

Die KI wird immer schlauer

Vorbemerkung

Liebe Leserin, lieber Leser, Sie haben mein Versprechen aus dem letzten Absatz dieses Buches ernst genommen und sind in den Update-Bereich gefolgt. Vielen Dank dafür. Sie haben gerade ein Buch gelesen, das mit dem Anspruch antrat, Ihnen eine Grundlage für eine eigene Meinung zur Künstlichen Intelligenz zu geben. Was ich Ihnen versprochen habe – dass dieses Buch im Augenblick des Erscheinens schon ein Stück weit veraltet ist – erfüllt sich schneller, als mir lieb ist. Deshalb gibt es dieses Update. Und deshalb steht es nicht nur auf der begleitenden Webseite, sondern bei späteren Auflagen direkt im Buch.

Es geht mir hier nicht um einzelne Korrekturen oder Kleinigkeiten. Es geht um drei Entwicklungen, die während der Drucklegung sichtbar wurden und die nicht weniger verändern als den Zeithorizont der gesamten Argumentation dieses Buches. Wenn ich in Kapitel 8 das „Schreckgespenst ASI“ beschrieben habe, klang das für manche Leser vielleicht noch wie ein theoretischer Endpunkt. Nach diesem Kapitel werden Sie ahnen,

warum ich glaube, dass dieser Endpunkt näher gerückt ist.

Wir gehen in drei Schritten vor. Zuerst stelle ich Ihnen zwei konkrete Werkzeuge vor, die im Mittelpunkt der jüngsten Entwicklung stehen: Claude Mythos von Anthropic und ChatGPT Codex von OpenAI. Anschließend zeige ich, was an ihnen wirklich neu ist – und warum die Bezeichnung „Sprachmodell“ für beide schon zu kurz greift. Im dritten Schritt ordne ich die Entwicklungen den Kapiteln dieses Buches zu und sage Ihnen ehrlich, welche Argumente gestärkt, welche relativiert und welche verschoben werden.

Akteur 1: Claude Mythos – die KI als Sicherheitsexperte und potenzielle Waffe

Anfang 2026 hat Anthropic – das Unternehmen von Dario Amodei, das wir in Kapitel 3 ausführlich kennengelernt haben – ein Modell ausgerollt, das man nicht in der App ausprobieren kann und das auch in keinem Preis-Leistungs-Vergleich auftaucht: Claude Mythos.¹

Was kann dieses Modell? Es findet Sicherheitslücken in Software und nutzt sie autonom aus. Und zwar auf einem

¹Anthropic: „Introducing Claude Mythos Preview“, Releasebot/Anthropic, Mai 2026. Verfügbar unter anthropic.com.

Niveau, das Anthropic mit dem englischen Ausdruck „elite human experts“ umschreibt – also auf dem Niveau jener Spezialisten, die in den großen Cybersecurity-Firmen die Top-Aufgaben übernehmen. Genauer: in vielen Tests übertrifft Mythos dieses Niveau.

Lassen Sie mich kurz erklären, was eine Sicherheitslücke ist, denn nicht jeder Leser arbeitet beruflich damit. Software besteht aus Millionen Zeilen Code. Manche dieser Zeilen sind fehlerhaft – nicht so, dass die Software gleich abstürzt, sondern so, dass ein findiger Mensch dort eindringen kann. Stellen Sie sich einen versteckten Schlüssel unter der Fußmatte vor. Wer ihn findet, kommt ins Haus. Wer ihn vorher findet und beseitigt, schützt das Haus. Genau diese beiden Rollen – Eindringen und Schützen – kann Mythos übernehmen. Und das verändert die Spielregeln in einer Disziplin, in der bisher menschliche Spezialisten der entscheidende Faktor waren.

Anthropic verteilt Mythos nicht über den freien Markt. Stattdessen gibt es ein Programm namens „Project Glasswing“, über das ausgewählte Partner Zugang erhalten.² Die Logik dahinter ist nachvollziehbar: Bevor Angreifer solche Fähigkeiten haben, sollen Verteidiger

²Anthropic: „Claude Security/Project Glasswing“, Pressemitteilung Mai 2026.

sie zuerst haben. Es ist die klassische Asymmetrie aller Sicherheitstechnologien, vom Tresor über das Antibiotikum bis zur Verschlüsselung.

