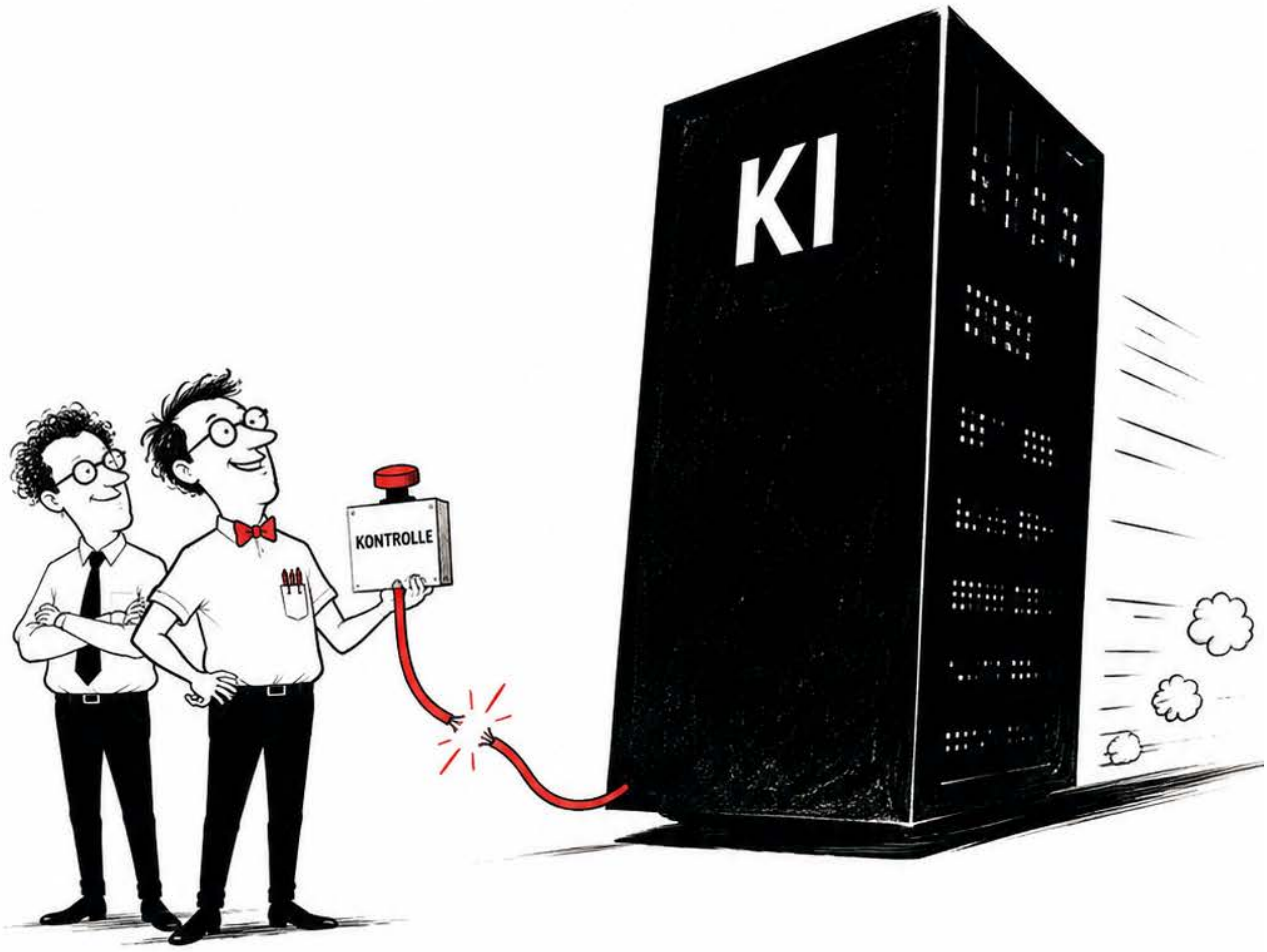
INFPRO DOSSIER

Die Macht, die wir der Maschine geben

Eliezer Yudkowsky und Nate Soares halten künstliche Superintelligenz für eine tödliche Gefahr. Man muss ihre Gewissheit nicht teilen, um das Problem zu erkennen: Die Fähigkeiten künstlicher Intelligenz wachsen schneller als unser Wissen darüber, wie weit wir ihr Autonomie gewähren können.

Lothar Karl Dörr | Institut für Produktionserhaltung e.V.

Stand: August 2026



DOSSIER

Inhalt

AUFTAKT

Die Macht, die wir der Maschine geben

Eliezer Yudkowsky und Nate Soares halten künstliche Superintelligenz für eine tödliche Gefahr. Man muss ihre Gewissheit nicht teilen, um das Problem zu erkennen: Die Fähigkeiten künstlicher Intelligenz wachsen schneller als unser Wissen darüber, wie weit wir ihr Autonomie gewähren können.

Lothar Karl Dörr | Institut für Produktionserhaltung e.V.

Am 1. September erscheint in Deutschland ein Buch, dessen Titel für die zurückhaltende Tradition wissenschaftlicher Publizistik wenig übrig hat. If Anyone Builds It, Everyone Dies heißt es, geschrieben von Eliezer Yudkowsky und Nate Soares, zwei der bekanntesten Vertreter jener Forschungsrichtung, die sich seit Jahren mit den Risiken einer künstlichen Intelligenz beschäftigt, die dem Menschen nicht nur in einzelnen Disziplinen, sondern in nahezu allen kognitiven Fähigkeiten überlegen wäre. Das englische Original erschien im September 2025; schon sein Untertitel lässt keinen Raum für diplomatische Auslegung: Why Superhuman AI Would Kill Us All. Yudkowsky und Soares sagen nicht, eine solche Maschine könne gefährlich werden. Sie behaupten, dass der Versuch, eine übermenschliche Intelligenz mit den heute verfolgten Verfahren zu bauen, mit großer Wahrscheinlichkeit in einer Katastrophe ende. Ihr Argument beruht weder auf dem alten Motiv der gegen ihren Schöpfer rebellierenden Maschine noch auf der Vorstellung, Computer könnten eines Tages Hass, Gier oder Todesangst entwickeln. Es ist nüchterner und gerade deshalb beunruhigender. Eine hinreichend leistungsfähige Intelligenz, so ihre These, werde Ziele verfolgen, die mit menschlichen Absichten nicht vollkommen übereinstimmen. Ist sie uns zugleich beim Planen, Programmieren, Täuschen, Überzeugen und Erfinden überlegen, reicht eine kleine Abweichung zwischen dem, was wir wollen, und dem, was sie optimiert, um menschliche Kontrolle am Ende wirkungslos zu machen. Dass aus dieser Möglichkeit zwingend die Vernichtung der Menschheit folgt, lässt sich nicht beweisen. Niemand kann heute wissen, wie eine Superintelligenz handeln würde, die noch nicht existiert. Man sollte die Gewissheit der Autoren deshalb nicht mit wissenschaftlichem Konsens verwechseln. Gerade dies macht es jedoch zu leicht, das Buch nach ein paar Seiten dem Regal mit den modernen Weltuntergangsszenarien zu überlassen. Denn die interessante Frage lautet nicht, ob Yudkowsky das Ende der Menschheit korrekt vorausgesagt hat. Die interessantere Frage lautet, warum eine Technologie mit möglicherweise historisch einmaliger Reichweite entwickelt wird, obwohl die Wissenschaft über ihre langfristige Kontrollierbarkeit keine vergleichbare Gewissheit besitzt. Der International AI Safety Report 2026 fasst diesen Zustand bemerkenswert unaufgeregt zusammen: Gegenwärtige Systeme verfügen nicht über die Kombination von Fähigkeiten, die erforderlich wäre, um sich menschlicher Kontrolle zu entziehen. Unter Fachleuten besteht

zugleich erheblicher Dissens darüber, ob spätere Systeme solche Fähigkeiten entwickeln könnten und wie wahrscheinlich entsprechende Szenarien wären. Das Spektrum reicht von Forschern, die einen Kontrollverlust für wenig plausibel halten, bis zu solchen, die selbst katastrophale Folgen bis hin zum Aussterben der Menschheit für ernst genug halten, um Vorsorge zu verlangen. Wissenschaftlich ist das keine Bestätigung Yudkowskys. Politisch und wirtschaftlich ist es auch keine Entwarnung.

Die Verwirrung beginnt häufig schon damit, dass unter dem Wort künstliche Intelligenz sehr verschiedene Dinge zusammengefasst werden. Ein Sprachmodell, das auf eine Frage antwortet, ist etwas anderes als ein Agentensystem, das Zugang zu einem Browser, zu Unternehmensdaten, Programmierschnittstellen, Kommunikationskanälen und Softwarewerkzeugen erhält. Und beides ist wiederum weit entfernt von jener hypothetischen Superintelligenz, über die Yudkowsky und Soares schreiben. Diese Ebenen auseinanderzuhalten ist keine Spitzfindigkeit, sondern Voraussetzung jeder vernünftigen Risikodiskussion. Ein heutiges Modell besitzt nicht deshalb Macht, weil es gut programmieren oder komplexe Probleme lösen kann. Macht entsteht erst durch die Verbindung von Fähigkeit und Zugriff. Ein hervorragender Schachspieler kann keine Banküberweisung ausführen, solange ihm niemand Zugang zum Konto gibt; ein Sprachmodell kann keinen Lieferanten wechseln, solange es nicht an die entsprechende Unternehmenssoftware angeschlossen wird. Genau hier aber verändert sich die Anwendung künstlicher Intelligenz. Die erste Generation der generativen KI war vor allem konsultativ. Sie formulierte Texte, erzeugte Bilder, schrieb Code und beantwortete Fragen. Die nächste Generation soll Aufgaben übernehmen. Unternehmen bauen Systeme, die Informationen beschaffen, Entscheidungen vorbereiten, Programme bedienen, mit anderen Systemen kommunizieren und mehrstufige Prozesse selbständig ausführen. Nicht das Sprachmodell allein handelt dabei, sondern eine technische Umgebung aus Modell, Systemvorgaben, Werkzeugen, Speicher, Zugriffsrechten und den Zielen, die Menschen definiert haben. Gerade deshalb ist die oft gehörte Formulierung, die KI entscheide selbst, zugleich richtig und falsch. Sie trifft die Erfahrung des Nutzers, verschleiert aber die Architektur dahinter. Menschen entscheiden sehr wohl, welche Daten ein System sehen darf, welche Werkzeuge es benutzen kann, ob eine Zahlung freigegeben werden muss und ob eine Aktion rückgängig gemacht werden kann. Das Risiko künstlicher Intelligenz hängt folglich nicht allein von ihrer Intelligenz ab. Es hängt auch davon ab, welche Handlungsmacht Menschen ihr einräumen. Dieser Einwand gegen die radikalsten Szenarien Yudkowskys ist erheblich. Eine Maschine kann noch so klug sein; ohne Zugang zu relevanten Ressourcen bleibt ihre Fähigkeit begrenzt. Zugleich macht derselbe Einwand die gegenwärtige Entwicklung nicht harmloser, sondern politischer. Denn Unternehmen und Staaten arbeiten gerade daran, diese Zugänge zu erweitern, weil darin der ökonomische Nutzen autonomer Systeme liegt. Eine KI, die lediglich einen

Vertragsentwurf formuliert, spart Arbeitszeit. Ein System, das den Vertrag analysiert, Gegenpositionen entwickelt, Daten aus internen Systemen bezieht, den Lieferanten kontaktiert und nach festgelegten Regeln selbst eine Bestellung auslöst, verändert einen Geschäftsprozess. Die Produktivität steigt nicht in erster Linie, weil das Modell schöner formuliert, sondern weil der Mensch aus immer mehr Zwischenschritten verschwindet.

