, *The Glass Cage* (2014).
5. Betriebsverfassungsgesetz: Pflicht zu Interessenausgleich und Sozialplan bei Betriebsänderungen (§§ 111, 112); Beteiligung des Betriebsrats bei der Gestaltung von Arbeitsabläufen „einschließlich des Einsatzes von Künstlicher Intelligenz" (§ 90), erleichterte Hinzuziehung eines Sachverständigen zur Beurteilung von KI (§ 80 Abs. 3) und Mitbestimmung bei Auswahlrichtlinien, auch wenn KI sie aufstellt (§ 95 Abs. 2a), eingeführt durch das Betriebsrätemodernisierungsgesetz 2021. Ergänzend Verordnung (EU) 2024/1689 (EU AI Act), Artikel 26 Abs. 7: Information der Beschäftigtenvertretung vor dem Einsatz eines Hochrisiko-Systems am Arbeitsplatz.
6. Institut der deutschen Wirtschaft / KOFA, Jahresrückblick 2024 (2025): rund 487.000 offene Stellen für qualifizierte Fachkräfte blieben rechnerisch unbesetzt, weil bundesweit keine passend Qualifizierten verfügbar waren. Institut für Arbeitsmarkt- und Berufsforschung (IAB): Das Erwerbspersonenpo-

tenzial sinkt bei realistischer Zuwanderung bis 2035 um etwa 2,9 Millionen, ohne Zuwanderung um rund 7,2 Millionen (rechnerisches Kontrafaktum); stark zuwanderungsabhängig.

7. Zur dualen Ausbildung: 328 staatlich anerkannte Ausbildungsberufe (BIBB, 2024), überwiegend auf standardisierte fachliche Ausführung ausgerichtet. Die Einschätzung, dass eine stark auf solche Ausführung gebaute Wirtschaft der KI-Automatisierung eher exponierter gegenübersteht, ist eine begründete eigene Position dieses Buches, gestützt auf die Aufgabenlogik der Automatisierung (Anm. 1): Was regelhaft und standardisiert ist, wird zuerst billig.

Kapitel 14 — Der Fahrplan

Lernziele Nach diesem Kapitel können Sie

- die Kernaussage des Buches und sein Gerüst aus dem Stand wiedergeben und auf Ihr Haus beziehen;
- mit einer Selbstdiagnose ehrlich bestimmen, wo Ihr Unternehmen bei den drei Knappheiten und beim Umbau steht;
- die ersten neunzig Tage planen, ohne in ein großes Programm oder in die Pilotitis zu verfallen;
- den einen kleinen Schritt benennen, mit dem Sie am Montag beginnen.

Am Ende eines solchen Buches ist die Versuchung groß, ein großes Programm aufzusetzen: eine KI-Strategie, eine Taskforce, ein Budget mit vielen Nullen. Die Erfahrung aus Kapitel 2 rät davon ab. Große Pläne und beeindruckende Pilotprojekte scheitern auf dieselbe Weise; nur etwa sechs von hundert Unternehmen führen einen spürbaren Ergebnisbeitrag auf ihre KI zurück.[1] Was die wenigen von der Mehrheit trennt, ist nicht der größere Plan, sondern der eine Arbeitsablauf, der wirklich umgebaut wurde. Dieser letzte Teil macht aus dem Gelesenen eine Reihenfolge: zuerst ehrlich bestimmen, wo Sie stehen, dann die ersten neunzig Tage, dann der Schritt für Montag.

Das Buch in einem Absatz

Auf den Punkt gebracht: KI alleine verschafft noch keinen Vorsprung. Den Vorsprung schafft erst die Führung, die daraus etwas macht. Das Werkzeug ist für jeden käuflich; knapp bleibt, was sich nicht kaufen lässt. Drei solche knappen Güter benennt dieses Buch. Das Urteil, gute von schlechter Arbeit zu unterscheiden (Kapitel 8). Den Kontext, die eigenen Daten und nachprüfbaren Rückkopplungen, an denen die Maschine im eigenen Haus besser wird als bei jedem anderen (Kapitel 9). Und den Wert je Recheneinheit, die Disziplin, aus Rechenleistung messbaren Geschäftswert zu ziehen, statt Budget zu verbrennen (Kapitel 10, Return on Tokens). Zur Rendite werden diese drei erst durch den Umbau: eine Organisation, die Entscheidungen dorthin verlegt, wo das Wissen sitzt (Kapitel 11); eine Governance, die KI beaufsichtigt und verantwortet (Kapitel 12); und ein menschlicher Übergang, der ehrlich geführt sein will (Kapitel 13). Wer dieses Gerüst im Kopf hat, kann sich selbst prüfen.