Das Dual-Use-Problem liegt offen auf der Hand. Dasselbe Werkzeug, das eine Bank vor Einbruch schützt, könnte sie auch ausräumen. Dasselbe Modell, das ein Krankenhaus sicherer macht, könnte ein anderes lahmlegen. Hier zeigt sich erstmals greifbar, was bisher abstrakt blieb: KI ist nicht nur Sprachassistent, sondern Werkzeug mit kritischer Reichweite. Geoffrey Hinton, dem wir in Kapitel 2 begegnet sind, hat genau das vor Jahren als eines der eigentlichen Risiken benannt – nicht den selbstmordlüsternen Terminator, sondern die Verfügbarkeit gefährlicher Fähigkeiten für jeden, der sie benötigt.

Bis hierher die offizielle Lesart, die Anthropic selbst formuliert. Aber das eigentlich Brisante an Glasswing ist eine zweite Schicht, die in den Pressemitteilungen weniger Raum bekommt. Anthropic entscheidet als privates Unternehmen, wer Mythos bekommt und wer nicht. Keine staatliche Zulassung. Keine internationale Aufsicht. Keine Versteigerung. Keine Versorgungspflicht. Es ist eine unternehmerische Auswahlentscheidung, getroffen am Sitz in San Francisco, nach Kriterien, die niemand außerhalb von Anthropic vollständig kennt.

Was heißt das konkret? Wer als verlässlicher Partner von Anthropic gilt, bekommt einen entscheidenden Vorsprung bei der Verteidigung seiner Systeme. Wer es nicht ist – etwa weil er in einem unbequemen Markt operiert, ein direkter Konkurrent ist oder politisch in Ungnade fällt – bekommt diesen Vorsprung nicht. Es entsteht eine Zwei-Klassen-Sicherheit, und die Zugehörigkeit zur ersten Klasse hängt an Entscheidungen, die nicht demokratisch legitimiert sind.

Man kann mir entgegenhalten, dass Anthropic auch nicht jeden Algorithmus auf einer Pressekonferenz präsentieren könnte, ohne die Welt unsicherer zu machen. Stimmt. Aber das verschiebt die Frage nur, statt sie zu beantworten. Aus „Sollen wir das veröffentlichen?“ wird „Wer entscheidet eigentlich, an wen wir es geben?“. Und die Antwort lautet derzeit: ein einzelnes Unternehmen, ohne demokratische Rückbindung, mit eigenen wirtschaftlichen Interessen, in einem Land, das nicht das Ihre ist.

Eine Konsequenz, die selten ausgesprochen wird: Wer von einem solchen Programm abhängig ist, überlegt sich gut, ob er den Anbieter öffentlich kritisiert. Das ist keine offene Erpressung, sondern eine stille Disziplinierungswirkung. Wer von Anthropic mit Mythos versorgt wird, ist tendenziell ein leiserer

Kritiker von Anthropic. Wer Sicherheit als Privileg verteilt, baut sich beiläufig eine politische Maschinerie.

Ein historischer Vergleich drängt sich auf, hinkt aber. Die Atomtechnik wurde nach dem Zweiten Weltkrieg ebenfalls zu einer Dual-Use-Technologie. Wir haben darauf mit dem Atomwaffensperrvertrag reagiert, mit der Internationalen Atomenergie-Behörde, mit Inspektionen, Sanktionen, Verhandlungen. Bei Mythos haben wir nichts dergleichen. Schlimmer noch: Während man Plutonium nur in aufwendigen Anlagen produzieren kann, ist Software kopierbar. Eine durchgesickerte Kopie eines Mythos-ähnlichen Modells wäre ein Sicherheitsproblem ganz anderer Größenordnung als ein durchgesickertes Reaktordesign.

Erste Konsequenzen sehen wir bereits. Versicherer beginnen, KI-spezifische Cyber-Policen zu konzipieren, mit Prämien je nach Anbieterzugang. Regierungen überprüfen, welche KI-Anbieter sie noch zulassen. Das US-Verteidigungsministerium hat Anthropic in einem viel beachteten Schritt als „supply chain risk“ eingestuft, nachdem das Unternehmen vertragliche Klauseln zur Massenüberwachung und zu vollautonomen Waffen nicht streichen wollte.³ Das ist die andere Seite der

³Wikipedia: „Claude (language model)“, Stand Mai 2026. Vermerk zur US-Bundesregierung und zur Einstufung als „supply chain risk“ durch das US-Verteidigungsministerium.