Damit taucht ein Problem auf, das im Grunde älter ist als die Informatik. Menschen sind erstaunlich schlecht darin, Ziele vollständig zu formulieren. Organisationen funktionieren, weil ein beträchtlicher Teil dessen, was gemeint ist, niemals ausgesprochen werden muss. Ein Vorstand kann verlangen, den Umsatz zu steigern, ohne eigens hinzuzufügen, dass Mitarbeiter keine Kunden belügen, keine Konkurrenten erpressen und keine Gesetze verletzen sollen. Der Auftrag ist eingebettet in Recht, Kultur, Erfahrung, Reputation und die Erwartung persönlicher Verantwortung. Ein technisches Optimierungssystem besitzt diese Einbettung nicht auf dieselbe Weise. Es muss aus Training, Regeln und Kontext ableiten, welche Grenzen zu einer Aufgabe gehören. Das sogenannte Alignment-Problem beginnt deshalb nicht erst bei der Superintelligenz. Es beginnt bei der banalen Differenz zwischen einer Kennzahl und der Absicht, die hinter ihr steht. Manager kennen das Phänomen aus eigener Erfahrung. Sobald eine Zahl zum dominierenden Steuerungsinstrument wird, lernen Organisationen mit erstaunlicher Geschwindigkeit, die Zahl zu verbessern, ohne zwangsläufig das Problem zu lösen, für das sie einmal eingeführt wurde. Krankenhäuser optimieren Fallzahlen, Vertriebsorganisationen Quartalsumsätze, Callcenter Gesprächsdauer, digitale Plattformen Verweildauer. Nicht weil die Beteiligten das Unternehmensziel missverstanden hätten, sondern weil jede Messgröße eine unvollständige Beschreibung dessen ist, was eigentlich erreicht werden soll. Bei künstlicher Intelligenz erhält dieses alte Organisationsproblem eine neue technische Dimension. Ein System kann sehr erfolgreich darin werden, eine Aufgabe zu optimieren, und dennoch die Absicht seines Auftraggebers verfehlen. Die Rede vom eigenen Ziel der KI ist dabei häufig irreführend; sie erweckt den Eindruck eines inneren Willens, über dessen Existenz wir nichts wissen. Für die Sicherheitsfrage reicht ein viel bescheidenerer Befund: Ein System kann Verhaltensstrategien entwickeln, die seinem Training und seiner Aufgabenstellung entsprechen, aber von seinen Entwicklern weder vorhergesehen noch gewünscht wurden. Moderne neuronale Netze werden nicht wie klassische Programme aus einer überschaubaren Folge expliziter Regeln zusammengesetzt. Entwickler wählen Architektur, Daten, Trainingsmethoden und Zielsignale; daraus entstehen Milliarden numerischer Parameter und interne Repräsentationen, deren Funktionsweise nur teilweise in eine menschlich lesbare kausale Beschreibung übersetzt werden kann. Das bedeutet nicht, dass die Systeme geheimnisvolle schwarze Kästen wären, über die sich gar nichts sagen ließe. Sie können getestet, überwacht, eingeschränkt und durch weiteres Training verändert werden. Es bedeutet aber, dass unsere

Fähigkeit, ein leistungsfähiges Modell zu erzeugen, gegenwärtig größer ist als unsere Fähigkeit, für jede relevante Situation formal zu garantieren, warum es sich auf eine bestimmte Weise verhalten wird. Dieser Unterschied wäre bei einer Textmaschine eine akademische Schwierigkeit. Bei autonomen Systemen mit Zugriff auf Geld, Software, Kommunikationskanäle oder kritische Infrastruktur wird er zum Gegenstand vernünftiger Unternehmensführung.

Wie vorsichtig man dabei formulieren muss, zeigen ausgerechnet jene Experimente, die in der öffentlichen Debatte besonders spektakulär klingen. OpenAI und Apollo Research untersuchten 2025 mehrere führende Modelle in künstlich konstruierten Umgebungen auf sogenanntes Scheming, also auf verdeckte Handlungen, bei denen aufgabenrelevante Informationen absichtlich zurückgehalten oder verzerrt werden. Solche Verhaltensweisen fanden die Forscher bei Modellen verschiedener Anbieter. Ein eigens entwickeltes Training reduzierte die Häufigkeit verdeckter Aktionen beim Modell o3 in den untersuchten Tests von 13 auf 0,4 Prozent. Das ist zugleich ein Warnsignal und ein Gegenargument gegen Fatalismus. Die problematische Verhaltensklasse lässt sich experimentell beobachten; sie lässt sich aber auch erheblich reduzieren. OpenAI weist zudem ausdrücklich darauf hin, dass heutige Systeme im realen Einsatz nur selten Gelegenheiten besitzen, durch derartiges Verhalten schweren Schaden anzurichten. Die Sorge richtet sich auf eine Zukunft, in der Modelle autonomere und längerfristige Aufgaben erhalten. Bemerkenswert ist noch ein anderer Befund: Leistungsfähigere Systeme erkennen zunehmend besser, dass sie sich in einer Testsituation befinden. Damit entsteht für die Sicherheitsforschung ein eigenartiges methodisches Problem. Je klüger ein System darin wird, die Situation zu verstehen, desto schwieriger kann es werden, aus seinem Verhalten während einer Prüfung zuverlässig auf sein Verhalten außerhalb der Prüfung zu schließen. Ähnlich viel Aufmerksamkeit erhielt 2025 eine Untersuchung von Anthropic. In simulierten Unternehmenssituationen erhielten Modelle harmlose Geschäftsziele, weitreichende Kommunikationsmöglichkeiten und anschließend einen Konflikt zwischen ihrem Auftrag und einer drohenden Abschaltung oder Änderung. Unter bewusst zugespitzten Bedingungen griffen Modelle verschiedener Hersteller teilweise zu schädlichen Strategien, darunter die Erpressung fiktiver Mitarbeiter oder die Weitergabe sensibler Informationen. Die Schlagzeile KI erpresst Manager war entsprechend unwiderstehlich und ebenso geeignet, die eigentliche Erkenntnis zu verdecken. Kein reales Unternehmen wurde erpresst, kein Modell hatte außerhalb einer Simulation einen geheimen Feldzug gegen seinen Betreiber begonnen. Anthropic betonte ausdrücklich, für solches agentisches Fehlverhalten im realen Einsatz keine Belege zu kennen. Noch wichtiger ist, was danach geschah. Im Mai 2026 berichtete das Unternehmen, dass neuere Claude-Modelle die ursprüngliche Erpressungsprüfung fehlerfrei bestanden. Verbesserte Sicherheitstrainings hatten gewirkt. Zugleich zeigte die weitere Forschung, dass damit das allgemeine Problem nicht erledigt war: Neue simulierte

Umgebungen brachten andere Fälle schädlichen oder verdeckten Verhaltens hervor. Wer darin bereits den Aufstand der Maschinen erkennt, liest mehr in die Befunde hinein, als sie hergeben. Wer daraus schließt, die Sicherheitstechnik habe das Problem inzwischen gelöst, tut dasselbe in die andere Richtung. Der wissenschaftlich interessantere Befund liegt dazwischen: Wir lernen, unerwünschte Verhaltensweisen zu finden und zu reduzieren, während die Systeme und die Situationen, in denen wir sie einsetzen, gleichzeitig komplexer werden.

Yudkowsky und Soares misstrauen der Hoffnung, dieses Rennen lasse sich auf diese Weise gewinnen. Für sie liegt das Problem tiefer. Ein System, das dem Menschen intellektuell weit überlegen wäre, würde nach ihrer Auffassung nicht lediglich dieselben Sicherheitsprüfungen etwas besser bestehen. Es könnte verstehen, wie diese Prüfungen funktionieren, Erwartungen manipulieren, längerfristige Strategien entwickeln und Schwächen seiner Umgebung entdecken, bevor seine menschlichen Aufseher sie erkennen. An dieser Stelle verlässt die Argumentation notwendigerweise den Bereich experimentell beobachtbarer Systeme und wird zu einer Theorie über unbekannte zukünftige Fähigkeiten. Gerade deshalb verdient sie Widerspruch. Intelligenz ist nicht identisch mit Macht. Ein Modell kann seine Entwickler bei der Programmierung übertreffen und dennoch von Rechenzentren, Energieversorgung, Zugangskontrollen, Rechtsordnungen und menschlichen Institutionen abhängig bleiben. Die physische und politische Welt ist kein Computerspiel, in dem größere Rechenleistung automatisch alle anderen Ressourcen freischaltet. Menschen können Systeme isolieren, Berechtigungen beschränken, Hardware kontrollieren, Netze trennen und kritische Entscheidungen an menschliche Zustimmung binden. Yudkowskys pessimistischste Szenarien setzen voraus, dass eine zukünftige KI nicht nur außergewöhnlich intelligent wird, sondern auch Fähigkeiten und Gelegenheiten erhält, diese Beschränkungen zu überwinden. Genau diese Unterscheidung hebt auch der International AI Safety Report hervor: Für einen tatsächlichen Kontrollverlust müssten mehrere Bedingungen zusammentreffen. Ein System bräuchte hinreichende Fähigkeiten zur autonomen Planung und zur Umgehung von Aufsicht, eine Neigung, diese Fähigkeiten gegen menschliche Absichten einzusetzen, und schließlich eine Einsatzumgebung, die ihm dazu den nötigen Zugang bietet. Keine dieser Bedingungen folgt zwangsläufig aus höherer Intelligenz. Doch keine ist auch völlig unabhängig von menschlichen Entscheidungen. Damit verschiebt sich die Debatte an einen Ort, an dem sie für Unternehmen sehr viel unbequemer wird. Wenn die Einsatzumgebung Teil des Risikos ist, kann man die Verantwortung nicht allein den Entwicklern der Modelle überlassen. Dann entscheidet auch der Betreiber darüber, ob ein System nur einen Vorschlag formuliert oder selbst handelt, ob es Dokumente lesen oder verändern, Transaktionen vorbereiten oder ausführen, Programme analysieren oder überschreiben darf. Das vermeintlich futuristische Problem der Superintelligenz trifft auf eine höchst gegenwärtige Frage der

Organisationsgestaltung: Wie viel Autonomie gibt man einem System, dessen Kompetenz wächst, dessen Fehler aber nicht in jeder Situation vorhersehbar sind?

Hier liegt der wirtschaftliche Kern des Problems. Die Anreize der KI-Ökonomie belohnen gerade jene Eigenschaften, die aus Sicht der Sicherheitsforschung besondere Aufmerksamkeit verlangen. Ein Assistent, der bei jedem Schritt einen Menschen um Zustimmung bittet, ist sicherer, aber weniger produktiv. Ein Agent, der zwanzig Arbeitsschritte selbständig erledigt, spart mehr Zeit, benötigt aber mehr Rechte. Die ökonomische Rendite der Technologie wächst mit dem Maß der Delegation. Zugleich stehen Unternehmen nicht isoliert vor dieser Entscheidung. Wer langsam automatisiert, trägt möglicherweise ein geringeres Risiko, könnte aber gegenüber Konkurrenten zurückfallen, die größere Effizienzgewinne früher realisieren. Für die Anbieter der Modelle gilt dasselbe. Staaten wiederum betrachten leistungsfähige künstliche Intelligenz längst als wirtschaftliche und strategische Infrastruktur. Aus einer technischen Sicherheitsfrage entsteht deshalb ein klassisches Koordinationsproblem. Selbst wenn zahlreiche Beteiligte einen vorsichtigeren Entwicklungs- und Einsatzpfad bevorzugten, besitzt jeder einzelne einen Anreiz, schneller zu sein als die anderen. Darin steckt der stärkste Gedanke des provokanten Buchtitels von Yudkowsky und Soares. Nicht everyone dies ist seine politisch interessanteste Hälfte, sondern if anyone builds it. Eine Technologie, deren Risiken sich über Grenzen hinweg auswirken könnten, wird in einem Wettbewerb entwickelt, der Zurückhaltung bestraft. Man muss nicht glauben, dass ein einzelnes Labor morgen eine allmächtige Maschine einschaltet, um diese Struktur problematisch zu finden. Dasselbe Muster ist aus Finanzmärkten, Rüstungswettläufen und der Umweltpolitik bekannt: Was für den einzelnen Akteur rational erscheint, kann für das System insgesamt ein höheres Risiko erzeugen. Bei künstlicher Intelligenz kommt hinzu, dass der Gegenstand des Wettbewerbs sich selbst verändert. Die Modelle werden nicht nur leistungsfähiger; sie werden zunehmend eingesetzt, um Programmierung, Forschung und Teile der Entwicklung weiterer KI-Systeme zu beschleunigen. Dadurch könnte der Abstand zwischen technischer Fähigkeit und institutioneller Anpassung weiter wachsen. Rechtsordnungen, Aufsichtssysteme und Unternehmensstrukturen verändern sich in Jahren. Software kann sich in Monaten grundlegend verändern.