Die Selbstdiagnose

Die folgende Diagnose ist zum Teilen gedacht. Legen Sie sie Ihrem Führungskreis vor, jeder antwortet zuerst für sich, dann vergleichen Sie. Antworten Sie je Frage mit Ja, Teils oder Nein, und werten Sie ein Weiß-nicht als Nein. Es geht weniger um eine Punktzahl als um die Stellen, an denen die Antworten auseinanderfallen oder unangenehm ausfallen. Zu jeder der sechs Fragen gehört eine Gegenprobe: das, was eine beschönigte Antwort verrät.

Urteil. Können Ihre Leute an einer konkreten Aufgabe aus dem eigenen Haus sagen, ob eine KI-Ausgabe gut oder schlecht ist, und woran sie das festmachen? *Gegenprobe*: Wenn der einzige Beleg eine beeindruckende Vorführung ist, prüfen Sie nicht, Sie glauben — die Demo-Falle aus Kapitel 6.

Kontext. Welche Ihrer Daten sind wirklich einzigartig und für eine wertvolle Aufgabe relevant, und fließt aus ihrer Nutzung eine Rückmeldung zurück, die das System besser macht?

Gegenprobe: Ein Data Lake, den niemand ordnet und prüft, verursacht Kosten und schafft keinen Vorsprung (Kapitel 9).

Wert je Recheneinheit. Kennen Sie für Ihre wichtigste KI-Anwendung eine Zahl je Vorgang, etwa die Kosten je gelöstem Fall, oder nur die Gesamtrechnung am Monatsende? *Gegenprobe*: Wer allein den Token-Preis drückt, optimiert das Falsche; der Wert entsteht im Zähler, beim Nutzen (Kapitel 10).

Organisation. Haben Sie zuletzt einen Arbeitsablauf um die KI herum neu gebaut und dabei eine Entscheidung näher an die Front verlegt, oder die KI an einen unveränderten Ablauf geheftet? *Gegenprobe*: Ist keine Entscheidung gewandert, war es Anbau, nicht Umbau (Kapitel 11).

Governance. Können Sie Ihr autonomstes System im Ernstfall anhalten, und steht für jedes Ergebnis ein Mensch gerade? *Gegenprobe*: Eine menschliche Kontrolle, die nur abnickt, ist keine (Kapitel 12).

Menschen. Haben Sie ehrlich benannt, welche Rollen die KI wachsen lässt, welche sie wandelt und welche sie beendet? *Gegenprobe*: Wer Stellenabbau als bloße Entlastung verkauft, verliert das Vertrauen, das den Umbau trägt (Kapitel 13).

Die wertvollsten Erkenntnisse stecken in den Fragen, bei denen Ihr Kreis zögert oder streitet. Dort liegt Ihre Arbeit. Nehmen Sie die Frage, bei der die Antworten am weitesten auseinanderlagen, mit in die folgenden neunzig Tage: Sie sagt Ihnen, was der erste Fall beweisen muss.

Die ersten neunzig Tage

Die ersten neunzig Tage haben ein bescheidenes Ziel: an einem einzigen Fall zu lernen, wie Umbau in Ihrem Haus geht. Nicht alles auf einmal, ein Fall.

Wochen eins und zwei: wählen. Suchen Sie einen Arbeitsablauf, der häufig vorkommt, lästig genug ist, um die Mühe zu lohnen, und dessen Ergebnis sich messen lässt. Benennen Sie einen Verantwortlichen für das betriebliche Ergebnis, nicht für die Vorführung, und die eine Zahl, an der Sie in drei Monaten Erfolg oder Misserfolg ablesen (Kapitel 2 und 11).

Wochen drei bis acht: bauen. Bauen Sie diesen Ablauf um die Frage herum, wo das Wissen sitzt und wer entscheidet, statt die KI an die alte Form zu heften. Beginnen Sie mit dem billigsten Modell, das die Prüfung an Ihrer eigenen Beispielsammlung besteht (Kapitel 6 und 10). Klären Sie früh, welche Daten das System sehen darf und auf welcher Rechtsgrundlage (Kapitel 12). Wo Arbeitsplätze berührt sind, holen Sie die Mitbestimmung von Anfang an dazu (Kapitel 13).