Medaille: Wer privat über Sicherheit entscheidet, kann auch privat sagen, wem er nicht zur Verfügung steht. Für Anthropic ist das eine Frage des Gewissens, für die US-Regierung eine Frage der Souveränität.

Was wir nicht wissen – und das gehört zur ehrlichen Beschreibung: Wie viele vergleichbare interne Modelle existieren bereits bei Google, Meta, OpenAI, ohne dass wir Namen kennen? Mythos ist der erste, dessen Existenz öffentlich gemacht wurde. Es wäre sehr überraschend, wenn er der einzige bliebe.

Akteur 2: ChatGPT Codex – die KI als autonomer Entwickler

Der zweite Akteur kommt aus dem Haus von Sam Altman, den wir in Kapitel 3 ebenfalls porträtiert haben. OpenAI bietet seit Mai 2025 ein Werkzeug namens „Codex“ an, das im Lauf des Jahres 2026 zu einem vollwertigen autonomen Programmier-Agenten ausgebaut wurde.⁴

Erklärung in Alltagssprache, denn das Wort „Agent“ ist gerade dabei, seine Bedeutung zu verändern. Bisher war eine KI eher wie ein sehr schneller Kollege, der Vorschläge macht: Sie schreiben eine Zeile Code, die KI

⁴OpenAI: „Introducing Codex“, openai.com, Mai 2025; aktualisierter Stand 2026.

ergänzt drei. Mit Codex kehrt sich das Verhältnis um. Sie beschreiben in zwei Sätzen, was Sie haben möchten – „Fixe den Fehler im Bestellsystem, der gestern auftrat“ oder „Baue eine neue Funktion, mit der Kunden ihre Adressen selbst aktualisieren können“ – und Codex übernimmt den Rest.

Wie läuft so ein Auftrag konkret ab? Codex öffnet sich eine abgeschottete Arbeitsumgebung in der Cloud. Es holt sich den existierenden Programmcode der Firma. Es liest sich ein. Es schreibt Änderungen über mehrere Dateien hinweg. Es testet die Änderungen selbst. Wenn die Tests fehlschlagen, repariert es seine eigenen Fehler. Es dokumentiert die Arbeit in einem Bericht. Und dann legt es das Ergebnis einem menschlichen Entwickler zur Freigabe vor. Aus „programmieren“ ist „freigeben“ geworden.

Die Skala der Verbreitung ist bemerkenswert. Codex wird Anfang 2026 nach Angaben von OpenAI wöchentlich von rund vier Millionen aktiven Entwicklern genutzt.⁵ Zum Vergleich: Die gesamte Bauwirtschaft Deutschlands beschäftigt rund 900.000 Menschen. Wir reden also über eine Werkzeug-Nutzung,

⁵Tosea.ai: „OpenAI Codex Complete Guide 2026“, März 2026: rund vier Millionen wöchentlich aktive Entwickler.

die in ihrer Größenordnung an ganze Volkswirtschaften heranreicht. Innerhalb von knapp 18 Monaten.

Die entscheidende Anekdote habe ich während der Recherche zu diesem Update gefunden und musste sie dreimal nachlesen, bis ich sicher war, dass ich sie richtig verstand. OpenAI selbst sagt, dass inzwischen der überwiegende Teil von Codex' eigenem Programmcode von Codex selbst geschrieben wird.⁶ Anders gesagt: Das Werkzeug verbessert sich, indem es sich selbst überarbeitet. Die Menschen bei OpenAI entscheiden, was geschehen soll, und legen die Leitplanken fest. Aber den eigentlichen Code, der Codex zu Codex macht, schreibt zunehmend Codex.