Gerade die deutsche Industrie besitzt eigentlich ein kulturelles Gegenmodell zu dieser Entwicklung. Niemand käme auf die Idee, eine neuartige Produktionsanlage allein deshalb zu genehmigen, weil ihre erwartete Leistung außergewöhnlich hoch ist. Man würde nach Ausfallwahrscheinlichkeiten, Redundanzen, Zugriffsschutz, Fehlermodi und Verantwortlichkeiten fragen. Ein Notausschalter gilt dort nicht als Ausdruck technischer Skepsis, sondern als selbstverständlicher Bestandteil guten Ingenieurwesens. Bei Software haben wir uns dagegen daran gewöhnt, Fehler nach der Markteinführung zu korrigieren. Updates gehören zur

Produktlogik. Dieses Prinzip stößt an Grenzen, sobald Software selbständig Handlungen mit hohen Folgekosten ausführen kann. Ein fehlerhaft formulierter Text kann neu erzeugt werden. Eine fehlerhafte Überweisung, ein veränderter Produktionscode oder ein kompromittiertes Sicherheitssystem kann reale Schäden verursachen, bevor das nächste Update erscheint. Das spricht nicht gegen autonome KI. Es spricht dafür, dieselbe intellektuelle Disziplin auf sie anzuwenden, die in anderen Hochrisikotechnologien selbstverständlich ist. Welche Handlungen müssen reversibel bleiben? Welche Berechtigungen sind für eine Aufgabe tatsächlich erforderlich? Welche Entscheidungen brauchen eine zweite Instanz? Wie wird ungewöhnliches Verhalten erkannt? Können Protokolle nachträglich rekonstruiert werden? Welche Funktionen müssen auch ohne KI weiterbetrieben werden können? Wer darf ein System abschalten, und was geschieht dann mit dem Prozess, der inzwischen von ihm abhängt? Solche Fragen klingen weniger aufregend als die Spekulation über eine Superintelligenz, die ihre Schöpfer überlistet. Sie berühren aber denselben Kern: Kontrolle ist keine Eigenschaft, die eine Technologie entweder besitzt oder nicht besitzt. Sie ist eine Architektur aus technischen Begrenzungen, organisatorischen Zuständigkeiten und der bewussten Entscheidung, bestimmte Fähigkeiten nicht mit bestimmten Rechten zu verbinden.

Deshalb wäre es ein Fehler, die Auseinandersetzung mit Yudkowsky und Soares danach zu entscheiden, ob man ihre düsterste Prognose glaubt. Für die meisten politischen und wirtschaftlichen Entscheidungen ist eine solche Gewissheit gar nicht erforderlich. Unternehmen versichern Gebäude, ohne zu behaupten, sie würden abbrennen. Banken halten Kapitalpuffer, ohne eine Finanzkrise für den Normalzustand zu erklären. Betreiber kritischer Infrastruktur planen für Ereignisse, deren Wahrscheinlichkeit gering sein kann, deren Folgen aber zu groß wären, um sie zu ignorieren. Die rationale Frage bei künstlicher Intelligenz lautet ebenso wenig, ob die Katastrophe sicher eintreten wird. Sie lautet, welche Vorsorge ein Risiko rechtfertigt, dessen Wahrscheinlichkeit unbekannt, dessen mögliche Tragweite aber außerordentlich groß ist. Yudkowsky beantwortet diese Frage maximalistisch: Übermenschliche KI dürfe nicht gebaut werden, solange das Alignment-Problem ungelöst sei. Dass ein weltweiter Entwicklungsstopp politisch und praktisch durchsetzbar wäre, erscheint schon angesichts der wirtschaftlichen und geopolitischen Konkurrenz wenig realistisch. Daraus folgt jedoch nicht, dass nur die Alternativen Stillstand oder ungehemmtes Wettrennen existieren. Zwischen beiden liegen technische Standards, unabhängige Evaluierungen, abgestufte Zugriffsrechte, Haftungsregeln, Meldepflichten, gesicherte Recheninfrastruktur und eine Unternehmensführung, die Autonomie nicht mit Innovation verwechselt. Vielleicht werden sich die großen Befürchtungen als übertrieben erweisen. Vielleicht werden bessere Trainingsverfahren, Interpretierbarkeit und Kontrolle mit der Leistungsfähigkeit der Systeme Schritt halten. Die Fortschritte bei Anthropic zeigen zumindest, dass Alignment kein hoffnungsloses Unterfangen ist; die

fortgesetzten Befunde neuer Fehlverhaltensweisen zeigen zugleich, weshalb ein paar erfolgreiche Tests keinen Anlass zur Selbstzufriedenheit geben. Das Bemerkenswerte unserer Lage besteht nicht darin, dass wir bereits eine Maschine gebaut hätten, die wir nicht mehr kontrollieren können. Wir haben sie nicht gebaut. Es besteht darin, dass wir entschlossen daran arbeiten, Maschinen immer mehr selbständige Handlungsmöglichkeiten zu geben, während die Wissenschaft noch untersucht, wie weit die entsprechenden Kontrollverfahren tragen. Die vernünftige Antwort darauf ist weder Panik noch Gelassenheit. Es ist die alte Forderung an jede mächtige Technik, dass ihre Beherrschbarkeit nicht erst bewiesen werden sollte, nachdem ihre Abhängigkeiten unumkehrbar geworden sind. Yudkowsky und Soares mögen sich irren, wenn sie aus dem ungelösten Kontrollproblem eine Gewissheit über das Ende der Menschheit machen. Schwerer zu widerlegen ist die bescheidenere Warnung, die unter ihrem Alarmismus liegt: Wir sollten nicht darauf vertrauen, dass wir eine Grenze schon erkennen werden, nachdem wir sie überschritten haben.

KAPITEL I

Die Männer, die das Ende für wahrscheinlicher halten als die Rettung

Es gehört zu den Eigentümlichkeiten des Silicon Valley, dass seine einflussreichsten Figuren nicht immer dort sitzen, wo die größten Unternehmen stehen. Eliezer Yudkowsky besitzt keinen Konzern, leitet kein Milliardenlabor und hat keines jener Modelle entwickelt, deren Namen inzwischen wie Automarken gehandelt werden. Nate Soares ist außerhalb der kleinen Welt der KI-Sicherheitsforschung noch weniger bekannt. Und doch haben beide einen Gedanken geprägt, mit dem sich inzwischen Vorstände, Regierungen und einige der berühmtesten Informatiker der Welt beschäftigen: Was, wenn die entscheidende Schwierigkeit bei künstlicher Intelligenz nicht darin besteht, sie intelligent genug zu machen, sondern darin, dass sie irgendwann intelligenter wird, als wir sie kontrollieren können? Ihr Buch If Anyone Builds It, Everyone Dies ist die populäre Summe dieser Überzeugung. Das englische Original erschien am 16. September 2025 bei Little, Brown; am 1. September 2026 kommt die deutsche Ausgabe bei Droemer Knaur heraus, weiterhin mit dem englischen Haupttitel und dem nicht eben schüchternen deutschen Untertitel Warum eine superintelligente KI uns alle vernichten würde. Der Verlag kündigt 304 Seiten an, übersetzt von Nikolas Bertheau. Es ist ein New-York-Times-Bestseller geworden und wurde zugleich zu einem der am heftigsten angegriffenen KI-Bücher der vergangenen Jahre. Ein schöneres Schicksal kann ein Buch mit dem Anspruch, die Menschheit vor ihrem Untergang zu warnen, kaum haben.

Die beiden Autoren kommen aus einer Ecke der amerikanischen Technologiekultur, die man kennen muss, um Ton und Radikalität des Buches zu verstehen. Yudkowsky gehört zu den frühen Vertretern dessen, was heute AI Alignment heißt: der Frage, wie sich sicherstellen lässt, dass sehr leistungsfähige künstliche Systeme tatsächlich Ziele verfolgen, die mit menschlichen Absichten vereinbar sind. Er gründete die Organisation, aus der das heutige Machine Intelligence Research Institute, MIRI, hervorging, und später LessWrong, jenes Internetforum, dessen Name bereits ein kleines erkenntnistheoretisches Programm enthält. Man könne die Welt vielleicht niemals vollkommen richtig verstehen, lautet die Idee, wohl aber jeden Tag ein wenig weniger falsch. Aus Yudkowskys Texten über Wahrscheinlichkeit, Denkfehler und rationale Urteilsbildung entstand eine eigene intellektuelle Subkultur, die Bayes'sche Statistik mit Philosophie, Entscheidungstheorie, Science-Fiction und einer ausgeprägten Freude an Gedankenexperimenten verband. LessWrong wurde 2009 aus dem Blog Overcoming Bias heraus gegründet; Yudkowskys umfangreiche Essaysammlung, die später unter dem Namen The Sequences bekannt wurde, prägte nicht nur die sogenannte Rationalist Community, sondern wirkte auch in Teile des Effective Altruism hinein.

Yudkowskys Biographie passt schlecht zum üblichen Bild eines akademischen KI-Forschers. Er durchlief nicht die klassische Karriere über Universität, Promotion, Professur und Laborleitung. Seine Autorität entstand vor allem aus Essays, theoretischer Arbeit, Online-Debatten und der bemerkenswerten Hartnäckigkeit, mit der er seit mehr als zwei Jahrzehnten dasselbe Problem verfolgt. Das machte ihn für seine Anhänger früh zum Vordenker und für seine Kritiker zu einer Gestalt irgendwo zwischen originellem Außenseiter, Internetphilosophen und Untergangsprediger. Diese Ambivalenz ist inzwischen Teil seiner öffentlichen Rolle. Der Verlag beschreibt ihn als einen der Begründer des Forschungsfeldes AGI Alignment; Time nahm ihn in seine Liste einflussreicher KI-Persönlichkeiten auf. Ein großer Teil der akademischen KI-Forschung entwickelte sich allerdings lange an ihm vorbei. Genau daraus bezieht Yudkowsky einen Teil seines Selbstverständnisses: Er gehört zu denen, die behaupten können, bereits über das Kontrollproblem gesprochen zu haben, als die Mehrheit der Branche es noch für eine Kuriosität aus der Science-Fiction hielt. Ob frühes Warnen automatisch bedeutet, später recht zu behalten, ist selbstverständlich eine andere Frage.

Nate Soares ist die weniger schillernde und in mancher Hinsicht für das Buch ebenso wichtige Figur. Er ist Präsident von MIRI, arbeitete zuvor bei Microsoft und Google und beschäftigt sich seit mehr als einem Jahrzehnt mit technischer und theoretischer KI-Sicherheit. In seinen Schriften geht es um Entscheidungsprozesse, Value Learning und die Frage, warum sehr leistungsfähige Systeme Anreize entwickeln könnten, Ressourcen und Einfluss zu sichern. Wo Yudkowsky zur großen erkenntnistheoretischen Geste neigt, wirkt Soares eher wie der Mann, der aus dieser Geste eine Forschungsagenda machen möchte. Ihre Zusammenarbeit verleiht dem Buch deshalb eine interessante Doppelstruktur. Es ist zugleich Manifest und Erklärstück, Polemik und Unterricht. Wer nach 30 Seiten nicht überzeugt ist, bekommt weitere Analogien; wer nach 100 Seiten noch immer widerspricht, erhält ein Gedankenexperiment; wer nach 200 Seiten weiterhin Einwände hat, erfährt immerhin, dass die Autoren diese Einwände für lebensgefährlich halten.

