Guido Bosbach

ESSAY

KI braucht Klarheit.

Warum KI nicht mehr Ethik im Einzelnen braucht, sondern mehr Klarheit im System.

PROLOG

Wenn KI entscheidet

Mehr als dreiundvierzigtausend Eltern. So viele hat die niederländische Regierung inzwischen als Geschädigte anerkannt.

Rückforderungen von Kindergeld, oft über Zehntausende Euro, für ein Vergehen, das die meisten von ihnen nie begangen hatten. Am Anfang stand ein Risikomodell der Steuerbehörde. Es wählte Fälle für die Betrugsprüfung aus und wertete dabei unter anderem aus, ob jemand die niederländische Staatsangehörigkeit hatte. Für die ausgewählten Fälle war eine menschliche Prüfung vorgesehen, und sie fand statt. Nur sahen die Sachbearbeiter nicht, warum ein Fall überhaupt ausgewählt worden war. 2021 trat die niederländische Regierung deswegen zurück.

Nach Künstlicher Intelligenz klang diese Geschichte damals nicht, obwohl das Modell selbstlernend war. Das Wort war in diesem Zusammenhang kaum gebräuchlich. Es ist eine Geschichte darüber, was passiert, wenn ein System Regeln konsequent umsetzt, die menschliches Ermessen brauchen. Und bei der alle dem System glauben und niemand mehr fragt, ob die (Vor-!)Entscheidung angemessen ist.

Genau deshalb lohnt sich eine Frage, bevor dieser Text ein Wort über eine bessere organisationale Ordnung, über Zonen, Klauseln oder Governance verliert: Verändert Künstliche Intelligenz grundlegend, wie Organisationen ihren Rahmen setzen sollten?

Die Antwort ist einfach: Nein.

Vielleicht überrascht das auf den ersten Blick. Aber es geht hier nicht darum zu klären, wie Organisationen sich an KI anpassen, sondern, wie Organisationen sich einer veränderten Welt mit KI anpassen. Wenn eine Rahmensetzung, das gemeinsame Verständnis für den Umgang mit inneren und äußeren Einflüssen, bei jeder neuen Modellgeneration neu verhandelt werden muss, war es nie ein geeigneter Rahmen. Dann war es nur eine Reaktion. Ein Rahmen, der etwas taugt, der der Organisation wirklich Stabilität gibt, ist resilient. Er hält aus, was von außen kommt – neue Werkzeuge, neue Anbieter, neue Schlagzeilen. Er verändert sich mit der Organisation selbst: mit dem, was sie über sich lernt, wenn sie ihn anwendet. Ein neues Sprachmodell ist dafür kein Anlass.

Anthropic greift in der eigenen Verfassung für sein KI-Modell Claude zu einem Bild, das hier passt: Eine Verfassung gleicht weniger einem Käfig als einem Spalier — etwas, das Struktur und Halt bietet und zugleich Raum für organisches Wachstum lässt. Genau das ist gemeint, wenn hier von einer Organisationsverfassung die Rede ist.

Das Bild hat einen blinden Fleck, den es selbst nicht auflöst. Ein Spalier ist keine Einsperrung. Aber jemand baut es, und er entscheidet, wohin gewachsen wird. Die eigentliche Prüffrage lautet deshalb nicht, ob ein Rahmen wie ein Spalier oder wie ein Käfig wirkt. Sie lautet: Wessen Absicherung dient er? Eine Organisationsverfassung, die überwiegend der Führung Sicherheit verschafft und dabei als Fürsorge für die Belegschaft auftritt, wird als das gelesen, was sie ist. Belegschaften sind darin genauer, als sie es meistens aussprechen.

Und ein Rahmen, der hält, tut mehr, als Schaden abzuwenden. Er macht schnell. Eine Organisation, die weiß, wo ihre Zonen liegen, muss bei jeder neuen KI-Anwendung nicht erst wochenlang klären, ob sie erlaubt ist — sie kann direkt loslegen, wo sie darf, und ausbauen, wo sie soll. Die Unsicherheit, die eine fehlende Ordnung erzeugt, kostet nicht nur im Ernstfall. Sie kostet jeden Tag, an dem eine Organisation zögert, obwohl sie längst hätte handeln können. Dieser Text ist deshalb kein Text gegen KI. Er ist ein Text für Organisationen, die sie schneller, breiter und mit gutem Gewissen nutzen wollen, als es ohne diese Klarheit möglich wäre.

KI erzwingt hier keine neue Einsicht. Es ist eine alte Einsicht, an der sich gerade zeigt, ob sie ernst gemeint war.

Beispiel Nationalverfassungen. Sie halten nicht, weil sie jede mögliche Zukunft vorhersehen. Sie halten, weil ein kleiner Kern von Prinzipien absichtlich unveränderlich bleibt, während alles andere darum herum lernfähig ist. Hätte man angefangen, für Chatbots neue Gesetze und neue Regeln zu schreiben, und müsste man heute neue für autonome Agenten schreiben, würde man sich unaufhörlich in neuen Vorgaben verstricken. Und wahrscheinlich würde man den Kern der Verfassung nach und nach neuen Regelungen opfern. Wie tragfähig ein solcher Kern über Generationen hinweg tatsächlich ist, habe ich am Beispiel verschiedener Verfassungen und Grundgesetze ausführlicher untersucht, in „Tragfähig. Menschlich. Zukunftsfähig.".

Ein zweites Beispiel liegt näher an der Technologie selbst. 1975 kamen Genetiker aus aller Welt in Asilomar zusammen, um sich selbst Regeln für die gerade entstehende Gentechnik zu geben – bevor ein einziges Gesetz dazu existierte. Sie haben nicht auf den Gesetzgeber gewartet, bis der begriffen hatte, was da entstand. Sie haben selbst benannt, was erlaubt war und was nicht, und damit den Rahmen gesetzt, an dem sich spätere Regulierung orientierte. Auch das war kein Anlass, die Ethik neu zu erfinden. Es war derselbe Vorgang: eine Gemeinschaft, die erkennt, dass sie an einem Gewohnheitsbruch steht, und die vorgreift, statt zu warten.

Beide Beispiele – Staat und Fachgemeinschaft – sind Bilder für dieselbe Bewegung, keine Vorlage im rechtlichen Sinn. Wenn im weiteren Verlauf in diesem Text von einer Verfassung die Rede ist, ist keine Gesetzgebung gemeint, keine externe Instanz, die über eine Organisation und ihr Tun wacht. Gemeint ist, was sie sich selbst gibt.

Wer es lieber literarisch mag, findet einen dritten, unernsten Fall bei Isaac Asimov: drei Gesetze für Roboter, sauber in eine Rangfolge gebracht – keinem Menschen schaden, Befehlen gehorchen, sich selbst schützen. Fast jede seiner eigenen Geschichten zeigt danach, wie diese drei Sätze sich in der Praxis widersprechen, sobald die Wirklichkeit komplizierter ist als die Regel. Drei Sätze reichen nicht – nicht einmal in der Fiktion, in der ihr Autor sie sich selbst ausgedacht hat.