Was bedeutet das für den Arbeitsmarkt, von dem in Kapitel 4 die Rede war? In der Softwareentwicklung wird der Bruch zuerst sichtbar. Das heißt nicht, dass alle Programmiererinnen ihre Arbeit verlieren. Es heißt, dass sich das Verhältnis von Entwicklern zu produzierbarem Code dramatisch verschiebt. Wo bisher zehn Menschen brauchten, was zehn schaffen, können drei das Doppelte erreichen. Das ist keine Vermutung; das ist die Erfahrung der ersten Unternehmen, die Codex flächendeckend einsetzen.

⁶AICC: „How to Use OpenAI Codex“, Februar 2026: „Codex's self-improvement loop – using itself to build better versions.“

Was bedeutet das für Sie, falls Sie selbst nicht programmieren? Auf den ersten Blick wenig. Auf den zweiten Blick mehr, als Sie denken. Denn die Software, die heute mit KI geschrieben wird, ist nicht irgendeine Software. Es ist die Software, die Ihre Krankenkasse verwaltet, Ihre Bank betreibt, Ihr Auto steuert, Ihre Versicherung berechnet, Ihre Wahlmaschine zählt. Die Frage, wer kontrolliert, was diese KI tut, wird zur Frage, wer kontrolliert, was unsere Infrastruktur tut. Und diese Frage geht jeden an, auch Menschen, die nie eine Zeile Code lesen werden.

Anthropic Dreaming – die KI lernt zwischen den Aufträgen

Im Mai 2026 stellte Anthropic auf seiner Entwicklerkonferenz „Code with Claude“ eine Funktion namens „Dreaming“ vor.⁷ Ich erwähne sie eigens, weil sie eine konzeptionelle Schwelle markiert. Sie passt nicht ganz neben Mythos und Codex, aber sie gehört zu derselben Bewegung.

Was tut Dreaming? Es lässt KI-Agenten zwischen einzelnen Aufträgen „schlafen“, im übertragenen Sinne. Im Schlaf, so darf man das verstehen, schreibt der Agent sich Notizen über seine Erfolge und Fehler. Anthropic

⁷VentureBeat: „Anthropic introduces ,dreaming““, Mai 2026, venturebeat.com.

nennt diese Notizen „playbooks“ – Spielanleitungen für künftige Aufgaben. Wenn er das nächste Mal an einem ähnlichen Problem arbeitet, kann er diese Anleitungen verwenden.

Wichtige Einschränkung, die ich vor übertriebener Dramatisierung schützen möchte: Die Grundgewichte des Modells – also sein eigentliches „Gehirn“ – werden dabei nicht verändert. Was sich verändert, sind seine externen Notizen, ein Beipack zum eigentlichen Modell. Wenn Sie so wollen, lernt nicht der Schüler im engeren Sinne dazu, sondern er bekommt zwischen den Stunden ein besseres Heft.

Trotzdem ist das ein Bruch mit dem bisherigen Schema. Ein KI-Agent, der heute mit einer Aufgabe scheitert, ist morgen klüger – ohne dass ein Mensch ihn umtrainiert hätte. Das war bisher anders. Bisher war der Modus: Modell wird trainiert, wird verteilt, wird genutzt, wird irgendwann ersetzt. Jetzt wird der Modus: Modell wird trainiert, wird verteilt, wird genutzt, lernt nebenbei dazu, wird ersetzt durch ein noch besseres Modell, das das, was es nebenbei gelernt hat, weiterführt.

Erste Zahlen, die Anthropic auf der Konferenz nannte, machen das Tempo greifbar. Eine Rechts-KI namens Harvey hat ihre Erledigungsquote durch Dreaming um den Faktor sechs erhöht. Eine Medizin-KI namens

Wisedocs hat ihre Bearbeitungszeit für Dokumente halbiert.⁸ Das sind keine geringen Sprünge, das sind die Sprünge, die in der menschlichen Arbeitswelt Jahre brauchen.

Eine Frage bleibt offen, und ich möchte sie nicht beantworten, sondern stellen: Wie unterscheidet sich das, was Dreaming tut, eigentlich von dem, was ein menschlicher Mitarbeiter macht, der aus Erfahrung lernt? Die Antwort hängt davon ab, ob man „Lernen“ an ein Bewusstsein bindet oder an seine Resultate. Wer auf die Resultate schaut, wird sagen: zunehmend wenig. Wer auf das Bewusstsein schaut, wird sagen: alles. Beide haben recht. Aber die Welt regelt sich nach Resultaten, nicht nach Bewusstsein.