Wochen neun bis zwölf: messen und entscheiden. Messen Sie den Wert je Vorgang, nicht die Menge der Ausgaben. Entscheiden Sie dann nüchtern, ob Sie verbreitern, nachbessern oder beenden. Ein ehrliches Beenden ist billiger als jede Fortsetzung des Falschen — und es lehrt Sie mehr über den zweiten Fall als ein geschöntes Weiterlaufen. Was bleibt, tragen Sie in ein Risikoregister ein, mit einem Verantwortlichen je Eintrag (Kapitel 12). Erst wenn dieser eine Fall trägt, nehmen Sie den nächsten.

> **Stolperfalle — Das große Programm** Der teuerste Fehler des Anfangs ist, zu groß zu beginnen: eine unternehmensweite KI-Initiative, ausgerufen im Vorstand, mit Dutzenden Vorhaben gleichzeitig. Sie bindet Aufmerksamkeit, erzeugt Vorführungen und erreicht selten den Wirkbetrieb. Ein einziger umgebauter Ablauf, der nachweislich Wert schafft, überzeugt die Organisation mehr als jede Strategiefolie, und er lehrt Sie, was im zweiten Fall anders laufen muss.

Für den deutschsprachigen Raum passt dieser bescheidene Weg besonders gut. Gerade der kleinere Betrieb, dem es an Zeit und IT-Leuten fehlt, kommt mit einem einzigen, sauber umgebauten Ablauf weiter als mit jedem großen Programm. Und der Standort als Ganzes gewinnt weniger durch ein nationales Leuchtturmprojekt als durch viele solche stillen Umbauten in tausenden Betrieben, die ihren Datenschatz heben und ihre Abläufe neu denken. Der Rückstand bei der Verbreitung, der diesen Standort bremst (Kapitel 1 und 2), schließt sich Ablauf für Ablauf.

Was am Montag zu tun ist

Neunzig Tage beginnen mit einem Montag, und der verlangt weniger, als man denkt. Tun Sie an diesem ersten Tag nur eines: Finden Sie heraus, wo in Ihrem Haus die Leute KI bereits benutzen, ohne dass es jemand angeordnet hat. Diese stille, inoffizielle Nutzung zeigt Ihnen genauer als jede Strategiesitzung, wo die Arbeit hakt und welcher Ablauf sich als Erster zu bauen lohnt. Sie müssen dafür nichts kaufen und niemanden einstellen. Sie müssen hinsehen und zuhören.

Ein letztes Wort

Dieses Buch ist mit KI geschrieben und von einem Menschen verantwortet, als gelebtes Beispiel dessen, was es lehrt. Seine Quellen sind angegeben, damit Sie nachprüfen können; nehmen Sie nichts unbesehen, auch nicht von hier. Genau das ist die Haltung, auf die es ankommt: weder dem Versprechen noch der Abwehr zu glauben, sondern selbst zu prüfen und dann zu handeln. Es ist die Haltung des nüchternen Frühanwenders aus Kapitel 1.

Den Vorsprung, von dem dieses Buch handelt, kauft man nicht. Man führt ihn herbei, mit Urteil, mit dem eigenen Kontext, mit Disziplin im Wert je Recheneinheit und mit dem Mut, das Unternehmen darum herum umzubauen. Das ist mehr Arbeit als ein Einkauf und weniger Zauber als ein Versprechen. Es ist die Arbeit, die übrig bleibt, wenn die Technik für alle dieselbe ist. Fangen Sie am Montag an.

Das Wichtigste in 60 Sekunden

- Die Kernaussage: KI alleine verschafft keinen Vorsprung; den schafft die Führung, die daraus etwas macht. Knapp sind Urteil, Kontext und Wert je Recheneinheit; zur Rendite werden sie erst durch den Umbau.
- Kein großes Programm. Große Pläne und Pilotprojekte scheitern gleichermaßen; was trägt, ist ein einziger umgebauter Ablauf mit einem Verantwortlichen und einer Zahl.
- Diagnostizieren Sie ehrlich entlang der sechs Dimensionen, und achten Sie auf die unangenehmen Antworten, nicht auf die Punktzahl.
- Die ersten neunzig Tage: wählen, bauen, messen und entscheiden — verbreitern, nachbessern oder beenden.
- Am Montag genügt ein Schritt: herausfinden, wo Ihre Leute KI längst nutzen, und dort den ersten Ablauf suchen.

Merksatz Den Vorsprung kauft man nicht, man führt ihn herbei — einen Ablauf nach dem anderen, nicht ein Programm auf einmal.