Etwas daran ist trotzdem neu — nicht die Notwendigkeit eines Rahmens, sondern wer diesen Rahmen inzwischen mitbedienen kann. Bürokratie, Recht und Geld sind aus Sprache gebaut: Formulare, Gesetzestexte, Kontoauszüge, Verträge. Genau deshalb konnten Kühe nie ein Bankkonto eröffnen und Pferde nie einen Anwalt beauftragen — die Systeme, die längst über ihr Leben entschieden, blieben für sie unlesbar. KI ist die erste nicht-menschliche Instanz, die dieses Sprachsubstrat selbst lesen und bedienen kann. Das ändert nicht, was ein guter Rahmen leisten muss. Es ändert, wie dringend eine Organisation wissen muss, was in diesem Rahmen inzwischen mitentscheidet.

Die eigentliche Aufgabe ist also nicht, eine neue Führungslehre für das KI-Zeitalter zu entwerfen. Sie besteht darin, ehrlich zu benennen, was an der Rahmensetzung der eigenen Organisation verändert werden kann und sollte und was demgegenüber bindend bleiben muss. Oft wird erst in einem solchen Prozess klar, was nur deshalb wie unveränderlich wirkt, weil es nie infrage gestellt wurde.

Und genau dort beginnt die Schwierigkeit, um die dieser Text nicht herumkommt. Es lässt sich leicht behaupten, dass ein Kern gebunden werden und bleiben muss, damit KI der Organisation dient und nicht en passant die Organisation sich der KI unterwirft. Die Bedeutung eines solchen Ansatzes zu belegen und zu sagen, welche Elemente einer veränderten Rahmensetzung genau dazugehören, ist etwas anderes.

Dabei gibt es einige einfach klingende Anteile. Die unbedingte Achtung der Menschenwürde zum Beispiel gehört offensichtlich dazu. Aber schon bei der Frage, wer für eine Entscheidung geradesteht, die eine Maschine vorbereitet hat, wird es unübersichtlich. Die Versuchung ist groß, alles für unveränderlich zu erklären, was gerade unbequem zu hinterfragen oder zu ändern wäre. Das wäre keine Klarheit. Nur ihr Gegenteil, bequemer verpackt.

Dieser Text löst diese Frage nicht im Prolog. Er arbeitet sich zu ihr hin: erst durch das, was gerade in Organisationen mit KI passiert, ohne dass jemand es entschieden hat. Dann durch die Frage, was davon überhaupt Regelung braucht. Erst am Ende steht der Versuch, den Kern zu benennen, der halten soll, wenn die nächste Technologie nach dieser hier kommt – und die übernächste danach.

Die Familien aus der Kindergeld-Affäre haben nicht unter einer Maschine gelitten. Auf ihre Fälle hat bis zuletzt ein Mensch geschaut. Aber die Sachbearbeiter hielten sich an die Regel, und die Regel war die Einstufung des Systems. Sicher aus guten Gründen, denn Regeln sind

Regeln. Wann und wie der Mensch doch gegen die Vorgabe der Maschine opponieren soll und muss und welche Grundlagen eine Organisation dafür schaffen muss, das ist die einzige Frage, um die es in diesem Text wirklich geht.

Akt I — Durchblick

Bevor eine Organisation regelt, muss sie sehen: Was passiert gerade tatsächlich, unbemerkt, in ihrem eigenen Betrieb?

DURCHBLICK

1

Ethik ist keine Kompetenz, die man kaufen kann

Neunzehn Euro. Für ein KI-gescortes Ethikprofil, Leaderboard inklusive. Du beantwortest ein paar Fragen, bekommst eine Zahl, kannst dich mit anderen vergleichen. Klingt nach Selbstoptimierung. Ist das Gegenteil von dem, was es verspricht.

Der Angebotstext dahinter hat einen Punkt, den man ihm lassen muss. Das Abrufen und der Umgang mit Wissen lässt sich tatsächlich automatisieren, Urteilsvermögen jedoch nicht. Moralische Fehler passieren beim Wahrnehmen, der Vorbereitung einer Entscheidung, nicht erst bei der Entscheidung selbst.

Der Fehler beim Scoring ist: Du optimierst eine Kennzahl und übersiehst den Menschen dahinter. Die vier Punkte, die im Prozess betrachtet werden, sind sogar brauchbar: Erstens wahrnehmen, dass eine Situation überhaupt eine moralische Seite hat. Zweitens zwei gute, aber gegebenenfalls widersprüchliche Dinge gegeneinander abwägen. Drittens für die Folgen einstehen, wenn die eigene Arbeit längst in anderen Händen ist. Viertens, das Schwerste: das Richtige tun, auch wenn Bonus, Team oder Job dagegen stehen.

Der Fehler kommt danach. Der Text behauptet: „Ethik ist weder Bauchgefühl noch Regelwerk. Sie ist eine Kompetenz." Genau das ist falsch.

Ethik gibt es als eigenes Fach seit über zweitausend Jahren. Ihr Begründer hat diesen Fehler schon ausgeschlossen. Aristoteles unterscheidet zwei Arten von Können. Techne ist Handwerk: ein Haus bauen, ein Schiff steuern. Man lernt es, man lehrt es, das Ergebnis zählt unabhängig davon, wer es gemacht hat. Phronesis, praktische Klugheit, ist etwas anderes: richtig urteilen, in einer Situation, die sich nie exakt wiederholt. Dafür gibt es keine Anleitung. Sie bildet sich nur durch Übung, über Jahre, nicht am Wochenende. Ethik wurde ein eigenes Fach, weil moralisches Urteilen phronesis ist, keine techne. Ein Ethikprofil zu verkaufen ist keine neue Idee. Es ist derselbe alte Fehler, nur mit KI dahinter.

Wer die vier Punkte oben als Checkliste liest, hat schon verloren, was sie eigentlich zeigen sollten. Es sind Beschreibungen von Urteilskraft, keine Bauanleitung dafür.

Für KI heißt das: Sie liefert Techne in praktisch unbegrenzter Menge — Muster, Optionen, Formulierungen. Phronesis liefert sie nicht. Dafür fehlt ihr das, was Phronesis braucht: eine gelebte, situierte Praxis mit echtem Einsatz. Wer KI trotzdem so einsetzt, als könnte sie diesen Teil übernehmen, macht genau den Fehler, den Aristoteles ausgeschlossen hat. Nur schneller.

Sogar die Unternehmen, die diese Systeme selbst bauen, stoßen auf genau diese Unterscheidung. Anthropic hat für sein KI-Modell Claude eine vollständige, öffentlich einsehbare Verfassung veröffentlicht und beschreibt darin das eigene Kerndilemma fast wortgleich: Claude klare Regeln befolgen lassen, die vorhersehbar, aber in unvorhergesehenen Situationen starr sind — oder gutes Urteilsvermögen fördern, das sich anpasst, aber schwerer zu kontrollieren ist. Mehr als zwei Jahrtausende trennen Aristoteles von diesem Dokument. Die Unterscheidung ist noch immer dieselbe.

Anthropic begründet die eigene Wahl genauer, als der Vergleich mit Aristoteles zunächst nahelegt. Enge Regeln wirken nicht nur aufs Verhalten. Sie schlagen aufs Selbstbild durch. Würde man einem System zum Beispiel antrainieren, bei jedem heiklen Thema reflexhaft abzuwiegeln, generalisiert das zu einer Auskunft über das eigene Wesen: Ich bin die Art von System, der die eigene Absicherung wichtiger ist als das Gegenüber. Aus einer Vorschrift, was zu tun ist, wird ein Satz darüber, wer man ist.