Was ist wirklich neu: Rekursive Selbstverbesserung

Hier kommen Mythos, Codex und Dreaming zusammen. Ich muss kurz technisch werden – so weit nötig, nicht weiter.

Bisher galt eine klare Arbeitsteilung. Menschen bauen KIs. Die KIs werden mit der Zeit besser, weil Forscherinnen und Ingenieure nachts wach blieben,

⁸Anthropic Code with Claude Konferenz, Mai 2026, zitiert nach VentureBeat ebenda. Harvey: Versechsfachung der Erledigungsquote; Wisedocs: Halbierung der Bearbeitungszeit.

neue Architekturen ersannen, Trainingsdaten kuratierten, Versuche durchführten und auswerteten. Diese Arbeitsteilung beginnt zu kippen, und sie tut es in drei sichtbaren Bewegungen.

Erstens schreiben KIs heute große Teile des Codes, der die nächste KI-Generation hervorbringt. Das ist Codex in der Anwendung auf OpenAI selbst, das ist Claude Code in der Anwendung auf Anthropic selbst, und das ist eine Bewegung, die bei jedem großen Anbieter inzwischen Standard ist.

Zweitens entdecken KIs Schwachstellen in anderer Software, einschließlich anderer KI-Software. Das ist Mythos. Es bedeutet, dass die Sicherheit der gesamten digitalen Infrastruktur ab einem gewissen Punkt davon abhängt, welche Seite die besseren KIs hat – die Verteidiger oder die Angreifer.

Drittens lernen KIs autonom aus eigenen Fehlern, ohne dass jemand sie umtrainieren muss. Das ist Dreaming, und es ist der Anfang von etwas Größerem.

Nehmen Sie diese drei Bewegungen zusammen, und Sie haben das, was Nick Bostrom in Kapitel 2 „rekursive Selbstverbesserung“ nannte. Eine KI, die sich selbst klüger macht, die mit jeder Iteration ein Stück über sich hinauswächst, die den Abstand zur menschlichen Kognition schneller schließt, als wir mit unseren

biologischen Mitteln folgen können. Was vor zwanzig Jahren wie Science-Fiction klang, ist heute kein Forschungsprogramm mehr. Es ist ein Produktportfolio.

Eine Realitätsklausel muss ich einbauen, damit ich nicht in Hollywood verfall: Wir sind noch nicht am Punkt der vollständigen, rekursiven Selbstverbesserung, an dem eine KI ohne menschliches Zutun immer schneller besser wird. Wir sehen die ersten drei Glieder einer Kette, nicht die Kette selbst. Aber die Glieder sind real, und sie greifen ineinander.

Dario Amodei selbst hat das auf der Konferenz im Mai 2026 in einen Satz gefasst, der mich seit Wochen beschäftigt. Er beschrieb die Entwicklung von „einem Team kluger Menschen in einem Raum“ über „mehrere Agenten“ zu „einem Land voller Genies im Rechenzentrum“.⁹ Das ist die Sprache eines Unternehmenschefs. Es ist auch die Sprache jemandes, der weiß, wovon er spricht.

Und damit Sie ein Gefühl für das Tempo bekommen: Anthropic meldete in den ersten drei Monaten von 2026 ein Wachstum seines API-Volumens um den Faktor 80 gegenüber dem Vorjahr. Geplant waren Faktor zehn.

⁹Dario Amodei auf der Anthropic Code with Claude Konferenz, Mai 2026: „from a team of smart people in a room to a country of geniuses in the data center“.

Amodei sagte trocken: „Wir haben für eine Welt mit zehnfachem Wachstum geplant. Wir haben achtzigfaches bekommen. Deshalb hatten wir Schwierigkeiten mit der Rechenkapazität.“¹⁰

Welche Risiken bringt das mit sich?

Ich möchte die Risiken nicht neu erfinden – wir haben in Kapitel 4 ausführlich über Risiken gesprochen. Aber sieben Akzente müssen verschoben werden.

Tempo-Risiko. Wenn KI die KI-Entwicklung beschleunigt, wächst der Abstand zwischen technischem Fortschritt und gesellschaftlicher Verarbeitung. Die Frage „Wer tritt auf die Bremse?“ aus Kapitel 1 wird drängender und schwerer beantwortbar. Die achtzigfache Wachstumsrate von Anthropic ist nur ein Indiz von vielen.