Quellen

1. Rückruf auf Kapitel 2. McKinsey & Company, *The State of AI* (Ausgabe 2025): rund 88 % der Unternehmen nutzen KI in mindestens einem Bereich, aber nur etwa 6 % („AI high performers") führen einen spürbaren Anteil ihres Betriebsergebnisses auf KI zurück. Die Zahl ist hier als Rückruf auf den in Kapitel 2 belegten Befund verwendet; sie steht für das Muster, nicht für eine zusätzliche Behauptung. Vor Druck mit der in Kapitel 2 zitierten Fassung abgleichen.

GLOSSAR

Quelle: konsolidiert aus den *Klartext*-Kästen der Kapitel (kanonische Definitionen, Manifest §D). Begriffe alphabetisch; die Zahl in Klammern nennt das Kapitel der Einführung.

GLOSSAR

Die wichtigsten Begriffe dieses Buches, in einem Satz erklärt — in der Reihenfolge des Alphabets. Wo ein Begriff zuerst ausführlich vorkommt, steht das Kapitel dahinter.

Abgestufte Autonomie / Geringste Rechte (Kap. 12) — *Abgestufte Autonomie:* Ein KI-System erhält nur so viel Handlungsspielraum, wie es sich an der eigenen Aufgabe nachweislich verdient hat — von „darf nur lesen und vorschlagen" über „darf mit Freigabe handeln" bis „darf in engen Grenzen selbst handeln". *Geringste Rechte:* Jedes Werkzeug eines Agenten wird so eng zugeschnitten wie möglich (das Prinzip der minimalen Berechtigung aus der IT-Sicherheit).

Agent (Kap. 5) — ein Modell, das ein Ziel verfolgt: Es plant Schritte, nutzt Werkzeuge und handelt in einer Schleife. Kurzform: Der Chatbot redet, der Agent tut.

Allzwecktechnologie (Basistechnologie) (Kap. 1) — keine Lösung für ein einzelnes Problem, sondern die Grundlage für viele; ihr Nutzen stellt sich verzögert ein, weil sich ganze Branchen erst um sie herum neu erfinden müssen (wie Dampf, Strom, Computer).

Benchmark-Verfall (Kap. 6) — öffentliche Tests verlieren mit der Zeit ihre Aussagekraft: durch Sättigung, durchgesickerte Testfragen und Goodharts Gesetz — wird eine Kennzahl zum Ziel, taugt sie nicht mehr als Kennzahl.

Drei Knappheiten, die — das Gerüst dieses Buches: **Urteil** (Kap. 8), **Kontext** (Kap. 9) und **Wert je Recheneinheit / Return on Tokens** (Kap. 10) — die drei Güter, die knapp bleiben, wenn

die Technik für alle käuflich ist. Zur Rendite werden sie erst durch den **Umbau** (Kap. 11–14).

Entfertigung (Kap. 13) — der schleichende Verlust von Können, wenn eine Fertigkeit dauerhaft an eine Maschine abgegeben wird; die unternehmensweite Form des Automatisierungsparadoxes. Sie bedroht gerade das knappe Gut Urteil.

Evaluation (Kap. 6) — das systematische Prüfen an einem vorab festgelegten Maßstab, auf anwendungsnahen Beispielen, ob ein KI-System gut genug funktioniert.

Evaluationsgetriebenes Vorgehen (Kap. 6) — den Erfolgsmaßstab festlegen, *bevor* gebaut wird; das KI-Pendant zum testgetriebenen Entwickeln.

Feinabstimmung (Kap. 4) — die Modellgewichte auf eigenen Beispielen weitertrainieren; ändert Stil, Format und Verhalten. Der aufwendigste Anpassungsweg — er bringt **Verhalten**.

Geschlossenes Modell (Kap. 7) — ein Modell in der Cloud des Anbieters, nur über eine Schnittstelle nutzbar, die Gewichte unter Verschluss (etwa GPT, Claude, Gemini).

Halluzination (Kap. 3) — eine überzeugend klingende, sachlich falsche Ausgabe; strukturelle Folge des Verfahrens (Wahrscheinliches ist nicht dasselbe wie Wahres), kein behebbarer Programmierfehler.

Inferenz (Kap. 4) — das Ausführen des fertigen Modells auf eine Anfrage; lernt nichts dazu. Sie ist die laufende, mit der Nutzung wachsende Betriebsrechnung des Anwenders — das, was Return on Tokens misst.

KI als Richter (Kap. 6) — ein Modell, das die Ausgaben eines anderen bewertet; nützlich, aber voreingenommen (für Reihenfolge, Länge, die eigene Modellfamilie) und schwankend. Zu eichen, festzuschreiben und gegen menschliche Urteile abzugleichen.