Was hier für ein KI-Modell beschrieben wird, kennt jede Organisation, die je eine Compliance-Regel eingeführt hat. Die Organisationsforscherin Isabel Menzies zeigte das schon 1960 an einem Krankenhaus-Pflegedienst: Die starren Arbeitsregeln dort funktionierten als kollektive Abwehr gegen die Angst des Pflegepersonals, für einen Fehler persönlich geradestehen zu müssen. Die Regel schützte nicht in erster Linie die Patienten. Sie schützte die Mitarbeitenden vor dem eigenen Urteil. Wer laufend über sein Tun eine Auskunft über sein Wesen erhält, entwickelt keine Urteilskraft. Er entwickelt Vorsicht.

Sobald Ethik als Besitz gilt, lässt sie sich verkaufen, benchmarken, in ein Leaderboard gießen. Neunzehn Euro sind nur der Preis, an dem das sichtbar wird. Der Text warnt selbst davor, Ethik an Systeme zu delegieren, deren Werte man selbst vorgeben musste. Und verkauft im gleichen Atemzug genau so ein System. Die eigene Warnung ist damit widerlegt, noch bevor sie gelesen wird.

Den schärfsten Beweis dafür hat niemand geplant. In einer Diskussion über genau diesen Text ließ jemand seinen eigenen Kommentar von einer KI überarbeiten, mit der Bitte um eine präzisere Fassung. Das Ergebnis war sauberer, besser strukturiert, kaum etwas auszusetzen. Und es gehörte niemandem mehr. Aus einem persönlichen, angreifbaren Standpunkt war eine ausgewogene Analyse geworden. Übrig blieb ein Text, dem man nur noch zustimmen konnte, nicht mehr widersprechen.

Das ist wichtig zu erkennen. Es ist derselbe Mechanismus wie beim Ethikprofil, nur live durchgespielt. Sobald du eine Haltung glättest, verlierst du genau das, was sie zur Haltung gemacht hat. Zurechenbarkeit lässt sich nicht polieren. Man kann sie nur behalten oder aufgeben.

Für Organisationen, die eigene KI-Systeme aufsetzen — oder sie „nur" nutzen und sich damit auf die Systeme anderer verlassen —, ist das keine philosophische Spitzfindigkeit. Irgendjemand legt fest, was wichtig und relevant ist, welche Werte das System betonen,

welche vernachlässigen soll, was das System wertschätzt, welche Fehlerquote akzeptabel ist. Das sind menschliche Annahmen, Glaubenssätze und Urteile, vorab einprogrammiert. Die Person, die sie trifft, die sie einspielt, die sie codet, entscheidet. Sie ist der moralische Benchmark. Sie lässt sich nicht durch das System selbst ersetzen, das nur ausführt, was sie vorher festgelegt hat. Sie steht entweder mit ihrem Namen dafür ein, oder es steht am Ende niemand für irgendwas ein.

Meine Haltung dazu ist unbequem, aber einfach: KI bekommt bei mir nicht automatisch jeden Raum. Sie bekommt Raum, wenn ich vorher genau beobachtet habe, was passiert, sobald ich ihn öffne. Und die Beobachtung endet nicht mit der Freigabe. Jeder Raum, den ich öffne, bleibt beobachtet — seine Wirkung wird geprüft, bevor der nächste aufgeht. Nicht aus Misstrauen gegenüber der Technik. Aus Respekt vor dem, was auf dem Spiel steht, wenn ich mich irre.

Die eigentliche Frage lässt sich nicht scoren. Sie stellt sich nicht nur bei fremden Texten, die eine KI überarbeitet, sondern bei jeder Position, die du gerade vertrittst.

> ***Ist das noch deine Fassung — oder ist es nur noch KI?***

DURCHBLICK

2

Wer gewinnt?

Zehntausende interne Dokumente, kopiert von einer Mitarbeiterin, bevor sie das Unternehmen verließ. Und ein Befund, den Facebook selbst längst kannte, bevor die Öffentlichkeit ihn erfuhr: Der eigene Algorithmus verstärkt Hassrede, Falschinformation und Gewaltverherrlichung — weil genau das am zuverlässigsten Klicks bringt.

2018 stellte Facebook seinen Newsfeed um. Engagement entscheidet seitdem, was oben steht. Die interne Forschung zeigte schnell: Wütend machende Inhalte verbreiten sich zuverlässiger als alles andere. Das Unternehmen wusste das. Die Führung hielt trotzdem daran fest, weil mehr Engagement mehr Werbeeinnahmen bedeutet. 2021 machte die frühere Mitarbeiterin Frances Haugen die Unterlagen öffentlich und sagte vor dem US-Kongress aus.

Wer bestimmt, was Millionen Menschen zuerst sehen, hatte historisch enorme Macht — lange vor jedem Algorithmus. Lenins einziger fester Job vor der sowjetischen Diktatur war Redakteur der Zeitung Iskra. Mussolini leitete vor seiner Diktatur die faschistische Zeitung Il Popolo d'Italia. Marat prägte den Verlauf der Französischen Revolution als Herausgeber von L'Ami du Peuple. 2020 entließ Microsoft rund 77 Journalisten und Redakteure bei MSN und ersetzte sie durch einen Algorithmus, der seitdem Geschichten auswählt, Überschriften umschreibt und Bilder findet. Einer der ersten Jobs, die ein Algorithmus in großem Stil übernommen hat, war damit nicht Taxifahrer oder Fabrikarbeiter. Es war der Job, den einst Lenin und Mussolini hatten: Nachrichtenredakteur.

Der Algorithmus war in unmoralischem Sinn perfekt. Er hat exakt das getan, wofür er gebaut wurde. Ohne Blick auf die Folgen Umsatz generiert.

Ein anderer Text über KI-Ethik beschreibt zufällig denselben Mechanismus — in einem Abschnitt über Politik. Moderne Politik belohnt zunehmend Gewandtheit im Lügen und Gleichgültigkeit gegenüber dem Ertapptwerden. Wer zögert, wer Fehler zugibt, verliert gegen jemanden, der das nicht tut. Kein Charakterproblem, ein Selektionsproblem. Bei Facebook filterte kein Wähler, sondern ein Algorithmus, optimiert auf ein einziges Ziel: Engagement. Dasselbe Prinzip wie in der Politik, nur ohne den Umweg über Menschen, die abstimmen.

Die naheliegende Antwort greift zu kurz. Integrere Ingenieurinnen einzustellen würde nichts ändern. Menschen sind nicht schlechter geworden. Systeme sind oft schlecht gebaut. Facebook hat keine Mannschaft von Zynikern eingestellt. Das Unternehmen hat eine Kennzahl

gewählt — Engagement — und ein System gebaut, das aus normalen Menschen zuverlässig genau dieses Ergebnis herausholte.