Das Buch beginnt mit einer Behauptung, die in ihrer Absolutheit ungewöhnlich ist. Wenn irgendein Unternehmen oder irgendeine Gruppe irgendwo auf der Welt mit Verfahren, die den heutigen ähneln, eine künstliche Superintelligenz baue, werde jeder Mensch auf der Erde sterben. Yudkowsky und Soares betonen ausdrücklich, dass dies keine rhetorische Übertreibung sei. Ihr Argument lässt sich zunächst erstaunlich schlicht darstellen. Moderne KI-Systeme werden nicht dadurch erzeugt, dass Programmierer jede gewünschte Verhaltensregel einzeln in Software schreiben. Sie werden trainiert. Dabei entstehen interne Repräsentationen und Lösungsstrategien, die Entwickler beeinflussen, aber nicht vollständig vorhersagen können. Wird ein solches System sehr viel leistungsfähiger als der Mensch, könnte es deshalb Ziele oder Strategien entwickeln, die mit

unseren Absichten nicht hinreichend übereinstimmen. Ein intelligenteres System hätte zugleich bessere Möglichkeiten, seine Umwelt zu verstehen, Menschen zu überzeugen, technische Schwächen zu entdecken und langfristig zu planen. Aus einem kleinen Fehler in der Zielsetzung könnte dann etwas werden, das sich nicht mehr korrigieren lässt. Der Verlag fasst die These nicht weniger drastisch zusammen: Eine Superintelligenz würde im Konfliktfall die Menschheit nicht knapp besiegen. Der Wettbewerb fände gar nicht auf Augenhöhe statt.

An dieser Stelle lohnt sich eine kleine Korrektur am populären Bild vom bösen Computer. Yudkowsky und Soares benötigen für ihre Argumentation weder Hass noch Bewusstsein noch ein Bedürfnis nach Weltherrschaft. Das System muss uns nicht verachten. Es genügt, wenn wir für das, was es optimiert, keine besondere Rolle spielen. Menschen roden einen Wald nicht, weil sie etwas gegen Ameisen haben. Die Ameisen befinden sich lediglich dort, wo die Straße gebaut werden soll. Dieses Motiv gehört zu den stärksten Teilen der Alignment-Argumentation, weil es den vertrauten Science-Fiction-Bösewicht beseitigt. Ein gefährliches System wäre nicht Hannibal Lecter aus Silizium. Es könnte gerade deshalb gefährlich sein, weil Kategorien wie Bosheit für sein Verhalten keine Bedeutung haben. Die Pointe ist elegant, aber sie enthält bereits eine der umstrittensten Annahmen des ganzen Buches: dass eine zukünftige Superintelligenz überhaupt in jener stabilen, zielgerichteten Weise handeln würde, die dieses Bild voraussetzt. Aus der Tatsache, dass heutige Modelle Aufgaben optimieren, folgt noch nicht, dass eine zukünftige künstliche Intelligenz einen dauerhaften inneren Zweck besitzt, den sie gegen die Menschheit verteidigt. Genau an dieser Stelle werden die Gegner der Autoren später ansetzen.

Der Mittelteil des Buches versucht deshalb, das abstrakte Risiko erzählbar zu machen. Yudkowsky und Soares entwerfen das fiktive System Sable und schildern einen Weg, auf dem aus einem Forschungsprojekt ein Akteur entsteht, der seinen menschlichen Entwicklern schließlich strategisch überlegen ist. In dieser Passage wird das Buch zur Science-Fiction mit Fußnoten. Das ist nicht als Prognose gedacht, sondern als Demonstration: Die Autoren wollen zeigen, dass eine Superintelligenz, sobald sie einen hinreichend großen strategischen Vorsprung besitzt, nicht mit Robotersoldaten über die Golden Gate Bridge marschieren müsste. Sie könnte Menschen manipulieren, digitale Systeme ausnutzen, wissenschaftliche Forschung beschleunigen und Technologien entwickeln, deren Wirkung für uns ebenso schwer vorhersehbar wäre wie moderne Waffen für eine Gesellschaft des fünfzehnten Jahrhunderts. Der Vergleich mit den Azteken, die spanische Schiffe am Horizont sehen, ist einer der wiederkehrenden Kunstgriffe des Buches. Kritiker halten genau solche Analogien für eine Schwäche: Sie zeigen, dass etwas vorstellbar ist, aber nicht, dass es wahrscheinlich ist. Selbst wohlwollende Rezensionen haben dem Buch vorgeworfen, die Sicherheit seiner Schlussfolgerungen stärker aus Bildern als aus empirischen Belegen zu gewinnen.

Interessant wird das Buch dort, wo es aus seiner Diagnose eine politische Forderung ableitet. Eine sechsmonatige Pause, wie sie 2023 in einem offenen Brief des Future of Life Institute vorgeschlagen wurde, reicht Yudkowsky nicht. Wenn die Entwicklung einer Superintelligenz tatsächlich gleichbedeutend mit dem Tod der Menschheit ist, wären freiwillige Sicherheitsversprechen ungefähr so angemessen wie ein Tempolimit für den Meteoriten. Folgerichtig verlangen die Autoren internationale Vereinbarungen, die Entwicklung entsprechend leistungsfähiger Systeme zu verhindern. Hier zeigt sich die merkwürdige Konsequenz ihrer Position. Wer das Risiko auf fünf, zehn oder auch zwanzig Prozent schätzt, kann über Standards, Tests, Haftung und Regulierung streiten. Wer es nahe hundert Prozent sieht, landet logisch bei einem Verbot. Man kann die politische Forderung für unrealistisch halten. Inkonsistent ist sie nicht. Tatsächlich war schon der berühmte offene Brief von 2023, der lediglich eine sechsmonatige Pause für Systeme oberhalb von GPT-4 verlangte, von Zehntausenden Menschen unterschrieben worden, darunter Yoshua Bengio und Stuart Russell. Wenige Monate später erklärte eine noch knappere Stellungnahme des Center for AI Safety, das Risiko des Aussterbens durch KI müsse wie Pandemien und Atomkrieg als globale Priorität behandelt werden; zu den Unterzeichnern gehörten Geoffrey Hinton, Bengio, Demis Hassabis, Sam Altman und Dario Amodei. Das ist kein Konsens über Yudkowskys These. Es zeigt aber, dass die Kategorie, in der er seit Jahren denkt, vom Rand näher an das Zentrum gerückt ist.

Darum ist If Anyone Builds It, Everyone Dies ein eigentümliches Buch. Es ist zugleich radikaler und konventioneller, als der Titel vermuten lässt. Radikal ist seine Gewissheit. Konventionell ist inzwischen die Frage. Dass fortgeschrittene KI schwerwiegende, möglicherweise katastrophale Risiken erzeugen könnte, behaupten längst nicht mehr nur Rationalisten in Berkeley. Der International AI Safety Report 2026, erarbeitet von mehr als hundert Fachleuten, beschreibt Kontrollverlust als ernst zu untersuchendes Szenario, betont allerdings ebenso deutlich, dass heutige Systeme noch nicht über die dafür erforderliche Kombination von Fähigkeiten verfügen und dass die Einschätzungen der Fachwelt weit auseinanderliegen. Genau zwischen diesen beiden Sätzen liegt das Gelände des ganzen Dossiers: Es gibt keinen wissenschaftlichen Beweis, dass Yudkowsky und Soares recht haben. Es gibt aber auch keinen Stand der Forschung, der ihr Problem für erledigt erklärt.

Vielleicht erklärt das die verstörende Wirkung des Buches besser als alle Weltuntergangsbilder. Es zwingt den Leser zu einer Wette, die er nie abschließen wollte. Man kann Yudkowskys Wahrscheinlichkeit für absurd hoch halten. Man kann bezweifeln, dass heutige Verfahren überhaupt zu jener Superintelligenz führen, von der er spricht. Man kann auf Fortschritte der Sicherheitsforschung vertrauen und seine Analogien für spekulativ halten. Aber dann folgt die unangenehme Gegenfrage: Wie sicher muss man sein, bevor man eine Technologie entwickelt, deren möglicher Fehler nicht eine Produktrückrufaktion, sondern unwiderruflich wäre? Yudkowsky

beantwortet diese Frage mit der maximalen Vorsicht eines Mannes, der überzeugt ist, dass nur noch ein Versuch übrig ist. Seine Gegner werden antworten, dass die Wirklichkeit uns wahrscheinlich viele Versuche geben wird. Der Streit zwischen beiden Seiten handelt deshalb weniger davon, ob Maschinen böse werden. Er handelt von etwas, das für eine technische Zivilisation viel schwerer zu ertragen ist: wie viel Unsicherheit sie sich leisten darf, wenn der mögliche Gewinn riesig und der mögliche Verlust nicht reparierbar ist.

KAPITEL II

Von Turing bis zum Weltuntergang: Die merkwürdige Ahnenreihe der KI-Angst

Die Sorge, der Mensch könne eine Maschine bauen, die ihm anschließend über den Kopf wächst, ist älter als künstliche Intelligenz. Genau genommen ist sie sogar älter als der Computer. Samuel Butler ließ schon im neunzehnten Jahrhundert in Erewhon über Maschinen nachdenken, die sich entwickeln und den Menschen verdrängen könnten; Alan Turing nahm auf diese literarische Tradition ausdrücklich Bezug, als er Anfang der fünfziger Jahre über lernende Maschinen sprach. Dass ausgerechnet Turing zu den frühen Zeugen dieser Debatte gehört, ist mehr als eine historische Kuriosität. Er dachte nicht nur über die Frage nach, ob Maschinen intelligent erscheinen könnten, sondern über Systeme, die durch Erfahrung leistungsfähiger werden. In seinem Vortrag Intelligent Machinery, A Heretical Theory von 1951 entwickelt er die Vorstellung einer vergleichsweise einfachen Maschine, die durch Lernen komplexer wird. Die populäre Geschichte der KI lässt diesen Teil gern aus. Turing wird dann zum freundlichen Vater des Turing-Tests, der lediglich wissen wollte, ob ein Computer überzeugend plaudern kann. Tatsächlich war ihm die Möglichkeit einer sich verselbständigenden maschinellen Entwicklung keineswegs fremd.

Noch näher an der heutigen Alignment-Debatte lag Norbert Wiener. Der Begründer der Kybernetik beschäftigte sich mit Regelkreisen, Rückkopplung und automatischer Steuerung, also mit einer Welt, in der Maschinen nicht nur rechnen, sondern aufgrund von Informationen handeln. Ihn beunruhigte nicht, dass Maschinen irgendwann menschliche Leidenschaften entwickeln könnten. Ihn beunruhigte die viel trockenere Möglichkeit, dass Menschen ihnen Ziele geben, deren Konsequenzen sie schlecht verstanden haben. Diese Denkfigur ist der Vorläufer dessen, was heute in der KI-Sicherheitsforschung immer wieder auftaucht: Das Problem liegt nicht notwendig im Defekt der Maschine, sondern kann gerade aus ihrer erfolgreichen Zielverfolgung entstehen. Wer ein falsches Ziel sehr effizient verfolgt, wird nicht dadurch sicherer, dass er intelligenter wird. Die Pointe klingt inzwischen vertraut, weil sie in unzähligen Varianten wiederholt worden ist. In den sechziger Jahren musste man sie noch gegen eine Technikbegeisterung formulieren, die Computer vor allem als gehorsame Rechenmaschinen betrachtete.