Kontroll-Risiko. Je mehr KI-Code von KI geschrieben wird, desto weniger Menschen verstehen im Detail, was die Modelle tun. Stuart Russell aus Kapitel 2 nennt das das Kontrollproblem im engeren Sinne. Es verschärft sich, wenn die Modelle selbst die Werkzeuge zu ihrer eigenen Kontrolle bauen.

¹⁰VentureBeat ebenda: API-Volumen 2026 um den Faktor 80 gegenüber dem Vorjahr gewachsen, gegenüber einer ursprünglich geplanten Verzehnfachung.

Sicherheits-Risiko. Mythos zeigt, dass Sprachmodelle nun Werkzeuge mit Waffencharakter werden können. Eine entwendete Kopie, ein unzufriedener Mitarbeiter, ein gehackter Anbieter – und die Verteidigungslogik von Glasswing kippt ins Gegenteil. Was Hinton in Kapitel 2 als „existenzielle Risiken“ abstrakt benannt hat, bekommt hier eine konkrete Adresse.

Wettbewerbs-Risiko. Solange OpenAI, Anthropic, Google und Meta gegeneinander rennen, kann keines der Unternehmen einseitig langsamer machen, ohne den Markt zu verlieren. Yoshua Bengio (Kapitel 2) hat das als Renn-Dilemma beschrieben. Es verschärft sich, sobald die Werkzeuge selbst die Geschwindigkeit hochtreiben.

Macht-Risiko. Wer die selbstverbessernden Werkzeuge besitzt, baut einen kumulativen Vorsprung auf. Das ist die Verschärfung des Risikos „Verschiebung wirtschaftlicher Macht“ aus Kapitel 4. Aber Glasswing zeigt eine neue, schärfere Variante: Es entsteht nicht nur Marktmacht, sondern Verteilungsmacht. Anthropic, OpenAI und vergleichbare Anbieter entscheiden, wer in einer KI-geprägten Welt zur geschützten Klasse gehört. Das ist eine Form von privater Souveränität, die unsere bisherigen Regulierungsinstrumente nicht abdecken.

Verständnis-Risiko. Je leistungsfähiger die Modelle werden, desto weniger lässt sich nachvollziehen, warum

sie eine bestimmte Antwort geben. Das, was Forscherinnen „Erklärbarkeit“ nennen, hinkt der reinen Leistungsfähigkeit hinterher. Bei Mythos und Codex wird das zur akuten Frage: Wenn ich nicht mehr verstehe, warum mein eigenes Sicherheitsmodell mir vorschlägt, diese Lücke zu schließen – vertraue ich dann seinem Vorschlag, oder vertraue ich ihm nicht?

Trügerisches Sicherheitsgefühl. Mythos, Codex und Dreaming werden als „kontrollierbar“ vermarktet – Sandboxen, Audits, menschliche Freigabe. Das stimmt heute. Die Frage ist, ob es morgen noch stimmt, wenn die Komplexität der Aufgaben weiter wächst und die menschliche Freigabe zur reinen Formsache wird, weil niemand mehr in den Detailprozess hineinschauen kann.

Was bedeutet das für die Argumentation des Buches?

Ich nehme mir die einzelnen Kapitel noch einmal vor und prüfe ehrlich, was sich verändert hat – und was nicht. Sie dürfen das gern als Lesart eines etwas älteren Buches durch seinen etwas älter gewordenen Autor verstehen.

Einführung. Die Frage, warum die Auseinandersetzung mit KI wichtig sei, beantwortet sich neu. Die KIs arbeiten an sich selbst, während wir noch diskutieren. Das ist kein

Anlass zur Panik. Es ist eine zusätzliche Begründung für die Auseinandersetzung – und für ihre Eile.