Kontext (Knappheit II) (Kap. 9) — die eigenen, vertrauenswürdigen Daten und die nachprüfbaren Rückkopplungen aus dem eigenen Geschäft; die einzige der drei Knappheiten, die ein *Vermögenswert* ist (das Modell mieten alle, den Kontext besitzt nur einer). Ein Vorsprung aber nur, wenn die Daten einzigartig,

aufgabenrelevant und durch Rückkopplung erneuert sind — nicht durch bloßes Volumen.

Kontextfenster (Kap. 3) — die maximale Textmenge (Frage, Anhänge und Antwort zusammen), die ein Modell auf einmal berücksichtigt; sein Arbeitsgedächtnis, mit Obergrenze und einer Schwäche für das, was in der Mitte steht.

Lesen vs. Handeln / Schreibrechte (Kap. 5) — die zentrale Risikolinie: Lesende Aktionen sind harmlos; schreibende (senden, ausführen, ändern, zahlen) greifen in die Welt ein.

Marginales Risiko (Kap. 7) — der richtige Maßstab in der Missbrauchsdebatte: nicht das absolute Risiko eines offenen Modells, sondern das *zusätzliche* gegenüber bereits frei verfügbaren Werkzeugen.

Nachschlagen (RAG) (Kap. 4) — dem Modell zur Laufzeit relevante Stellen aus den eigenen Daten vorlegen; erdet die Antwort, hält sie aktuell und mindert (beseitigt aber nicht) die Halluzination. Es bringt **Wissen**.

Offenes Modell (Open-Weight) (Kap. 7) — ein Modell, dessen Gewichte zum Herunterladen und Selbstbetrieb freigegeben sind (etwa Llama, Mistral, DeepSeek, Qwen). „Offen" heißt meist nur „Gewichte verfügbar", selten „quelloffen" — die Trainingsdaten bleiben verschlossen.

Organisationskapital (Kap. 2) — der unsichtbare, firmenspezifische Wert in Abläufen, Routinen, Können, Datenordnung und Struktur; er entscheidet über den Ertrag einer Technik. Nicht kaufbar, schwer zu kopieren — der eigentliche Vorsprung, den ein Wettbewerber nicht über Nacht nachbildet.

Post-Training / RLHF (Kap. 4) — der Schritt, der ein Rohmodell zum anweisungstreuen Assistenten macht (Instruktions-Tuning plus Lernen aus menschlichen Präferenzurteilen). Hier sitzen Ton, Leitplanken — und die Neigung zur Gefälligkeit.

Prompt (Kap. 4) — die Anweisung samt Beispielen; sie steuert das Modell ohne Änderung der Gewichte (Lernen im Kontext). Die billigste Form der Anpassung.

Prompt-Injection (Kap. 5) — das Einschmuggeln versteckter Anweisungen, die das Modell als Befehl ausführt; *direkt* durch

den Nutzer oder *indirekt* in fremdem Inhalt, den ein Agent liest. Strukturell, weil Anweisung und Daten denselben Kanal teilen.

Return on Tokens → siehe **Wert je Recheneinheit**.

Souveränität / Drei Betriebsarten (Kap. 7) — drei Wege, ein Modell zu betreiben: (a) geschlossenes Modell aus der Cloud, (b) offenes Modell über eine fremde Schnittstelle, (c) offenes Modell selbst betrieben. Nur (c) hält die Daten vollständig im eigenen Haus. Souveränität fragt, unter wessen Hoheit Daten und Rechenleistung liegen — nicht, in welchem Land der Server steht.

Temperatur (Kap. 3) — der Einstellregler für die Zufälligkeit der Token-Auswahl: niedrig heißt nüchtern und vorhersehbarer, hoch heißt verspielter. Erklärt, warum dieselbe Frage zwei verschiedene Antworten bekommt.

Token (Kap. 3) — der kleinste Textbaustein (meist ein Wortteil), in dem ein Modell liest, schreibt und abgerechnet wird. Faustregel: 100 Tokens entsprechen im Englischen rund 75 Wörtern; Deutsch braucht etwa 30 bis 65 Prozent mehr.

Training (Kap. 4) — der einmalige, teure Lernvorgang, in dem ein Modell aus großen Textmengen entsteht; Sache des Modellbauers, nicht des Anwenders.