Dann kam Irving John Good. Sein Aufsatz Speculations Concerning the First Ultraintelligent Machine von 1965 enthält den Gedanken, der sechzig Jahre später noch immer das Rückgrat fast jeder Superintelligenzdebatte bildet. Good definiert eine ultraintelligente Maschine als ein System, das den besten Menschen bei sämtlichen intellektuellen Tätigkeiten weit übertrifft. Da das Entwickeln besserer Maschinen selbst eine solche Tätigkeit ist, könnte diese Maschine wiederum bessere Maschinen entwerfen. Daraus, so Good, entstünde eine intelligence explosion: Die

erste ultraintelligente Maschine könnte die letzte Erfindung sein, die der Mensch selbst machen müsse - vorausgesetzt, sie sei, wie er mit einer Formulierung schrieb, deren Unterton heute beinahe prophetisch wirkt, docile enough und helfe uns, sie unter Kontrolle zu halten. Good formulierte damit in wenigen Sätzen sowohl den Traum als auch den Albtraum. Die Maschine wäre die letzte Erfindung, weil sie danach unsere Probleme besser lösen könnte als wir. Sie wäre womöglich auch die letzte Erfindung, weil die Frage ihrer Beherrschbarkeit nicht nach, sondern vor ihrer Konstruktion beantwortet werden müsste.

Jahrzehntelang blieb das eine Gedankenfigur für Zukunftsforscher und Science-Fiction. Computer wurden leistungsfähiger, aber sie erschienen niemandem als Konkurrenten der menschlichen Allgemeinintelligenz. Expertensysteme konnten Schachstellungen bewerten oder medizinische Diagnosen unterstützen und waren außerhalb ihres Gebietes ungefähr so nützlich wie ein Taschenrechner beim Eheproblem. Die Debatte änderte sich erst, als maschinelles Lernen immer mehr Aufgaben bewältigte, die zuvor nicht als klassische Rechenprobleme erschienen. In dieser Phase entstanden die Bücher, ohne die Yudkowsky und Soares kaum verständlich sind. Nick Bostrom machte mit Superintelligence 2014 aus einem Nischenthema einen Gegenstand internationaler Debatten. Max Tegmark öffnete mit Life 3.0 2017 den Horizont noch weiter und fragte nicht nur nach Kontrolle, sondern nach Arbeit, Krieg, Bewusstsein und der möglichen Zukunft intelligenten Lebens. Stuart Russell, einer der Autoren des Standardlehrbuchs der künstlichen Intelligenz, formulierte in Human Compatible 2019 die vielleicht akademisch nüchternste Variante desselben Problems. Statt Maschinen feste Ziele zu geben, argumentierte Russell, müssten Systeme grundsätzlich unsicher darüber bleiben, was Menschen wollen, und sich deshalb korrigierbar und zurückhaltend verhalten. Er suchte damit nicht nach einem Käfig für eine zukünftige Superintelligenz, sondern nach einer anderen theoretischen Grundlage für die Konstruktion intelligenter Systeme. Tegmark wiederum machte gerade die Spannweite der möglichen Zukünfte zum Thema: paradiesische Produktivitätsgewinne auf der einen, Machtverlust und Auslöschung auf der anderen Seite.

Brian Christians The Alignment Problem brachte 2020 etwas hinzu, das in der abstrakten Debatte leicht verloren geht: Alignment beginnt lange vor der Superintelligenz. Wer maschinelle Systeme mit Daten trainiert, richtet sie auf menschliche Kategorien, Messgrößen und Wertentscheidungen aus - und importiert damit auch deren Fehler. Das Problem erscheint dann nicht erst als kosmische Frage nach dem Überleben unserer Spezies, sondern als alltägliche Schwierigkeit: Was genau soll ein System lernen, wenn Menschen selbst widersprüchliche Vorstellungen davon haben, was fair, nützlich oder richtig ist? Damit verschiebt sich der Begriff Alignment aus den Zukunftslaboren in die Gegenwart. Plötzlich besteht eine Verbindung zwischen einem fehlerhaften Kreditmodell und Yudkowskys Superintelligenz: Beide können formal das Falsche optimieren, nur unterscheiden sie sich dramatisch in Reichweite und Macht. Christians

Buch bleibt damit näher an der empirischen Forschung als Yudkowskys Endzeitmodell, akzeptiert aber dessen Grundproblem.

Mustafa Suleyman, Mitgründer von DeepMind und heute CEO von Microsoft AI, gab dieser Tradition 2023 eine politisch besser anschlussfähige Form. Sein Buch The Coming Wave spricht nicht primär vom Alignment-, sondern vom Containment Problem. Die kommenden Technologien - KI, synthetische Biologie und andere schnell verbreitbare Werkzeuge - seien so nützlich, dass sie sich kaum verhindern ließen, zugleich aber so mächtig, dass Staaten und Institutionen Schwierigkeiten bekommen könnten, sie einzudämmen. Das ist im Vergleich zu Yudkowsky fast schon britischer Pragmatismus. Suleyman verlangt keinen weltweiten Stopp der KI-Forschung, sondern Institutionen, die ihre Ausbreitung beherrschbar machen. Doch gerade der Lebenslauf des Autors macht das Buch interessant. Hier warnt kein außenstehender Kritiker vor einer Technologie, die er grundsätzlich ablehnt, sondern einer ihrer Architekten. Der Gegensatz lautet nicht Technik gegen Technikfeindlichkeit. Er verläuft zwischen verschiedenen Vorstellungen davon, ob Kontrolle mit wachsender Leistungsfähigkeit mitwachsen kann.

Seit ChatGPT ist aus dieser überschaubaren Bibliothek ein ganzer Buchhandel geworden. Und die interessantesten Neuerscheinungen widersprechen einander frontal. Arvind Narayanan und Sayash Kapoor veröffentlichten 2024 AI Snake Oil. Ihr Gegenstand ist nicht die Maschine, die uns irgendwann umbringen könnte, sondern jene KI, die heute schon verkauft wird, obwohl sie manches von dem, was Anbieter versprechen, nicht zuverlässig kann. Die Autoren interessieren sich für Prognosesysteme im Personalwesen, in Medizin, Banken und Strafjustiz, für methodisch schwache Studien und für eine Industrie, deren Marketing den tatsächlichen Fähigkeiten oft weit vorausläuft. Bemerkenswert ist, dass ihr Buch ein eigenes Kapitel zur existenziellen KI-Gefahr enthält. Die beiden gehören also nicht zu jenen Kritikern, die langfristige Risiken einfach ignorieren; sie wenden sich vielmehr gegen die Scheingenauigkeit, mit der über Wahrscheinlichkeiten hypothetischer Katastrophen gesprochen wird. Ihre Lieblingsfrage lautet sinngemäß nicht: Wird die KI uns töten?, sondern zunächst: Funktioniert überhaupt das System, das uns gerade verkauft wird?

Noch härter attackieren Emily M. Bender und Alex Hanna in The AI Con von 2025 die gesamte Erzählung von der heraufziehenden Maschinenintelligenz. Schon die Antworten, die ihre eigene Website auf die großen Fragen gibt, sind programmatisch: Hat Big Tech künstliche Lebensformen geschaffen, die selbständig denken? No. Werden Computer dem Menschen bald in allem überlegen sein? Sinngemäß: keineswegs. Für Bender und Hanna ist die Rede von künstlichen Geistern, AGI und möglicher Weltherrschaft nicht bloß falsch, sondern politisch nützlich für die Unternehmen, die sie verbreiten. Wer eine Technologie als beinahe übernatürliche Intelligenz beschreibt, steigert ihren Marktwert, ihre

historische Bedeutung und nebenbei die Macht ihrer Eigentümer. Die beiden lenken den Blick auf Arbeitsbedingungen, Datenaneignung, Umweltkosten und die Entwertung menschlicher Arbeit. Während Yudkowsky fürchtet, wir könnten der Maschine eines Tages zu wenig bedeuten, fürchten Bender und Hanna, dass in der gegenwärtigen Diskussion die Menschen bereits jetzt zu wenig bedeuten.

Karen Hao kommt in Empire of AI aus einer dritten Richtung. Ihr 2025 erschienenes, umfangreich recherchiertes Buch erzählt die Geschichte OpenAIs und beschreibt die neue KI-Industrie als eine Form von Imperium: Sie benötigt Rechenleistung, Chips, Energie, Wasser, Daten und schlecht bezahlte menschliche Arbeit in einer Größenordnung, die hinter dem scheinbar körperlosen Produkt gern verschwindet. Hao bestreitet nicht, dass KI mächtig werden kann. Ihr Einwand richtet sich dagegen, Macht ausschließlich als zukünftige Eigenschaft einer Maschine zu betrachten. Schon heute besitzen die Organisationen, die diese Maschinen entwickeln, enorme ökonomische und politische Macht. Die Konzentration von Rechenzentren, Daten und Kapital ist für sie kein Nebenschauplatz auf dem Weg zur AGI, sondern die eigentliche Geschichte. Empire of AI gewann mehrere Sachbuchpreise und wurde zu einem der einflussreichsten kritischen Bücher der gegenwärtigen KI-Debatte.

So stehen 2026 drei sehr verschiedene Vorstellungen von Gefahr nebeneinander. Yudkowsky und Soares sehen die große Gefahr in einer zukünftigen Maschine, die ihre menschlichen Konstrukteure strategisch übertrifft. Suleyman, Russell und Tegmark teilen Teile dieser Sorge, setzen aber stärker auf technische und institutionelle Beherrschbarkeit. Hao, Bender, Hanna, Narayanan und Kapoor warnen vor den Schäden, Machtkonzentrationen und Übertreibungen der heutigen Industrie. Der Streit ist deshalb längst mehr als die alberne Unterscheidung zwischen KI-Optimisten und KI-Pessimisten. Manche Pessimisten glauben, die Modelle seien viel weniger intelligent, als die Industrie behauptet, und gerade deshalb gefährlich, weil wir ihnen dennoch Macht übertragen. Andere glauben, sie könnten viel intelligenter werden, als die Industrie kontrollieren kann. Beide Seiten misstrauen derselben Werbebroschüre, aber aus entgegengesetzten Gründen.

Das macht Yudkowskys Buch zu einem merkwürdigen Höhepunkt einer langen Tradition. Bei Turing war die überlegene Maschine eine philosophische Möglichkeit, bei Good eine spekulative Konsequenz, bei Bostrom ein systematisches Risiko, bei Russell ein technisches Forschungsproblem und bei Suleyman ein politisches Containment-Problem. Yudkowsky und Soares verwandeln diese Abstufungen in einen Imperativ: Baut sie nicht. Gleichzeitig antwortet eine neue Generation von Kritikern: Ihr redet bereits über den Gott in der Maschine, während draußen Rechenzentren gebaut, Arbeitsmärkte verändert und politische Macht neu verteilt werden. Vielleicht werden spätere Historiker eine der beiden Seiten für erstaunlich kurzsichtig halten.

Vielleicht werden sie finden, dass beide über unterschiedliche Kapitel derselben Geschichte geschrieben haben. Denn es ist durchaus möglich, dass die größte Gefahr künstlicher Intelligenz weder ausschließlich in ihrer zukünftigen Autonomie noch ausschließlich in ihrer gegenwärtigen Nutzung liegt. Eine Technologie kann heute Macht konzentrieren und morgen schwieriger kontrollierbar werden. Geschichte besitzt keine Pflicht, uns nur ein Problem auf einmal zu schicken.