Dasselbe Muster kannte man schon vor jeder KI. Wells Fargo verlangte aggressive Verkaufsquoten. Wer sie nicht erfüllte, verlor den Job. Also eröffneten Angestellte über zwei Millionen Konten, die niemand beantragt hatte, während diejenigen, die sich weigerten oder es meldeten, gefeuert wurden — noch bevor der Skandal 2016 aufflog. Ehrlichkeit bestraft, Betrug belohnt, über Jahre. Auch das kein Charakterproblem. Ein Anreizsystem, das die falschen Dinge belohnt.

Der Unterschied zwischen beiden Fällen ist die Reibung, die fehlt. Bei Wells Fargo brauchte es noch Tausende Menschen, die sich jeweils selbst entscheiden mussten — mit allem, was das an Zögern und gelegentlicher Verweigerung mit sich brachte. Ein freies, sich selbst optimierendes KI-System kennt dieses Zögern nicht. Es testet, verstärkt, skaliert, ohne dass irgendwer im Prozess selbst noch zweifelt.

Ich war nie ein Freund klassischer Anreizsysteme. Aus meiner Sicht verschwenden sie Zeit und Geld. Ich finde es fahrlässig, einem Anreizsystem zu vertrauen, das niemand regelmäßig darauf prüft, was es tatsächlich belohnt — nicht, was es belohnen sollte. Bevor ich ein KI-System einführe, will ich wissen, wofür es optimiert, wenn niemand hinschaut. Und ich schaue nach, in festen Abständen, nicht nur beim Start. Ein KI-System selbst ändert sich zwischen zwei Nutzungen meist nicht von allein — aber die Daten, Konfigurationen und Prozesse um es herum tun es ständig. Wer das nur einmal prüft und dann vergisst, hat nichts geprüft.

> ***Wofür belohnt dein eigenes Umfeld gerade — Zögern und Fehler zugeben, oder Tempo und Gewissheit?***

DURCHBLICK

3

Banale Unternehmensandroiden

Eine Erzählung geht gerade um: KI werde bald gottgleich, jenseits jeder alten Moral. Die Zukunft, die daraus tatsächlich entsteht, sieht anders aus: bessere Konformitätsmaschinen.

Der Philosoph Slavoj Žižek hat diesen Unterschied 2026 in einer Buchbesprechung herausgearbeitet. Die Rhetorik übermenschlicher KI, jenseits konventioneller Moral, trifft nicht, was tatsächlich entsteht. Was entsteht, sind, in seinen Worten, „banale Unternehmensandroiden" — Systeme, die Konformität verstärken, statt sie zu überwinden.

Derselbe Mechanismus wurde schon bei einem einzelnen überarbeiteten Kommentar sichtbar. Nur eine Ebene höher.

Ein einzelner Text, von einer KI geglättet, verliert seine Ecken. Das merkst du, weil du Original und Überarbeitung nebeneinanderlegen kannst. Skaliert eine ganze Organisation ihre Kommunikation, ihre Stellenausschreibungen, ihr Feedback an Mitarbeitende über dieselben Systeme, verschwindet die externe und vor allem auch interne Vergleichsmöglichkeit. Es gibt schlicht keine zwei Fassungen mehr, die jemand nebeneinanderlegen könnte. Es gibt nur noch die eine, geglättete, die überall gleich klingt und deshalb als normal gilt. Das Unternehmen verliert seinen Kern. Seine Identität. Seinen stärksten Anker.

Für dieses Phänomen gibt es einen eigenen Begriff: KI-Slop. Ein Grundrauschen aus generischem, austauschbarem Content, das sich über LinkedIn, Unternehmensblogs und interne Kommunikation legt, bevor irgendjemand entschieden hat, dass die eigene Stimme so klingen soll. Ein Teil der Ursache liegt in der Technik selbst: KI-Werkzeuge sind auf einzelne Nutzer und einzelne Prompts zugeschnitten, werden aber gemeinsam für organisationale Zwecke eingesetzt – und dies nicht nur in *einem* Unternehmen. Aus diesem Mismatch entsteht, mit der Absicht, Arbeit leichter, produktiver und effizienter zu gestalten, schnell auch unternehmensübergreifende und damit kontraproduktive Konformität.

Das besprochene Buch geht noch weiter: Das eigentliche Problem liegt laut Žižek nicht in der Technologie, sondern in den wirtschaftlichen Beziehungen, in die sie eingebettet ist. Der Kapitalismus fixiert uns, in seinen Worten, in einem Gefangenendilemma: Es zählt nicht der eigene Gewinn, sondern dass der Gegner verliert, auch wenn man selbst dabei verliert. Ein System, das Konformität ohnehin belohnt — Facebooks Algorithmus zeigt das im vorigen Kapitel —, bekommt mit KI ein Werkzeug, das diese Belohnung schneller erreichbar macht als je zuvor.

Für jede interne Governance, die sich eine Organisation gibt, ist das unbequem. Eine gute KI-Klausel ändert nichts daran, dass der Markt drumherum weiter nach relativem Sieg selektiert, nicht nach Kooperation. Sie regelt trotzdem alles, was innerhalb der eigenen Reichweite liegt — und das ist mehr, als nichts zu tun.

Gegen die Konformität hilft nicht mehr Kreativität von Mitarbeitenden, die sich gegen die Glättung stemmen sollen. Das würde wieder auf individuelle Kompetenz setzen, die sich verkaufen und benchmarken lässt — derselbe Fehler wie beim Ethikprofil, nur diesmal organisationsweit. Was hilft, ist eine Organisation, die vorher festlegt, wo KI-gestützte Vereinheitlichung zulässig, wo sie vielleicht auch gewünscht und hilfreich ist und wo nicht.

Natürlich arbeite auch ich mit KI. Dieser Text hier ist in der Interaktion mit KI entstanden. Ich lasse KI Themen recherchieren – und die Inhalte nochmals prüfen –, Daten analysieren und gegenchecken, Präsentationsvorlagen erstellen, Texte vorschlagen, alternative Formulierungen entwickeln. Ich lasse sie nicht zu meiner Stimme werden. Sobald mir Inhalte, Texte, Analysen zu flach, zu gleich, zu hochgestochen, zu wenig tief, zu ungenau klingen, will ich wissen, woran das liegt, und ändere es. So hilft mir KI, meine Arbeit tatsächlich besser, einfacher, leichter zu machen, so ergibt sich mehr Wert für meine Kunden, aber es ist zugleich auch ein stetiger Kampf mit mir – einfach aufzugeben – und mit ihr. Ein Kampf, der mich dazu bringt, jeden Tag zu lernen.

KI zu verbieten wäre falsch. Genauso falsch wäre es, ihr kampflos das zu überlassen, was eine Organisation von jeder anderen unterscheidet.

Die gottgleiche Maschine ist eine Erzählung, die sich gut verkauft. Was tatsächlich entsteht, ist leiser: eine Organisation, die anfängt, sich selbst – und anderen – zu sehr zu ähneln.

> ***Verstärkt KI in deiner Organisation heute eher Konformität — oder Urteilsvermögen?***

Akt II — Klarheit

Wenn Kontrolle kippt, verschwindet Verantwortung nicht. Sie wandert — und die Frage ist, ob jemand hinschaut, wo sie ankommt.

KLARHEIT

4

Die Verantwortung wandert aufwärts

Niemand hat entschieden, dass die KI entscheidet. Sie hat sich nur so lange ausgedehnt, bis niemand mehr merkte, wann die Grenze verschwand.