Umbau, der (Kap. 11–14) — der Mechanismus, der die drei Knappheiten in Rendite verwandelt: die Organisation (Entscheidungen dorthin, wo das Wissen sitzt), die Governance (beaufsichtigen, prüfen, stoppen können) und der menschliche Übergang (ehrlich geführt). Sein Gegenteil ist der Anbau: KI an unveränderte Abläufe heften.

Urteil (Knappheit I) (Kap. 8) — die Fähigkeit, gute von schlechter Arbeit zu unterscheiden und zu entscheiden, was überhaupt gebaut gehört; knapp, weil unkäuflich und nur durch Übung und Nähe zur Sache erwerbbar. Je billiger die Erzeugung wird, desto wertvoller wird das Urteil.

Vervollständigungsmaschine (Kap. 3) — das tragende Laienmodell für ein Sprachmodell: Es erzeugt Stück für Stück das wahrscheinlichste nächste Token, statt Fakten abzurufen. Es *erzeugt* Antworten, es *weiß* sie nicht.

Werkzeug (Tool / Function Calling) (Kap. 5) — eine vom Modell bei Bedarf aufrufbare Fähigkeit: Websuche, Code-Ausführung, Datenbankabfrage, E-Mail, Schnittstelle.

Wert je Recheneinheit / Return on Tokens (Knappheit III) (Kap. 10) — der Geschäftswert, den man je verbrauchter Recheneinheit herausholt (Nutzen geteilt durch Tokens); die Disziplin, aus billiger Rechenleistung messbaren Wert zu ziehen. Ein Bruch — der Hebel liegt meist im Zähler (dem Wert aus Urteil und Kontext), nicht im Nenner (dem Token-Preis). Fallende Stückpreise senken die Rechnung nicht; gesteuert wird über den *Wert je Vorgang*, nicht über die Kosten je Anfrage.

Wissenshierarchie (Kap. 11) — die Organisation als Apparat zur Bewirtschaftung knappen Fachwissens: Routine wird unten gelöst, Ausnahmen steigen zu selteneren Fachleuten auf. Ihre Form ergibt sich aus zwei Kosten — dem Aufwand, eine Frage weiterzureichen, und dem Aufwand, Wissen selbst vorzuhalten. KI verbilligt das Fachwissen an der Front, also dünnt die mittlere Informationsschicht aus.

Wissensstichtag (Kap. 4) — der Zeitpunkt, an dem das Trainingswissen eines Modells endet; alles Spätere kennt es nicht. Der Hauptgrund fürs Nachschlagen.

Zusammengesetzte Fehler (Kap. 5) — bei mehrstufigen Aufgaben multiplizieren sich die Zuverlässigkeiten der einzelnen Schritte (95 % hoch 10 ergibt rund 60 %); lange autonome Ketten sind deshalb prinzipiell zerbrechlich.

QUELLEN UND NACHPRÜF-HINWEIS

Dieses Buch steht und fällt mit seiner Nachprüfbarkeit. Jede Kernbehauptung ist an eine benannte Quelle gebunden: an eine begutachtete Studie, eine amtliche Statistik, einen Arbeitsbericht oder eine namentlich zugeordnete Position. Die Belege stehen als nummerierte Endnoten am Ende des jeweiligen Kapitels, mit Jahr und Fundstelle, und sind hier nicht ein zweites Mal abgedruckt.

Vier Bitten an Sie:

1. **Datierte Zahlen sind datiert, weil sie altern.** Modellpreise, Kontextfenster, Leistungsmerkmale und der Stand der Regulierung — besonders der EU AI Act — ändern sich im Monatsrhythmus. Wo eine Zahl einen Stand trägt, prüfen Sie ihn vor jeder Entscheidung am aktuellen Wert nach.
2. **Umstrittene Zahlen sind als umstritten gekennzeichnet.** Wo eine populäre Statistik (etwa zu gescheiterten KI-Projekten) auf schmaler Datengrundlage steht, nenne ich den Vorbehalt. Übernehmen Sie sie nicht als exakten Wert.
3. **Die Quellen sind Eingang, nicht Autorität.** Auch dort, wo ich eine prominente Stimme zitiere, folge ich ihr nicht blind; wo ich widerspreche, sage ich es und begründe es.
4. **Melden Sie Fehler.** Sollte Ihnen ein Beleg nicht standhalten, schreiben Sie mir. Korrekturen fließen in spätere Auflagen ein. Das ist kein Service, sondern die Konsequenz des Versprechens aus dem Vorwort.