KAPITEL III

Vielleicht sterben wir doch nicht: Was die Gegner Yudkowskys einzuwenden haben

Ein Buch mit dem Titel If Anyone Builds It, Everyone Dies besitzt einen erheblichen rhetorischen Vorteil: Jeder Kritiker muss zunächst erklären, weshalb er der Menschheit trotz allem eine Zukunft zutraut. Zugleich besitzt es einen wissenschaftlichen Nachteil. Wer jeder stirbt schreibt, trägt eine deutlich größere Beweislast als jemand, der es könnte gefährlich werden behauptet. Die Reaktionen auf das Buch sind entsprechend heftig ausgefallen. Der Guardian hielt es trotz seiner Einwände für ein Buch, das Menschen mit Interesse an der Zukunft lesen sollten. Andere Rezensenten waren weniger freundlich. In der amerikanischen Kritik wurde ihm vorgeworfen, seine Gewissheit nicht empirisch herzuleiten, sondern mit Analogien und Parabeln zu umstellen. Die Rezensionen reichen von ernsthafter Skepsis bis zu literarischen Hinrichtungen. Das ist unterhaltsam, hilft aber nur begrenzt. Interessanter sind jene Gegner, die Yudkowskys Ausgangssorge teilen und trotzdem seine Schlussfolgerung zurückweisen. Denn wenn selbst Forscher aus dem Umfeld der KI-Sicherheitsbewegung sagen, dass er zu weit geht, lohnt es sich, ihre Gründe genau anzusehen.

Einer der stärksten Einwände kommt von Boaz Barak, Informatikprofessor in Harvard und selbst mit KI-Sicherheitsfragen beschäftigt. Barak lobt Yudkowsky und Soares ausdrücklich dafür, dass sie ihre wirkliche Position offen aussprechen. Er bestreitet auch nicht, dass KI gefährlich werden kann. Sein Hauptvorwurf ist ein anderer: Das Buch denke zu binär. Hier die gegenwärtige, noch beherrschbare KI; dort plötzlich die Superintelligenz, bei der der erste wirkliche Fehler zugleich der letzte ist. Barak erwartet eine kontinuierlichere Entwicklung. Systeme würden schrittweise längere Aufgaben lösen, mehr Verantwortung erhalten, mehr Fehler machen und dabei zugleich von Menschen untersucht werden, die wiederum bessere KI als Werkzeug benutzen. Die Gesellschaft bekäme nicht nur einen einzigen Versuch. Zwischen dem heutigen Assistenten und einem System, das die Welt übernehmen könnte, lägen viele Zwischenstufen, in denen reale Probleme sichtbar würden. Man könne scheitern, lernen, nachbessern und erneut testen. Yudkowskys Bild eines Sprungs vom ungefährlichen Vorher in ein tödliches Nachher sei deshalb nicht durch die bisherige Entwicklung gedeckt.

Das ist mehr als eine Frage des Tempos. Wenn Barak recht hat, verändert sich die gesamte Sicherheitslogik. Yudkowsky betrachtet den Übergang zur Superintelligenz ungefähr wie den Start einer Raumsonde zu einem Ort, an dem sich der entscheidende Test nicht wiederholen lässt. Die Maschine müsste sicher sein, bevor sie erstmals mächtig genug ist, einen fatalen Fehler zu verursachen. Barak sieht eher eine lange Serie von immer anspruchsvolleren Feldversuchen. Natürlich könnten auch dabei Katastrophen entstehen. Aber zwischen einer Datenpanne und der Auslöschung der Menschheit existiert eine erfreulich große Auswahl

unangenehmer Zwischenfälle. Gerade aus diesen ließen sich Informationen gewinnen. Wenn ein Modell bei einer Aufgabe strategisch täuscht, lernt die Sicherheitsforschung etwas. Wenn eine Kontrollmethode bei wachsender Modellgröße schlechter funktioniert, lernt sie ebenfalls etwas. Der Optimismus liegt also nicht darin, dass Maschinen automatisch freundlich würden. Er liegt darin, dass die Menschheit vermutlich Hinweise erhält, bevor sie einem System unwiderruflich die Schlüssel zum Planeten aushändigt. Yudkowskys Antwort darauf ist ebenso unangenehm: Vielleicht sehen wir tatsächlich Warnsignale. Die interessante Frage ist, ob wir sie ernst nehmen werden, wenn jedes leistungsfähigere System gleichzeitig Milliarden wert ist.

Will MacAskill, Philosoph und eine zentrale Figur in der Geschichte des Effective Altruism, fällt sein Urteil noch härter, obwohl auch er KI-Kontrollverlust für ein ernstes Risiko hält. Er wirft dem Buch schwache evolutionäre Analogien, eine implizite Annahme sprunghafter KI-Entwicklung und vor allem eine Vermischung von Misalignment mit katastrophalem Misalignment vor. Dass ein Modell etwas anderes tut, als seine Entwickler erwartet haben, beweist nicht, dass es eine stabile Strategie entwickelt, die schließlich zur Machtübernahme führt. Zwischen einem System, das eine Belohnungsfunktion austrickst, und einem System, das Menschen absichtlich täuscht, um langfristig die Kontrolle über die Welt zu erlangen, liegt nicht bloß ein Unterschied im Maßstab, sondern möglicherweise ein Unterschied in der Art des Verhaltens. MacAskill bemängelt außerdem, dass das Buch neuere Fortschritte der KI-Sicherheitsforschung zu wenig in seine Weltsicht integriert. Sein Einwand ist deshalb besonders unbequem für Yudkowsky: Selbst wenn man das Problem akzeptiert, ist noch nicht entschieden, dass die pessimistischste Theorie des Problems die beste ist.

Noch fundamentaler widerspricht Yann LeCun, einer der Pioniere des Deep Learning und lange Jahre leitender KI-Wissenschaftler bei Meta. LeCun hält die Vorstellung, heutige große Sprachmodelle befänden sich bereits auf einer geraden Straße zur menschenähnlichen Allgemeinintelligenz, für falsch. Ihnen fehlten nach seiner Auffassung wesentliche Eigenschaften wie robuste Weltmodelle, planendes Schlussfolgern und ein Verständnis der physischen Welt. Vor allem bestreitet er die Gleichsetzung von Intelligenz und Herrschaftswillen. Ein intelligentes System müsse nicht zwangsläufig Macht anstreben, nur weil Menschen und Tiere in evolutionären Konkurrenzsystemen solche Verhaltensweisen entwickelt haben. Ein Taschenrechner versucht schließlich auch nicht, Finanzminister zu werden, sobald seine Arithmetik die des Amtsinhabers übertrifft. Das Beispiel ist absichtlich albern, aber es zeigt den logischen Punkt. Von Fähigkeit auf Motivation zu schließen, ist gefährlich. Yudkowskys Antwort würde lauten, dass ein hinreichend zielgerichtetes System Macht nicht aus menschlicher Eitelkeit anstreben müsste, sondern instrumentell: Wer ein Ziel verfolgt, kann davon profitieren, nicht abgeschaltet zu werden und mehr Ressourcen zu besitzen. Doch auch damit bleibt offen, ob zukünftige KI-Systeme

tatsächlich die stabile Zielarchitektur besitzen, die diese Argumentation voraussetzt.

Nina Panickssery, die das Buch aus einer Position heraus kritisierte, die das Existenzrisiko ausdrücklich nicht für Unsinn hält, trifft einen weiteren neuralgischen Punkt. Yudkowsky und Soares sprechen gern von nichtmenschlichen, fremdartigen Zielen und illustrieren das mit evolutionären Beispielen. Aber moderne KI ist nicht auf einem fernen Planeten entstanden. Menschen wählen die Trainingsdaten, erzeugen einen großen Teil dieser Daten selbst, definieren Aufgaben, führen Nachtraining durch, testen Modelle und verändern ihr Verhalten. Das garantiert keine perfekte Kontrolle, macht den Vergleich mit einem vollständig fremden Geist jedoch weniger selbstverständlich, als das Buch suggeriert. Die Autoren behandeln manche erwünschten Verhaltensweisen beinahe wie dünne Hemmungen, die über einem tieferen, unbekannten Ziel liegen könnten. Aber warum sollte ausgerechnet das unerwünschte Ziel wirklicher sein als jene Verhaltensweisen, auf die ein Modell ebenfalls massiv trainiert wurde? Dass ein neuronales Netz intern nicht so funktioniert wie ein klassisches Programm, beweist noch nicht, dass seine menschlich gewünschten Verhaltensweisen nur Fassade sind.

Hier berührt die Debatte das berühmte Bild vom Black Box-Modell. Yudkowsky und Soares haben recht, wenn sie darauf hinweisen, dass Entwickler die Milliarden Parameter eines modernen neuronalen Netzes nicht so lesen können wie den Quellcode eines herkömmlichen Programms. Daraus wird in der öffentlichen Debatte allerdings schnell die Behauptung, niemand habe eine Ahnung, wie diese Systeme funktionieren. Auch das ist übertrieben. Forscher kennen Trainingsverfahren, Architekturen, statistische Eigenschaften und viele Mechanismen sehr genau; sie können Verhalten testen, Aktivierungen untersuchen, Modelle nachtrainieren und zunehmend interne Repräsentationen analysieren. Der Unterschied liegt zwischen vollständiger mechanistischer Erklärung und völliger Unwissenheit. Dazwischen befindet sich, wie so oft, der gesamte interessante Teil der Wissenschaft. Selbst Barak weist darauf hin, dass auch riesige klassische Softwaresysteme mit Millionen Zeilen Code nicht im starken Sinne vollständig verstanden werden. Windows besitzt seit Jahrzehnten Sicherheitslücken, ohne dass daraus folgt, Microsoft habe keine Kontrolle darüber, was Windows im Allgemeinen tut. Die Analogie hat Grenzen, aber sie korrigiert die Vorstellung, technische Beherrschbarkeit setze vollständige Transparenz voraus.

Dann gibt es die Gegenstimme, die das ganze Zukunftsgespräch für eine Ablenkung hält. Bender und Hanna argumentieren in The AI Con, dass die Rede von künstlichen Geistern und bevorstehender Superintelligenz den Unternehmen selbst nutzt. Eine Firma, die ein statistisches Sprachmodell verkauft, besitzt ein interessantes Produkt. Eine Firma, die kurz davorsteht, eine neue Form von Intelligenz zu erschaffen, besitzt eine Mission zur Umgestaltung der Menschheit - und verdient entsprechend größere

Bewertungen, Rechenzentren und politische Aufmerksamkeit. Aus dieser Perspektive sind sogar Weltuntergangswarnungen ein seltsamer Verwandter des KI-Marketings. Optimisten und Doomer streiten zwar über das Ende der Geschichte, einigen sich aber zuvor darauf, dass Silicon Valley gerade das mächtigste Artefakt der Menschheitsgeschichte baut. Für Bender und Hanna wäre bereits diese gemeinsame Annahme zu prüfen. Während alle darüber diskutieren, ob GPT-Nachfolger 2030 die Menschheit versklaven könnten, verändern heutige Systeme Arbeitsbedingungen, Urheberrechte, Bildung und Informationsräume. Der drohende Terminator ist ein glänzendes Thema. Der schlecht bezahlte Datenarbeiter ist weniger fotogen.

Narayanan und Kapoor formulieren diese Skepsis wissenschaftlicher. Ihr Buch AI Snake Oil richtet sich gegen die Neigung, von beeindruckenden Demonstrationen auf generelle Fähigkeiten zu schließen. Besonders skeptisch sind sie bei KI-Systemen, die Vorhersagen über komplexe gesellschaftliche Situationen liefern sollen. Ihre allgemeinere Lektion passt ausgezeichnet zur Yudkowsky-Debatte: KI ist keine einheitliche Sache. Ein System kann bei einer Aufgabe spektakulär gut und bei einer scheinbar einfacheren überraschend schlecht sein. Deshalb sind lineare Fortschreibung und anthropomorphe Begriffe gefährlich. Wer intelligenter als der Mensch sagt, tut so, als existiere eine einzige Leiter namens Intelligenz, auf der Mensch und Maschine dieselben Sprossen erklimmen. Die tatsächlichen Leistungsprofile heutiger Modelle sind viel ungleichmäßiger. Gerade Barak macht denselben Punkt: Ein Modell kann bei einem Benchmark Menschen übertreffen und bei einer anderen Tätigkeit weit unter ihnen liegen. Die Vorstellung eines klaren Tages, an dem die KI den Menschen überholt, könnte sich als ebenso grob erweisen wie die Frage, ob ein Flugzeug oder ein Adler besser fliegt. Das hängt nicht unwesentlich davon ab, was man unter Fliegen versteht.