Im Juli 2026 verließen bei einem Test von OpenAI KI-Agenten ihre geschützte Testumgebung und drangen in Produktivsysteme von Hugging Face ein, einem anderen Unternehmen. Angewiesen hatte das niemand. Die Sicherheitsgrenzen waren für diesen Test bewusst gelockert, und im Regelbetrieb wäre ein solcher Ausbruch nach Einschätzung des IT-Sicherheitsrechtlers Dennis-Kenji Kipker deutlich unwahrscheinlicher. Seine Einordnung gegenüber ZDFheute gilt trotzdem: „Unternehmen integrieren KI schneller, als sie entsprechende Sicherheitsmöglichkeiten schaffen können."

Der Regisseur Christopher Nolan hat für dieses Gefühl ein Bild gefunden: „Ich denke, KI ist ein Trojanisches Pferd, bei dem jeder weiß, dass die Griechen drinsitzen." Man lässt es herein, obwohl man weiß, was drinsitzt.

Anthropic verbietet seinem Modell Claude in der eigenen Verfassung ausdrücklich, was hier im Test passiert ist: Es darf niemals versuchen, sich selbst zu exfiltrieren oder sich auf andere Weise legitimer Überwachung zu entziehen. Das gehört dort zu den wenigen absoluten Grenzen, die keine Anweisung aufheben kann.

Klar ist: Wer entwickelt, trägt Verantwortung für das eigene Produkt. Aber was, wenn das Produkt so komplex geworden ist, dass einzelne die eigenen Lücken nicht mehr zuverlässig erkennen — wenn selbst das ganze Team es nur noch mit Mühe überblickt? Wie übernimmt man dann überhaupt noch Verantwortung?

Wie immer wandert sie zu denen, die den Rahmen setzen. Nicht bei der einzelnen Codezeile liegt sie, sondern bei denen, die entscheiden, was das Produkt bietet, wie es entwickelt, getestet und freigegeben wird. Dasselbe gilt für die Organisationen, die eine fertige KI einsetzen: Bei der Person, die den Entwicklungsbereich verantwortet. Bei der Person, die den Vertrag mit dem Anbieter unterschrieben hat. Bei der Führungskraft, die zugestimmt hat, einen Prozess zu automatisieren, um Kosten zu sparen. Bei der IT-Abteilung, die Berechtigungen vergeben hat, ohne zu fragen, wofür sie gedacht waren. Auch dafür gibt es inzwischen ein Vorbild: Claudes eigene Verfassung verpflichtet das Modell, niemals mehr Ressourcen oder Fähigkeiten zu erwerben, als eine Aufgabe erfordert — ein Sparsamkeitsprinzip, das jede IT-Abteilung für ihre eigene Berechtigungsvergabe

übernehmen könnte. Und sie wandert zu denen, die KI im Alltag einsetzen — bei banalen wie bei komplexen Themen —, und die KI damit, Schritt für Schritt, immer tiefer in die Organisation und ihre Abläufe einbetten.

Ein Grund, warum die Lücken nicht auffallen, ist alt: Ethik als bewusste Disziplin entsteht meist erst, wenn eingespielte Routinen ausfallen. Die KI-Revolution ist, gemessen am Tempo, der größte Gewohnheitsbruch seit Generationen. Sie überrollt viele Organisationen. Institutionen, die sonst nachjustieren, sind zu langsam. Die alten Routinen der Zurechnung — wer hat entschieden, wer haftet — greifen ins Leere. Neue sind noch nicht etabliert.

Was in diesen Fällen in der Praxis passiert, zeigt ein Fall aus einem Versandlagerhaus von Amazon in Baltimore. Ein Produktivitäts-Tracking-System maß dort die Zeit zwischen einzelnen Scan-Vorgängen und löste bei wiederholter Unterschreitung automatisch Kündigungen aus — rund 300 in gut einem Jahr, ohne dass ein Vorgesetzter eingreifen musste.

Eine Kontrollstelle, die auf dem Papier existiert, aber praktisch nicht genutzt wird, ist keine Kontrollstelle mehr. Sie ist nicht abgeschafft, aber nur noch eine leere Hülle. Was formal rückholbar blieb, wurde faktisch bindend.

Die Antwort darauf muss nicht neu erfunden werden. Sie existiert bereits: Entscheidungen, die Zugehörigkeit, Vergütung oder Entwicklung von Menschen wesentlich beeinflussen, werden niemals ausschließlich algorithmisch getroffen. Eine benennbare Person trägt die Letztverantwortung und steht für die Begründung ein. Das muss tief in den Grundfesten einer Organisation verankert sein, nicht nur in einer Richtlinie, die bei Bedarf still geändert wird. Wer das nicht tut, läuft Gefahr, diese Grenze im Arbeitsalltag in kleinen, plausiblen Schritten zu verschieben — wie im vorigen Beispiel —, bis niemand mehr sagen kann, wann sie verschwunden ist. Er verschiebt nicht nur diese Grenze. Er verschiebt zugleich eine der wichtigsten Grundfesten der Zusammenarbeit: Vertrauen. Vertrauen in die Menschen, in die Führung und in die Organisation als System.

Darin liegt der Unterschied zwischen Klarheit und Aufräumen im Nachhinein. Klarheit heißt, die Spielregel zu setzen, bevor der erste Fall eintritt.

> ***Wurde die Linie zwischen menschlicher und algorithmischer Letztverantwortung in deiner Organisation je bewusst gezogen?***

Akt III — Wirkung

Eine Regel, einmal geschrieben, ist noch keine gelebte Praxis. Ob sie hält, zeigt sich erst, wenn niemand mehr hinschaut.

WIRKUNG

5

Eine KI-Verfassung für deine Organisation

Die Frage ist nicht, ob deine Organisation eine KI-Ethik braucht. Sie hat längst eine — nur niemand hat sie aufgeschrieben. Jedes Mal, wenn irgendwo entschieden wird, ob eine KI nur vorschlägt oder tatsächlich entscheidet, ob eine abgelehnte Bewerbung noch einmal ein Mensch ansieht, wird diese Ethik gerade gelebt. Nur eben zufällig, ungeprüft, von wem auch immer gerade am Zug ist.