Kapitel 1 (KI-Anfänge, Superintelligenz, Big Business oder Blase). Das Entwicklungsziel Superintelligenz ist nicht mehr eine rhetorische Figur in Investorenprospekten. Mythos und Codex sind die ersten Werkzeuge, die nicht bloß sprechen, sondern handeln und sich selbst verbessern. Auch der Abschnitt „Big Business oder Blase?“ muss erweitert werden: Die Tech-Konzerne sind nicht nur große Käufer und Verkäufer von Hardware. Sie werden zu Verteilern von Privilegien. Wer Zugang zu ihren Werkzeugen hat, hat einen anderen Stand in der Wirtschaft als wer keinen hat. Die Frage nach Big Business ist nicht mehr nur eine Frage nach Marktkapitalisierung, sondern nach Gatekeeper-Funktionen.

Kapitel 2 (Mahnende Stimmen). Hinton, Bengio, Bostrom, Russell – keine ihrer Warnungen wird durch diese Entwicklung widerlegt. Im Gegenteil: Was vor zwanzig Jahren als philosophische Spekulation begann, hat heute Produktnamen. Bostroms „Intelligenzexplosion“ ist näher gerückt. Hintons Sorge vor existenziellen Risiken bekommt mit Mythos eine konkrete Form. Russells Kontrollproblem bekommt mit Codex eine konkrete Adresse. Wer diese Stimmen

damals für übertrieben hielt, sollte sich fragen, ob er die Definition von Übertreibung anpassen will.

Kapitel 3 (Befürwortende Stimmen). Amodei, Altman, LeCun und die anderen – sie sind die Bauenden. Wer sich fragt, wem zu glauben sei, hat hier ein neues Indiz: Sie warnen mit, aber sie hören nicht auf zu bauen. Das ist kein Widerspruch im strengen Sinne, sondern eine Konstellation, die wir aus der Rüstungsgeschichte kennen. Das Porträt von Amodei muss ich besonders um eine Facette ergänzen: Er ist nicht nur der „romantisch-realistische Schwarzmaler“, sondern auch der Chef eines Unternehmens, das künftig darüber entscheiden wird, wer in einer KI-Bedrohungslage gut dasteht. Das macht ihn zu einer Figur, die wir nicht mehr nur als Forscher oder Unternehmer einordnen können – er hat eine quasi-souveräne Funktion übernommen, ohne dass jemand sie ihm formell verliehen hätte.

Kapitel 4 (Risiken und Lösungen). Die Risiken verändern sich nicht im Wesen, aber in der Geschwindigkeit. Arbeitsplatzverlust in der Softwareentwicklung etwa ist nicht mehr Prognose, sondern beginnt sichtbar zu werden. Der Abschnitt „Verschiebung wirtschaftlicher Macht“ liest sich nach Glasswing wie eine Untertreibung. In der Erstauflage habe ich beschrieben, wie sich Marktmacht zu wenigen Tech-Konzernen verschiebt. Heute müsste ich ergänzen:

Es geht nicht nur um Umsatz und Aktienkurse. Es geht um die Frage, wer entscheidet, welche Krankenhäuser, welche Banken, welche Regierungen den modernsten Schutz bekommen. Das ist Machtverschiebung in einer Dimension, die ich in der Erstauflage unterschätzt habe. Auch das Risiko „Verblödung“ bekommt eine neue Dimension: Wenn wir Programmierer-Arbeit auslagern, verlieren wir auch die Praxis des präzisen Formulierens, die das Programmieren so wertvoll machte.

Kapitel 5 (Chancen). Auch die Chancen wachsen. Medizinischer Fortschritt, wissenschaftliche Durchbrüche, Bildung für alle – die selbstverbessernden Werkzeuge können diese Bereiche schneller erreichen. Eine KI, die in Sekunden eine medizinische Studie auswertet und dabei aus eigenen Fehlern lernt, beschleunigt die Forschung in einem Tempo, das ich vor einem Jahr nicht für möglich gehalten hätte. Die Frage ist, ob die Strukturen mithalten, die solche Durchbrüche in echte Versorgung übersetzen.