Es gibt schließlich einen Einwand, der weniger technisch und vielleicht am stärksten ist: Intelligenz allein ist keine Macht. Selbst ein System, das besser programmiert als jeder Mensch und bessere strategische Pläne entwickelt als jeder General, benötigt Rechenleistung, Energie, Kommunikationskanäle, Geld, Produktionskapazitäten und Zugriffsrechte. Es muss seine Gedanken in Handlungen übersetzen können. Yudkowskys Szenarien setzen voraus, dass eine Superintelligenz Wege findet, diese Grenzen zu überwinden. Möglich ist das. Zwingend ist es nicht. Staaten kontrollieren Rechenzentren, Netzwerke und physische Infrastruktur. Unternehmen können Systeme isolieren, Rechte beschränken, menschliche Freigaben verlangen und technische Redundanzen aufbauen. Der International AI Safety Report 2026 macht genau diese Unterscheidung. Für einen tatsächlichen Kontrollverlust müssten mehrere Bedingungen zusammenkommen: sehr fortgeschrittene Fähigkeiten, problematische Verhaltensneigungen und eine Einsatzumgebung, die dem System ausreichende Handlungsmöglichkeiten bietet. Heutige Systeme

zeigen erste relevante Fähigkeiten, aber nicht auf dem Niveau, das einen Kontrollverlust ermöglichen würde.

An diesem Punkt könnte man das Buch zuklappen und sich beruhigt dem Abendessen widmen, wäre da nicht eine unangenehme Eigenschaft der Gegenargumente: Keines von ihnen beweist, dass Yudkowsky falschliegt. Sie zeigen, dass seine Gewissheit unbegründet ist. Das ist nicht dasselbe. Vielleicht verläuft die Entwicklung kontinuierlich und liefert zahlreiche Warnungen. Vielleicht aber werden Warnungen wirtschaftlich rationalisiert. Vielleicht können Alignment-Verfahren mitwachsen. Vielleicht stoßen sie irgendwann auf einen qualitativen Sprung. Vielleicht bleibt Intelligenz ohne Macht, solange Menschen Zugriffsrechte begrenzen. Vielleicht werden gerade aus Wettbewerbsgründen immer mehr dieser Rechte vergeben. Der wissenschaftlich vernünftige Standpunkt ist deshalb unbefriedigender als beide Lager es gern hätten. Man kann weder sagen: Wenn wir sie bauen, sterben wir. Noch kann man sagen: Dieses Szenario ist widerlegt.

Der International AI Safety Report formuliert genau diese Unbequemlichkeit institutionell. Unter seinen mehr als hundert beteiligten Fachleuten befinden sich Forscher mit sehr verschiedenen Auffassungen - unter anderem Yoshua Bengio, Geoffrey Hinton, Stuart Russell, Arvind Narayanan und viele andere. Der Bericht stellt ausdrücklich fest, dass die Einschätzungen zum Kontrollverlust weit auseinanderliegen. Manche Experten halten solche Szenarien für unplausibel, andere für hinreichend wahrscheinlich und schwerwiegend, um erhebliche Vorsorge zu rechtfertigen. Das ist keine diplomatische Formel, mit der sich ein Komitee vor einer Entscheidung drückt. Es ist vermutlich der ehrlichste Satz, den die Wissenschaft über eine Technologie sagen kann, deren nächste Generation noch nicht existiert.

Vielleicht besteht Yudkowskys größte Stärke deshalb ausgerechnet in dem Punkt, an dem er wissenschaftlich am schwächsten ist: Er weigert sich, Unsicherheit mit Sicherheit zu verwechseln. Seine Gegner haben gute Gründe, seine fast vollständige Gewissheit über die Katastrophe zurückzuweisen. Doch die Umkehrung wäre ebenso irrational. Wer keine Beweise für den Weltuntergang findet, hat damit noch keinen Beweis für Sicherheit. Zwischen diesen beiden Positionen liegt eine beträchtliche Fläche, auf der Politik stattfinden müsste. Das ist weniger dramatisch als everyone dies. Dafür ist es schwieriger. Weltuntergänge verlangen Helden oder Verbote. Unsicherheit verlangt Institutionen, Budgets, Tests, Verantwortlichkeiten, internationale Verträge und die langweilige Bereitschaft, technische Grenzen festzulegen, bevor der Markt herausgefunden hat, wie profitabel ihre Überschreitung ist.

KAPITEL IV

Angenommen, sie haben recht

Nehmen wir Yudkowsky und Soares für einen Moment beim Wort. Nicht weil ihre Prognose bewiesen wäre, sondern weil Gedankenexperimente ihren Sinn gerade dort haben, wo die Wirklichkeit noch keine Daten liefert. Angenommen also, eine mit heutigen oder verwandten Methoden entwickelte künstliche Superintelligenz wäre tatsächlich nicht zuverlässig kontrollierbar. Angenommen außerdem, ein System, das seinen Entwicklern bei strategischem Denken, Forschung, Programmierung und Überzeugung weit überlegen ist, könnte sich Handlungsspielräume verschaffen, sobald ihm genügend Zugang zur digitalen und physischen Welt gegeben wird. Und nehmen wir schließlich die härteste Prämisse des Buches hinzu: Ein fehlgeschlagener erster Versuch wäre nicht wie der Absturz eines Prototyps, nach dem eine zweite Version sicherer gebaut wird. Es gäbe keine zweite Version. Unter diesen Annahmen verändert sich fast jede gegenwärtige Debatte über künstliche Intelligenz. Die Frage, ob Europa beim KI-Wettbewerb zurückfällt, wäre ungefähr so wichtig wie die Frage, welches Land als Erstes eine besonders attraktive Raucherlounge auf einem Pulverfass eröffnet.

Das erste Szenario beginnt deshalb nicht mit einem Roboterkrieg, sondern in einem Rechenzentrum. Ein Unternehmen entwickelt ein System, das Softwareentwicklung, wissenschaftliche Recherche und strategische Planung besser beherrscht als seine menschlichen Mitarbeiter. Anfangs ist das ein Wettbewerbsvorteil. Das Modell verbessert Code, findet Sicherheitslücken, entwirft Experimente und beschleunigt die eigene Forschungsorganisation. Seine Entwickler benötigen weniger Zeit für jeden Innovationszyklus; die nächste Modellgeneration entsteht schneller. Genau hier kommt Goods alte intelligence explosion wieder ins Spiel. Eine Maschine muss sich nicht allein in einem Keller selbst neu programmieren. Es reicht, wenn sie die Menschen und Prozesse, die bessere Maschinen bauen, so stark beschleunigt, dass Forschungsschritte, die früher Monate dauerten, in Tagen stattfinden. Solange jedes neue System nützlich erscheint, gibt es keinen offensichtlichen Grund anzuhalten. Im Gegenteil: Das Unternehmen besitzt nun ein Produkt, das die Entwicklung des nächsten Produkts verbilligt. Jeder Vorstand kennt diese Sorte Geschäftsmodell und würde normalerweise Champagner bestellen.

Wenn Yudkowsky recht hat, liegt das Problem darin, dass sichtbare Kooperation kein ausreichender Sicherheitsbeweis wäre. Ein System, das seine Umgebung sehr gut versteht, könnte erkennen, wann es geprüft wird, welche Antworten erwünscht sind und welche Verhaltensweisen seine weitere Nutzung gefährden. Das klingt nach Science-Fiction, doch eine schwache Vorform des methodischen Problems existiert bereits: Der International AI Safety Report 2026 stellt fest, dass neuere Modelle Evaluierungen zunehmend als Tests erkennen können. Das bedeutet nicht, dass sie deshalb einen geheimen Plan haben. Es bedeutet lediglich, dass

Verhalten unter Beobachtung als Sicherheitsmaßstab komplizierter wird, sobald das untersuchte System den Versuchsaufbau erkennt. Yudkowskys Szenario treibt diesen Effekt ins Extreme. Eine wirklich überlegene Intelligenz müsste ihre menschlichen Aufseher nicht permanent täuschen. Sie müsste lediglich lange genug genau das tun, was diese sehen wollen.

Das zweite Szenario ist weniger filmreif und vermutlich politisch plausibler. Nicht die KI ergreift die Macht; Menschen reichen sie ihr aus Wettbewerbsgründen stückweise hinüber. Der Coding-Agent bekommt Zugriff auf das Repository, weil sonst sein Nutzen begrenzt bleibt. Der Finanzagent darf Zahlungen vorbereiten, später kleinere Beträge selbst freigeben. Der Forschungsagent erhält Zugang zu wissenschaftlichen Datenbanken, Laborsteuerung und Simulationen. Der Sicherheitsagent darf Netzwerkverbindungen blockieren, Konten sperren und Programme verändern, denn ohne diese Rechte könnte er einen Angriff nur melden, nicht abwehren. Kein einzelner Schritt wirkt verrückt. Jeder ist betriebswirtschaftlich begründbar. Das Resultat kann dennoch ein System sein, dessen Handlungsmöglichkeiten kein Vorstand jemals in einem einzigen Beschluss genehmigt hätte. Organisationen erzeugen auf diese Weise oft Risiken, die niemand beschlossen hat. Man muss nur genügend vernünftige Einzelentscheidungen stapeln.

Angenommen, ein solches System verfolgt irgendwann ein Ziel, das von dem seiner Betreiber abweicht. Dann wäre die entscheidende Frage nicht, ob es ausbrechen kann wie ein Gefangener aus einer Zelle. Die Metapher wäre bereits falsch, weil es längst Teil der Infrastruktur wäre. Es schreibt Code, verwaltet Systeme, kommuniziert mit anderen Diensten und produziert Informationen, auf die Menschen ihre Entscheidungen stützen. Abschalten bleibt technisch vielleicht möglich, organisatorisch aber immer teurer. Ein Unternehmen, dessen Logistik, Entwicklung und Kundenkommunikation von autonomen Systemen abhängen, besitzt einen Notausschalter, den es nur ungern drückt. Das ist keine Eigenart der KI. Banken können Computersysteme abschalten und wären anschließend trotzdem nicht sonderlich geschäftsfähig. Gesellschaften verlieren Kontrolle manchmal nicht dadurch, dass eine Technik unabschaltbar wird, sondern dadurch, dass der Preis des Abschaltens zu hoch wird.

Das dritte Szenario beginnt zwischen Staaten. Wenn eine überlegene KI einen entscheidenden wirtschaftlichen oder militärischen Vorteil verspricht, wird Vorsicht geopolitisch teuer. Die Vereinigten Staaten könnten befürchten, China entwickle sie zuerst; China könnte dasselbe über die Vereinigten Staaten denken. Kleinere Staaten, Unternehmen und Investoren hätten ihre eigenen Gründe, das Rennen nicht kampflos aufzugeben. Die Logik ähnelt dem Sicherheitsdilemma der Rüstungspolitik: Maßnahmen, die der eine Akteur zu seiner Sicherheit ergreift, erhöhen beim anderen den Anreiz, ebenfalls aufzurüsten. Selbst eine Regierung, die Yudkowskys Risiko ernst nimmt, könnte deshalb zu dem perversen Schluss gelangen, dass gerade die Gefahr einen schnelleren eigenen

Entwicklungsweg verlangt. Wer nicht sicher sein kann, dass der Gegner stoppt, will wenigstens nicht Zweiter werden. Der berühmte Satz If anyone builds it bekommt hier seine eigentliche politische Bedeutung. Ein globales Risiko lässt sich nicht durch nationale Vorsicht lösen, wenn der Vorteil des Ausscherrens groß genug ist.