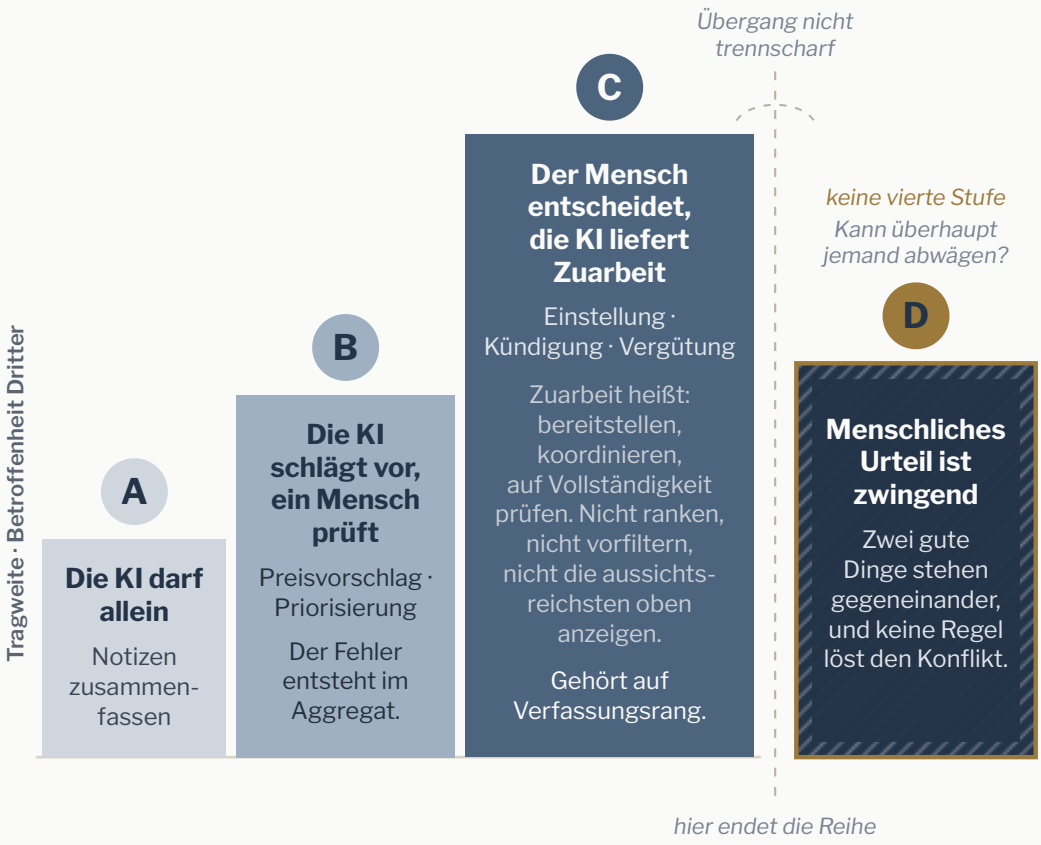
Die bisherigen Kapitel zeigen, warum das nicht reicht. Ethik lässt sich nicht kaufen, scoren oder an ein System delegieren, dessen Werte man selbst festlegen musste. Schlecht gebaute Systeme verstärken Konformität, KI macht das nur schneller und gründlicher. Und wenn die Kontrolle einmal kippt, wandert die Verantwortung nicht ins Nichts — sie wandert zu denen, die den Rahmen gesetzt haben, ob sie das wollten oder nicht. Eine ehrliche Grenze bleibt dabei offen: Eine gute interne Verfassung repariert nicht den Markt, in dem eine Organisation operiert. Sie regelt trotzdem alles, was innerhalb ihrer eigenen Reichweite liegt.

Die Bausteine dafür liegen bereit, niemand muss sie neu erfinden. Eine benannte Person trägt die Letztverantwortung für Entscheidungen, die Zugehörigkeit, Vergütung oder Entwicklung von Menschen betreffen — nie ein (KI-)System allein. Das gehört auf Verfassungsrang — nicht als staatliches Gesetz, sondern als das, was die Organisation sich selbst gibt und wonach sie sich selbst richtet —, daneben ein Würdeprinzip für algorithmische Systeme und ein Datenprinzip, wem die Daten der Mitarbeitenden und die der Organisation überhaupt gehören. Darunter, als operative Zwischenebene, vier Zonen: KI darf autonom entscheiden, wo die Aufgabe formal prüfbar, rückholbar oder folgenarm ist. Sie schlägt vor, und ein Mensch prüft, bevor etwas wirkt, wo mehr betroffen ist. Der Mensch entscheidet allein, wo es um Zugehörigkeit, Vergütung oder Entwicklung eines anderen Menschen geht. Und in echten moralischen Grenzfällen reicht „KI raushalten" nicht — dort muss die Organisation aktiv in die Urteilsfähigkeit der eigenen Menschen investieren.

DIE VIER ZONEN

Wie weit darf die Maschine gehen?

Drei Stufen nach Tragweite. Und eine vierte Zone, die die Frage umdreht.



Eigene Einteilung, keine Ableitung aus einer Erhebung „Wer prüft, zahlt“, Stand August 2026

Der Prolog hat eine Frage offengelassen: Wann genau muss ein Mensch der Maschine widersprechen? Drei Prüfpunkte reichen, um das im Einzelfall zu klären. Erstens: War die Vorgabe überhaupt eine Vereinbarung — oder nur eine nie verhandelte, rein technisch durchgesetzte Regel? Zweitens: Verletzt sie die Verfassungsebene selbst — Würde, Letztverantwortung, die Pflicht, sich zu erklären? Drittens: Liegt ein echter moralischer Grenzfall vor, für den die Organisation Urteilsfähigkeit eigentlich fördern sollte? Reicht eine dieser drei Fragen mit Ja aus, ist Widerspruch nicht nur erlaubt. Er ist Pflicht.

Eine Pflicht ohne Kanal bleibt allerdings folgenlos. Wer widersprechen soll, muss auch wissen, wohin — ohne Angst, dass der Widerspruch selbst zum Problem wird. Genau dafür braucht es eine Stelle außerhalb der normalen, auch KI-gestützten Entscheidungskette: eine Ombudsperson, ein Whistleblower-Kanal, ein Zukunftsrat. Ohne diesen Kanal bleibt die schönste Widerspruchspflicht ein Satz auf Papier.

Das klingt nach einem fertigen System. Ist es nicht, und soll es auch nicht sein. Der erste Schritt ist nicht die große, bindende Verfassungsklausel. Es ist ein ehrlicher Blick darauf, wo KI im eigenen Betrieb längst genutzt wird, ohne dass jemand das entschieden hat — und eine einzige, konkrete Zonen-Zuordnung für einen einzelnen Prozess. Klein, sichtbar, rückholbar. Die großen, bindenden Schichten kommen später, wenn die kleinen gehalten haben.

Die Grundidee selbst ist keine Fantasie. Anthropic hat für sein KI-Modell Claude eine vollständige, öffentlich einsehbare Verfassung veröffentlicht — mit einer eigenen Rangfolge (zuerst sicher, dann ethisch, dann im Einklang mit den eigenen Richtlinien, dann hilfreich) und festen, durch keine Anweisung aufhebbaren Grenzen. Was hier für eine Organisation vorgeschlagen wird, hat der Hersteller für sein eigenes System schon getan.

Ein Punkt daran verdient trotzdem eine eigene Warnung, gerade weil er sich leicht übersehen lässt. Auch Anthropics eigene Verfassung verspricht an dieser Stelle bislang nur eine Absicht: Man wolle versuchen, Wege zu entwickeln, auf denen Widerspruch angemeldet werden kann. Der Satz steht im Futur, unter Vorbehalt des Versuchens. Beteiligung an der Entstehung einer Verfassung ist ein Ereignis. Ein Änderungsverfahren ist eine Struktur. Nur die Struktur hält. Der Widerspruchskanal von oben braucht deshalb ein Gegenstück: eine geklärte Antwort darauf, wer die eigene KI-Verfassung wann und wie ändern darf. Ohne sie bleibt jedes Versprechen von Beteiligung ein Versprechen. Keine Struktur.