ÜBER DEN AUTOR

Die meisten, die über künstliche Intelligenz schreiben, setzen sie nicht selbst dort ein, wo ein Fehler wehtut. **Dr. Sven Jungmann** schon.

Er ist Arzt und baut mit seinem Unternehmen **aiomics** KI für Krankenhäuser: Software, die den Patientendatensatz korrekt hält, während Kliniken immer mehr Dokumentation an Maschinen übergeben — unter der DSGVO, im laufenden Betrieb, dort, wo ein Fehler nicht nur Geld kostet. Das Gesundheitswesen gilt als eine der reguliertesten und langsamsten Branchen überhaupt. In ihr hat aiomics Klinikverträge in sieben Monaten unterschrieben, wo der Branchenstandard bei achtzehn und mehr liegt.

Das Geld kam in einem Klima, in dem Investoren KI-Anwendungen zunehmend skeptisch prüfen, weil so viele davon austauschbar geworden sind: aiomics hat Kapital von renommierten internationalen Häusern eingeworben — darunter Norrsken Evolve aus Schweden und Calm/Storm aus Österreich — sowie von einer Reihe erfahrener Business Angels bis ins Silicon Valley. Die Wette dahinter ist dieselbe, die dieses Buch entfaltet: etwas zu bauen, das verteidigbar bleibt, wenn die Technik selbst zur Ware wird.

Wie viel das praktisch trägt, zeigt seine eigene Firma. Ein KI-Budget von einigen Tausend Euro im Monat ersetzt dort über 400.000 Euro, die früher pro Jahr in Personal, freie Mitarbeit und Software flossen — bei Ergebnissen, die sein Team und seine Kundschaft besser bewerten als zuvor. Das ist die Kennzahl, um

die das Buch kreist: der Wert, den man je Recheneinheit herausholt — im Buch: *Return on Tokens*.

Sven Jungmann kennt den Markt von beiden Seiten und über die ganze Spanne. Er hat im Corporate Venture Building Unternehmen mitgebaut und beraten, und er investiert selbst — vom Frühphasengeschäft bis zu Transaktionen in der Spätphase. Daraus entsteht der seltene Überblick dessen, der früh einsteigt und spät wieder aussteigt und alles dazwischen gesehen hat.

Seine Ausbildung erklärt die Spannweite: Medizinstudium, dann ein Master of Public Policy in Oxford (mit Systemdenken und Entrepreneurship), ein Entrepreneurship-Diplom in Cambridge, dazu Executive-Programme in Harvard, am MIT und in Stanford. 2017 zählte ihn das *Handelsblatt* zu den hundert klügsten Innovatoren Deutschlands und nannte ihn den „Big Data Doctor".

Dass er auch über das Scheitern schreiben darf, hat er sich verdient. Vor aiomics gründete er das Diagnostikunternehmen **Halitus**, warb eine Million Euro ein (unter anderem von Roche) und verkaufte eine Wohnung, um die Finanzierung zu strecken. 2025 stellte er die Firma ein. Aus diesem Kapitel, sagt er, habe er mehr über das Bauen, Finanzieren und Timing eines regulierten Geschäfts gelernt als aus den erfolgreichen. Ein Buch über Urteilsvermögen kann schlecht von jemandem stammen, der nur seine Siege zeigt.

Er hält Vorträge — auf Deutsch, Englisch und Französisch — vor Unternehmen wie **Pfizer**, **Roche**, **Bayer**, **Novartis**, **AstraZeneca**, **Siemens Healthineers** und **Julius Bär**, und hat Häuser wie **Johnson & Johnson**, **Audi Business Innovation** und die **Gothaer** beraten. Er lebt in Berlin.

Kontakt, Vorträge und aktuelle Veröffentlichungen: [svenjungmann.com / LinkedIn].

ÜBER AIOMICS

aiomics baut künstliche Intelligenz für eine der anspruchsvollsten Umgebungen, die es gibt: das Krankenhaus. Die Software sorgt dafür, dass der Patientendatensatz vollständig und korrekt bleibt — über den gesamten Behandlungsweg, von der Aufnahme bis zur Entlassung, unter der DSGVO und bei voller Datenhoheit des Hauses.

Das ist eine harte Probe für die Versprechen, die in diesem Buch verhandelt werden: unvollständige Daten, strenge Regulierung, knappe Zeit, und der Anspruch, dass das Ergebnis einer fachlichen Prüfung standhält. Genau deshalb stammen die Maßstäbe in diesem Buch aus der Wirklichkeit, nicht aus der Broschüre. aiomics wird von internationalen Investoren getragen (Vorwerk Ventures, Norrsken Evolve, Calm/Storm). Mehr unter **[aiomics.io]**.