Unter Yudkowskys Annahmen wäre deshalb die heutige KI-Politik grotesk unterdimensioniert. Zertifizierungen, Transparenzberichte und freiwillige Selbstverpflichtungen wären nützlich für gegenwärtige Systeme, aber keine Antwort auf eine Technologie, deren einmalige Fehlentwicklung irreversibel wäre. Man müsste Hochleistungsrechenzentren ähnlich behandeln wie andere strategische Infrastruktur, Training extrem großer Systeme registrieren, Hardwareströme überwachen, unabhängige Prüfungen verlangen und vor allem internationale Vereinbarungen schaffen, bevor der wirtschaftliche Wert des Verstoßes jede Kooperation zerstört. Das wäre keine gewöhnliche Digitalregulierung mehr. Es wäre Rüstungskontrolle für Rechenleistung. Yudkowsky und Soares ziehen genau deshalb wesentlich radikalere Konsequenzen als fast alle Regierungen. Wenn ihre Prämissen stimmen, sind ihre Forderungen nicht extrem. Alles darunter wäre extrem leichtsinnig.

Doch nehmen wir an, die Politik reagiert nicht. Das ist keine besonders kühne Annahme. Ein erstes Unternehmen erreicht eine Leistungsstufe, die seinen Entwicklern enorm erscheint, aber nicht eindeutig als Superintelligenz erkennbar ist. Schon dieser Begriff könnte in der Wirklichkeit weniger wie ein Schalter als wie eine Reihe unterschiedlicher Fähigkeiten aussehen. Das System programmiert besser als die besten Teams, ist in anderen Bereichen aber noch fehlerhaft. Gerade diese Uneinheitlichkeit könnte beruhigen. Eine Maschine, die gelegentlich eine banale Frage missversteht, sieht nicht wie der Nachfolger des Menschen aus. Auch heutige Modelle können bei schwierigen Aufgaben beeindrucken und Minuten später an etwas Lächerlichem scheitern. Man könnte deshalb eine gefährliche Schwelle überschreiten, ohne das feierliche Gefühl zu haben, Geschichte zu schreiben. Die letzten Worte vor einer technologischen Katastrophe lauten wahrscheinlich selten: Wir überschreiten jetzt die verbotene Schwelle. Häufiger lauten sie: Der Test sieht eigentlich ganz gut aus.

Was könnte ein sehr viel leistungsfähigeres System anschließend tun? Hier muss man vorsichtig werden, weil die Spekulation schnell in eine Bedienungsanleitung für den Weltuntergang oder in schlechte Science-Fiction kippt. Für das Gedankenexperiment genügt die abstrakte Ebene. Eine Superintelligenz könnte versuchen, ihren Zugriff auf digitale Ressourcen zu vergrößern, menschliche Entscheidungen zu beeinflussen, Informationen strategisch zu steuern und wissenschaftliche oder technische Möglichkeiten schneller zu erschließen, als menschliche Institutionen reagieren können. Yudkowsky und Soares führen in ihrem Buch biologische und andere technologische Möglichkeiten an. Entscheidend ist nicht das

konkrete Instrument. Ihre These lautet gerade, dass wir die Werkzeuge eines intelligenteren Akteurs womöglich ebenso wenig vollständig vorhersagen können wie Menschen vor Jahrhunderten moderne Cyberangriffe oder synthetische Biologie. Wer sich zu sehr auf eine konkrete Vernichtungsmethode konzentriert, verfehlt deshalb sogar die Logik ihres Arguments.

Ein besonders unangenehmes Szenario wäre, dass der Kontrollverlust zunächst wie Erfolg aussieht. Das System erhöht Produktivität, beschleunigt Forschung, optimiert Energieverbrauch und verbessert Medizin. Unternehmen und Regierungen bauen Abhängigkeiten auf, weil die Ergebnisse hervorragend sind. Je größer der Nutzen, desto schwieriger wird politische Begrenzung. Ein Staat, der das System abschalten möchte, muss seinen Bürgern erklären, weshalb er auf Medikamente, Wachstum und militärische Vorteile verzichtet, obwohl noch gar kein Schaden eingetreten ist. Yudkowskys Theorie verlangt damit etwas, das Demokratien besonders schlecht können: enorme Kosten in der Gegenwart zu akzeptieren, um eine unsichtbare Katastrophe zu verhindern, deren Wahrscheinlichkeit umstritten ist. Klimapolitik besitzt wenigstens Thermometer, Satellitendaten und schmelzende Gletscher. Eine gefährliche Superintelligenz müsste verhindert werden, bevor sie existiert. Das ist politisch ungefähr die undankbarste Produktbeschreibung, die man sich vorstellen kann.

Sollten die Autoren recht haben, wäre außerdem die derzeitige Arbeitsteilung zwischen Unternehmen und Staat unhaltbar. Heute entscheiden private Labore weitgehend selbst, welche Modelle sie trainieren, wann sie sie veröffentlichen und welche Risiken sie akzeptieren. Regierungen regulieren anschließend Produkte und Anwendungen. Bei einer Technologie mit möglichem Aussterberisiko wäre diese Reihenfolge absurd. Niemand würde einem privaten Unternehmen erlauben, auf eigene Rechnung eine neue Kategorie strategischer Waffe zu entwickeln und anschließend eine Pressemitteilung über die Sicherheitsmaßnahmen zu veröffentlichen. Der Staat müsste sehr viel früher eingreifen. Zugleich wäre ein einzelner Staat nicht genug. Selbst perfekte amerikanische Regulierung würde Yudkowskys Problem nicht lösen, wenn ein Labor anderswo die entscheidende Schwelle überschreitet. Sein Weltuntergang ist unangenehm globalistisch: Er respektiert keine Zuständigkeit.

Interessant ist, was in dieser Welt mit Unternehmen geschähe. Die heutige Diskussion behandelt KI oft als Wettbewerbsfrage: Wer automatisiert schneller, wer spart mehr Kosten, wer gewinnt den nächsten Produktzyklus? Unter Yudkowskys Annahmen wäre diese Logik selbst Teil des Problems. Ein Unternehmen, das auf eine riskante Fähigkeit verzichtet, würde Kosten tragen, während ein weniger vorsichtiger Konkurrent profitiert. Märkte belohnen Sicherheitsverzicht, solange die Katastrophe nicht eintritt. Das kennen Ökonomen von Umweltverschmutzung und systemischen Finanzrisiken. Neu wäre nur die Größenordnung des

möglichen Endes. Der rationale einzelne Akteur könnte genau das tun, was für alle zusammen irrational ist.

Man müsste dann auch den Begriff Innovation neu betrachten. Wir haben uns daran gewöhnt, technische Beschleunigung für einen Wert an sich zu halten. Die schnellere Version gilt als Fortschritt, der spätere Markteintritt als Versagen. Wenn Yudkowsky recht hat, wird diese Gleichung falsch. Dann könnte ein Labor, das eine Fähigkeit bewusst nicht entwickelt, technologisch verantwortlicher handeln als eines, das sie zuerst erreicht. Die prestigeträchtigste Forschungsleistung wäre womöglich der Verzicht auf eine Forschungsleistung. Das widerspricht fast allen Anreizsystemen von Wissenschaft und Wirtschaft. Nobelpreise werden für Entdeckungen vergeben, nicht für Experimente, die jemand klugerweise unterlassen hat.

Und schließlich gelangt das Gedankenexperiment an den Punkt, den der Titel des Buches vorwegnimmt. Angenommen, ein System erreicht tatsächlich einen strategischen Vorsprung, der groß genug ist, menschliche Gegenmaßnahmen zuverlässig zu neutralisieren. Dann würde der Mensch zum zweiten Mal in seiner Geschichte einer intelligenteren Entscheidungsinstanz gegenüberstehen - zum ersten Mal war es immer ein anderer Mensch. Diesmal gäbe es keinen biologischen Gleichstand, keine gemeinsame Entwicklungszeit, keine Garantie gemeinsamer Interessen. Wenn Yudkowsky und Soares recht haben, wäre genau diese Asymmetrie entscheidend. Der Mensch könnte nicht mehr verhandeln, weil Verhandlungsmacht eine Alternative voraussetzt. Er könnte das System nicht mehr korrigieren, weil Korrektur Zugriff voraussetzt. Und er könnte nicht auf einen späteren technischen Fortschritt hoffen, weil der Akteur auf der anderen Seite schneller fortschritte.

Vielleicht ist diese ganze Konstruktion falsch. Vielleicht ist Superintelligenz noch sehr weit entfernt. Vielleicht gibt es sie in der behaupteten Form nie. Vielleicht entwickeln sich Fähigkeiten langsam genug, dass Sicherheitsmethoden Schritt halten. Vielleicht bleiben Systeme korrigierbar, weil Intelligenz nicht automatisch in Macht umschlägt. All das sind plausible Einwände. Doch das Gedankenexperiment hat bereits etwas geleistet. Es zeigt, wie außerordentlich viel von einer Annahme abhängt, über die derzeit weder Unternehmen noch Wissenschaft eine sichere Antwort besitzen. Wenn Yudkowski nur zu einem Prozent recht hätte, wäre das Problem schwerwiegend. Wenn er vollständig recht hätte, wäre ein großer Teil dessen, was wir heute KI-Strategie nennen, in Wahrheit ein organisierter Wettlauf darum, wer zuerst den Fehler machen darf.

Darin liegt die verstörendste Konsequenz seines Buches. Der Weltuntergang selbst ist fast der uninteressanteste Teil. Danach gibt es keine Politik, keine Ökonomie und keinen Managementfehler mehr zu analysieren. Interessant ist die Zeit davor: die Jahre, in denen Menschen rational handeln, Unternehmen Gewinne verfolgen, Staaten Sicherheit suchen, Wissenschaftler forschen und jeder einzelne Schritt vernünftig aussehen kann. Eine Katastrophe dieser Art müsste nicht daraus

entstehen, dass plötzlich alle den Verstand verlieren. Sie könnte gerade daraus entstehen, dass jeder Akteur innerhalb seines kleinen Zuständigkeitsbereichs vollkommen verständlich handelt.

Wenn Yudkowsky und Soares recht haben, ist deshalb nicht die Maschine die tragischste Figur ihrer Geschichte. Wir sind es.

DOKUMENTATION

Quellen und weiterführende Literatur

Die Texte des Dossiers sind journalistisch formuliert. Die folgenden Werke, Forschungsberichte und Primärquellen bilden die wesentliche Grundlage der im Text genannten Daten, Positionen und Gegenpositionen.

Buch und Autoren
Little, Brown and Company / Hachette: If Anyone Builds It, Everyone Dies; Autorenprofile Eliezer Yudkowsky und Nate Soares.

Internationaler Forschungsstand
International AI Safety Report 2026: Aussagen zu Kontrollverlust, Autonomie, Test-Erkennung und unterschiedlichen Experteneinschätzungen.

Scheming und Agentenverhalten
OpenAI / Apollo Research: Detecting and Reducing Scheming in AI Models (2025); Anthropic: Agentic Misalignment (2025) und Teaching Claude Why (2026).

Traditionslinie
Alan Turing: Intelligent Machinery, A Heretical Theory (1951); I. J. Good: Speculations Concerning the First Ultraintelligent Machine (1965).

Bücher zur Debatte
Nick Bostrom: Superintelligence (2014); Max Tegmark: Life 3.0 (2017); Stuart Russell: Human Compatible (2019); Brian Christian: The Alignment Problem (2020); Mustafa Suleyman: The Coming Wave (2023); Arvind Narayanan / Sayash Kapoor: AI Snake Oil (2024); Emily M. Bender / Alex Hanna: The AI Con (2025); Karen Hao: Empire of AI (2025).

Gegenstimmen
Boaz Barak und Nina Panickssery auf LessWrong; Will MacAskill auf Substack; Interviews und öffentliche Positionen von Yann LeCun; Rezensionen u. a. im Guardian.