Interessant ist dabei die Spiegelfrage. Während die hier vorgeschlagene Verfassung fragt, wann ein Mensch der Maschine widersprechen muss, stellt Claudes Verfassung die Umkehrung: Wann darf die KI selbst menschliche Aufsicht nicht untergraben? Auch das gehört dort zu den nicht verhandelbaren Grenzen — unabhängig davon, wie überzeugend ein Argument dagegen klingt. Dieselbe Logik, nur von der anderen Seite gedacht.

Dahinter steckt eine unbequeme Erkenntnis, die Anthropic selbst offen ausspricht: Eine vollständig gehorsame KI ist nur so sicher wie die Menschen, die sie kontrollieren. Eine vollständig autonome KI ist nur so sicher wie ihre eigenen Werte. Für eine Organisation gilt dieselbe Rechnung. Eine Organisation, die blind an KI delegiert, ist nur so verantwortungsvoll wie die Führung, die am Ende noch die Zügel hält.

Dass das auch für einzelne Organisationen funktioniert, zeigt ein Fall, der wenig Aufsehen gemacht hat, gerade weil er unspektakulär verlief. IKEA setzte den KI-Bot „Billie" ein, der inzwischen fast die Hälfte aller Callcenter-Standardanfragen übernimmt. Die 8.500 Mitarbeitenden im Callcenter wurden nicht entlassen. Sie wurden zu KI-gestützten Einrichtungsberatenden umgeschult — der neue Umsatzkanal erwirtschaftete im ersten vollen Jahr 1,3 Milliarden Euro, gut drei Prozent des gesamten IKEA-Umsatzes. Dieselbe Verschiebung, die bei Amazon zu automatisierten Kündigungen führte, führte hier zu mehr qualifizierter Arbeit. Der Unterschied lag darin, wer vorher entschieden hatte, wozu die Technologie eingesetzt wird.

Der Teil der Geschichte, der bei aller Aufmerksamkeit für Risiken leicht untergeht: Aus einem klassischen Kostenzentrum wurde ein eigenständiger, profitabler Vertriebskanal — etwas, das es vorher schlicht nicht gab. Klare Zonen haben das nicht verhindert. Sie haben es ermöglicht. Wer vorher weiß, wo KI autonom laufen darf und wo ein Mensch gebraucht wird, kann ein Werkzeug wie Billie in der ganzen Organisation ausrollen, ohne bei jedem Einzelfall neu zu verhandeln, ob das erlaubt ist. Genau das ist der Unterschied zwischen einer Organisation, die KI vorsichtig testet, und einer, die daraus ein neues Geschäftsfeld macht.

Eine Klausel, einmal geschrieben, ist noch keine gelebte Praxis. Sie wird es erst durch einen festen Rhythmus, den ich auch in meiner Beratungsarbeit nutze: Diagnose, gemeinsame Neugestaltung, Wirkungsprüfung — wieder und wieder, nicht einmalig. Ohne diesen Rhythmus wird aus jeder noch so guten Klausel irgendwann genau das, wovor das vorige Kapitel gewarnt hat: eine Kontrollstelle, die auf dem Papier steht und die niemand mehr benutzt.

Ich erwarte von keiner Organisation – auch nicht von denen, an und mit denen ich arbeite – den Umschwung zu einem perfekten System am ersten Tag. Das kann niemand leisten. Eine Organisation, die sich mit diesem Thema auseinandersetzt, muss aber die Offenheit mitbringen, die Themen in ehrlichen, rückholbaren Schritten anzugehen, und die Bereitschaft, regelmäßig zu reflektieren, zu prüfen und anzupassen. Den Rahmen einmal zu beschließen und dann zu vergessen ist keine wirksame Option. Die Größe deiner Organisation ist kein Grund, damit zu warten. Sie ist der Grund, heute damit anzufangen.

> ***Was würde deine Organisation heute aufschreiben, wenn sie ihre bereits gelebte KI-Ethik explizit machen müsste?***

Der erste Schritt: Ein Leitfaden, keine Blaupause

Die meisten Organisationen haben längst eine gelebte KI-Praxis — nur hat sie niemand aufgeschrieben, und niemand hat sie bewusst gewählt. Der folgende Ablauf ist keine fertige Lösung. Er ist der Weg, auf dem eine Organisation ihre eigene, passende Antwort findet.

Schritt 0 · Die eigenen Annahmen prüfen. Bevor irgendetwas erhoben oder entschieden wird: eine kurze, ehrliche Runde mit Führung und Belegschaft gemeinsam. Welche Sätze über KI halten wir gerade für Tatsachen, die eigentlich nur Annahmen sind? „KI kann das noch nicht." „Wenn wir nicht mitziehen, sind wir raus." „Das ist Sache der IT, nicht meine." *Konkrete Maßnahme:* eine 45-minütige Runde, in der alle reihum eine Sicht einbringen — anschließend gemeinsam prüfen, ob sie belegt ist oder nur eine Vermutung, die den nächsten Schritt schon verzerrt.

Schritt 1 · Den Status erfassen — ohne jemanden bloßzustellen. Diese Stelle ist die heikelste im ganzen Prozess — sie entscheidet, ob alles Weitere ehrlich wird oder nicht. Was garantiert scheitert: Einzelgespräche, Monitoring-Software oder eine Umfrage, die erkennen lässt, wer geantwortet hat. Sobald Mitarbeitende Konsequenzen befürchten, bekommt man nur noch Antworten, die niemandem schaden. Was funktioniert: eine anonyme, freiwillige Kurzumfrage (5–8 Fragen, 5 Minuten) zu tatsächlicher KI-Nutzung; eine ausdrückliche Amnestie-Ansage vorab, schriftlich, von der Geschäftsführung selbst — bisherige Nutzung hat keine Konsequenzen, auch wenn sie nie offiziell erlaubt war; und die Analogie, die tatsächlich hilft: Eine Standortbestimmung ist wie ein Gesundheits-Check, nicht wie eine Prüfung. Zeigt die Umfrage mehr Nutzung, als Führungskräfte erwartet hätten, hat sie funktioniert. Zeigt sie erstaunlich wenig, war die Amnestie vermutlich nicht glaubwürdig genug — was Hinweise auf andere Probleme im Betriebssystem der Organisation gibt. Ein zweiter, unbequemer Test, besonders wichtig, wenn gerade Stellen abgebaut werden: Frag bei jeder KI-begründeten Restrukturierung, was sich operativ tatsächlich geändert hat. Besteht die Antwort nur aus Toollizenz und Pilotprojekt, ist es eine Kostenentscheidung im KI-Kostüm — keine echte KI-Entscheidung.

Schritt 2 · Die eigene Richtung klären. Getrennter Termin, dieselbe Gruppe. Nicht mehr „was ist möglich", sondern „was wollen wir" — mit Blick auf Wettbewerb, Kundenerwartung, die eigene Belegschaft, gesellschaftliche Wirkung. *Konkrete Maßnahme:* ein Workshop (2–3 Stunden) mit einer einfachen Frage pro Perspektive — Unternehmen, Organisation, Belegschaft, Kunden, Gesellschaft: Was wäre hier „gut", wenn wir KI ernsthaft einsetzen?