Dieses Buch ist keine Werbung für aiomics. Es nennt das Unternehmen nur dort, wo es um den Autor geht. Im Argument kommt es nicht vor — so, wie es das Versprechen aus dem Vorwort verlangt.

THE AIOMIC AGE — DER PODCAST

Ein Buch gibt eine Lage zu einem Zeitpunkt wieder. Künstliche Intelligenz aber bewegt sich schneller, als ein Buch erscheinen kann. Für alles, was nach dem Druck kommt, gibt es **The Aiomic Age** — den Podcast, mit dem Sven Jungmann und Karim Galzahr Führungskräfte strategisch über KI auf dem Laufenden halten.

Die Idee ist dieselbe wie die dieses Buches, nur in Bewegung: nicht jeder Neuigkeit hinterherlaufen, sondern das Signal im Lärm finden und es so einordnen, dass eine Entscheiderin damit etwas anfangen kann. Zwei Gastgeber bringen zwei Blickwinkel ein: der Unternehmer, der KI im Echtbetrieb einsetzt (**Sven Jungmann**), und der Investor mit dem Blick über die Fächergrenzen (**Karim Galzahr**). Sie streiten, prüfen und kommen zu einer begründeten Lesart — die Arbeit, die dieses Buch auf Papier leistet, fortgesetzt mit der Stimme.

Das Versprechen ist schlicht: Wer den Sendungen folgt, soll ein halbes Jahr später besser über künstliche Intelligenz urteilen können als zuvor — und sagen können, warum.

> **Wenn dieses Buch Ihr Urteil geschärft hat, hält der Podcast es scharf.** *The Aiomic Age* — überall, wo es Podcasts gibt, und unter **[the-aiomic-age.com / Link]**.

DANK

Ein Buch über das Beurteilen schuldet seinen größten Dank denen, die das Urteil des Autors geschärft haben. Mein Dank gilt den Kolleginnen und Kollegen bei aiomics, an deren Arbeit ich täglich lerne, was zwischen einer guten Demonstration und einem belastbaren Ergebnis liegt; den Menschen aus Investment, Beiräten und Unternehmen, mit denen ich über Jahre um die richtigen Fragen gerungen habe; und den vielen Zuhörerinnen und Zuhörern meiner Vorträge, deren skeptische Nachfragen mehr Kapitel dieses Buches geformt haben, als ihnen bewusst ist.

Den Forscherinnen und Forschern, deren Arbeiten dieses Buch belegen, gebührt ein Dank eigener Art: Ihre Sorgfalt ist der Grund, warum ich Ihnen, der Leserin und dem Leser, das Nachprüfen anbieten kann.

Und schließlich der Dank, der zum Gegenstand gehört. Dieses Buch ist mit künstlicher Intelligenz geschrieben und von einem Menschen verantwortet. Die Maschine hat unermüdlich recherchiert, formuliert und verworfen; geurteilt, geprüft und entschieden habe ich. Wer für die Fehler einsteht, die geblieben sind, ist deshalb leicht zu benennen: ich.

Namentlich danke ich denen, die mir über die Jahre Rat, Vertrauen und Widerspruch geschenkt haben — meinen Mentoren, Investoren und Beiräten **Karim Galzahr**, **Jeroen Tas**, **Young Sohn** und **Konstantin Othmer**.

EINLADUNG

Wenn dieses Buch Ihnen genützt hat, habe ich drei Bitten — und ein Angebot.

1. **Geben Sie es weiter** — an die eine Person in Ihrem Haus, die diese Fragen als Nächste entscheiden muss.
2. **Hinterlassen Sie bei Amazon eine ehrliche Rezension.** Zwei Minuten genügen. Auch eine kritische hilft der nächsten Leserin, das Buch richtig einzuordnen — und mir, die nächste Auflage besser zu machen.
3. **Melden Sie mir, was nicht standhält** — an **[E-Mail]**. Korrekturen fließen, namentlich verdankt, in spätere Auflagen ein.

Das Angebot: Unter **[URL]** finden Sie die Selbstdiagnose aus Kapitel 14 als Druckvorlage für Ihren Führungskreis. Der Podcast *The Aiomic Age* setzt fort, wo das Buch endet — überall, wo es Podcasts gibt.

Und die wichtigste Bitte zum Schluss, dieselbe wie am Anfang: **Prüfen Sie nach. Nehmen Sie nichts unbesehen.** Auch dieses Buch nicht.