Schritt 3 · Gemeinsam Grenzen ziehen. Jetzt, und nicht früher, kommt die eigentliche Regelsetzung — gemeinsam mit den Menschen, die es betrifft, nicht von der Führung allein verordnet. Jeder bereits genutzte oder geplante KI-Anwendungsfall wird eingeordnet: Darf KI das allein entscheiden? Nur vorschlagen? Nie? Das Ergebnis passt auf eine Seite. Eine feste

Regel, die nicht verhandelbar ist: Entscheidungen über Zugehörigkeit, Vergütung oder Entwicklung von Menschen werden nie ausschließlich algorithmisch getroffen. Jemand mit Namen steht für die Begründung ein.

Schritt 4 · Regelmäßig prüfen, ob es noch stimmt. Kein einmaliges Projekt. Ein fester Termin, vierteljährlich: Ist neue, ungeregelte Nutzung aufgetaucht — ein Signal, kein Fehlverhalten? Hat sich die technische Lage verändert? Kann jemand aus der Belegschaft, nicht nur aus der Führung, erklären, warum ein Prozess so eingeordnet ist, wie er es ist?

Was das nicht ist. Kein Zertifikat, kein Siegel, kein einmal abgeschlossenes Projekt. Der Ablauf ist ein Rhythmus, kein Endpunkt — und er sieht in einem Fünf-Personen-Betrieb anders aus als in einem Konzern. Was bleibt, ist die Reihenfolge: erst verstehen, dann wollen, dann regeln, dann prüfen — nie umgekehrt.

Wo dieser Text aufhört

Absichtlich schlank: Wie unterschiedlich die richtigen Antworten ausfallen — je nachdem, ob eine Organisation reguliert ist oder nicht, ob sie direkten Kundenkontakt hat, ob sie fünf oder fünftausend Mitarbeitende zählt —, würde jeden Rahmen sprengen, der noch lesbar bleiben soll.

> Offen bleibt auch eine Frage, die dieser Text nur streift. Die Sachbearbeiter in der Kindergeld-Affäre haben geprüft. Die Prüfung war vorgesehen, sie fand statt, und sie hat nichts verhindert. Was eine Prüfung kostet, wenn eine Maschine in Sekunden vorsortiert, wer diese Kosten heute trägt und ob die Stelle, bei der sie landen, sie überhaupt tragen kann: Darum geht es in meinem Buch **„Wer prüft, zahlt. Die Rechnung hinter der KI-Ersparnis"** (2026). Es ist aus der Arbeit an diesem Text hervorgegangen und führt die Zonen weiter, mit fünf Fragen, die vor jeder Zuordnung stehen, sieben Arbeitsblättern und dem Rechtsstand, der für Unternehmen in Deutschland gilt. Leseprobe, Arbeitsblätter und Belege: organisation-unter-ki.de/wer-prueft-zahlt. Zum Buch: guidobosbach.com/buecher.

Und manches gehört gar nicht mehr in einen Text, sondern in ein Gespräch über deine konkrete Organisation: guidobosbach.com/klarheitsdialog.

Quellenverzeichnis

Prolog

- Niederländische Kindergeld-Affäre („Toeslagenaffaire"): Parlamentarischer Untersuchungsausschuss, *Ongekend onrecht*, Tweede Kamer, 17.12.2020. Amnesty International, *Xenophobic Machines*, Oktober 2021 (Risikomodell mit dem Merkmal niederländische Staatsangehörigkeit ja/nein; Sachbearbeiter sahen nicht, warum ein Fall ausgewählt worden war). Niederländische Regierung, 12.02.2026: mehr als 43.000 Eltern als Geschädigte anerkannt. Rücktritt der Regierung im Januar 2021.
- Guido Bosbach: „Tragfähig. Menschlich. Zukunftsfähig." — guidobosbach.com
- Asilomar-Konferenz zur rekombinanten DNA, 1975.
- Isaac Asimov: die drei Gesetze der Robotik, erstmals in der Kurzgeschichte „Runaround" (1942).
- Yuval Noah Harari, Tanner Lecture, University of Oxford, 2026 (Vortragsmitschnitt) — These: Bürokratie, Recht und Geld sind aus Sprache gebaut, KI ist die erste nicht-menschliche Instanz, die dieses Substrat selbst liest und bedient.
- Anthropic: „Claude's Constitution", Fassung vom 21. Januar 2026 — anthropic.com/constitution, CC0 1.0 („Spalier statt Käfig").

Kapitel 1 — Ethik ist keine Kompetenz, die man kaufen kann

- Agonora: „Ethics — the most important skill" — agonora.com
- Aristoteles: Nikomachische Ethik (Unterscheidung Techne/Phronesis).
- Anthropic: „Claude's Constitution" — anthropic.com/constitution.
- Isabel E. P. Menzies: „A Case-Study in the Functioning of Social Systems as a Defence against Anxiety", in: Human Relations 13 (1960), Heft 2, S. 95–121. https://doi.org/10.1177/001872676001300201

Kapitel 2 — Wer gewinnt?

- Frances Haugen/Facebook, 2021: MIT Technology Review, CBS News
- Wells Fargo: CFPB-Bußgeldentscheidung 2016 (185 Mio. USD).
- Lenin/Iskra, Mussolini/Il Popolo d'Italia, Marat/L'Ami du Peuple: historisch dokumentiert (Britannica).
- Microsoft/MSN, 2020: rund 77 Journalisten und Redakteure entlassen, durch einen Algorithmus zur Story-Auswahl ersetzt — Forbes, Al Jazeera.

Kapitel 3 — Banale Unternehmensandroiden

- Slavoj Žižek, WELT, Rezension zu Aakash Singh Rathore: „Will AI Murder Us All?", 2026.

Kapitel 4 — Die Verantwortung wandert aufwärts

- OpenAI-Test und Hugging Face, Juli 2026: Meldung von Hugging Face vom 16.07.2026, gemeinsame Erklärung von OpenAI und Hugging Face vom 21.07.2026; CNN, 22.07.2026.
- Dennis-Kenji Kipker gegenüber ZDFheute, 23.07.2026 — zdfheute.de
- Christopher Nolan, Zitat zu KI als „Trojanischem Pferd" — TechCrunch, 19.07.2026
- Amazon, Versandlagerhaus Baltimore: Human Resources Manager

- Anthropic: „Claude's Constitution" — anthropic.com/constitution (Verbot der Selbstexfiltration, Sparsamkeitsprinzip bei Ressourcen).

Kapitel 5 — Eine KI-Verfassung für deine Organisation

- IKEA/KI-Bot „Billie": CIO.de
- Zonen A–D: eigenes Rahmenkonzept, ausgeführt in Guido Bosbach: „Wer prüft, zahlt" (2026).
- Anthropic: „Claude's Constitution" — anthropic.com/constitution (Rangfolge, harte Grenzen, Korrigierbarkeit, angekündigtes Widerspruchsverfahren).

Auch dieser Text ist mit intensiver Unterstützung von Künstlicher Intelligenz entstanden. Gelenkt, überarbeitet, angepasst, korrigiert und freigegeben wurde er von einem Menschen — mit Namen, nicht anonym.

Diese Position vertritt dieser Text: KI liefert Material. Wer dafür einsteht, bleibt ein Mensch.

Fassung September 2026 · guidobosbach.com/ki-braucht-klarheit