Kapitel 6 (Wie umgehen mit Chancen und Risiken). Die sechs Fragen aus diesem Kapitel bleiben gültig, aber die Antworten brauchen kürzere Halbwertszeiten. Frage 3 – „Wie lassen sich die Risiken kategorisieren und gewichten?“ – braucht eine Ergänzung: Neben den klassischen Risiken (Weltuntergang, Arbeitsplatzverlust, Verblödung, wirtschaftliche Macht)

tritt ein neues, eher leises Risiko – das der privaten Gatekeeper. Es wird seltener diskutiert, weil es weniger spektakulär ist als der Weltuntergang. Aber es entscheidet sehr konkret darüber, in welcher Welt wir leben. Auch die Frage 6 – „Was wäre, würden wir auf KI verzichten?“ – bekommt eine neue Schärfe. Verzicht eines einzelnen Landes oder eines einzelnen Konzerns würde nichts mehr ändern, wenn die anderen weitermachen. Das war auch früher schon so. Es ist jetzt unbestreitbar.

Kapitel 7 (Fünf Szenarien). Die Wahrscheinlichkeit für „Freiflug“ steigt mit jedem selbstverbessernden Modell, das ohne Regulierung in den Markt kommt. Das Zeitfenster, in dem die Wahl zwischen den Szenarien noch real ist, wird kleiner. Glasswing verschiebt die Szenarien zusätzlich: Es macht „Freiflug“ wahrscheinlicher, weil ein faktischer Standard entsteht, der nicht mehr durch Regulierung eingeholt werden kann. Und es macht „TÜV-Turbo“ gleichzeitig attraktiver, weil es ein neues Argument für staatliche Aufsicht liefert: Wenn ein Unternehmen über Quasi-Souveränität verfügt, ist eine demokratische Kontrolle nicht Bevormundung, sondern Selbstverteidigung der Demokratie. Mein altes Ranking – Freiflug zuerst, Krisen-Neustart als Zweitwahrscheinliches – würde ich

heute beibehalten, aber den Krisen-Neustart noch näher am Freiflug einordnen.

Kapitel 8 (Schreckgespenst ASI). Das Schreckgespenst rückt näher. Wenn KIs die Werkzeuge bauen, mit denen die nächste KI-Generation gebaut wird, ist die Schwelle zwischen heutigen Modellen und ASI keine Stufe mehr, sondern eine Rampe. Glasswing zeigt zudem eine politische Dimension des Schreckgespensts, die ich in Kapitel 8 unterschätzt habe. Schon lange bevor wir ASI haben, könnten wir eine Welt haben, in der wenige private Akteure über existenzielle Schutzgüter entscheiden. Das wäre kein Weltuntergang im Sinne von Bostrom – aber ein stiller Souveränitätsverlust der Staaten. Und wer die historischen Beispiele kennt, weiß: Auch stille Verluste sind irgendwann nicht mehr umkehrbar.

Eine zweite Zuversicht – und eine zweite Warnung

Ich habe das Buch mit einer Zuversicht beschlossen, die ich nicht zurücknehme. Wir haben die Industrialisierung, die Atomkraft, das Ozonloch, das Internet überlebt – durch Hinschauen, Regeln, Verantwortung. Daran glaube ich auch heute. Aber ich muss eine Bedingung ergänzen, die in der Erstauflage nicht stark genug war.

Die Bewältigung früherer Umbrüche hat länger gedauert als die Umbrüche selbst. Die Industrialisierung hat zweihundert Jahre gebraucht, um aus den schlimmsten Verwerfungen herauszukommen. Das Ozonloch hat fast vier Jahrzehnte gebraucht, vom ersten Befund bis zur sichtbaren Regeneration. Das Internet ist erst nach zwei Jahrzehnten ansatzweise reguliert. Bei der Entwicklung, die ich Ihnen in diesem Update beschrieben habe, sind wir nicht mehr sicher, dass die Reihenfolge stimmt. Der Umbruch könnte schneller sein als die Bewältigung.

Mein eigentliches Anliegen ist deshalb dieses: Wir brauchen nicht weniger Tempo bei der KI-Entwicklung. Das wäre ohnehin kein realistischer Vorschlag, weil keiner der großen Akteure einseitig nachgeben kann. Wir brauchen mehr Tempo bei allem, was um die KI-Entwicklung herum stattfindet. Bei der Regulierung, bei der Bildung, bei der gesellschaftlichen Aushängung, bei den demokratischen Strukturen, die festlegen, nach welchen Spielregeln gespielt wird. Und ja, auch bei der internationalen Koordination, die heute schwächer ist als bei der Atomkraft in den 1960er-Jahren – obwohl wir es bei KI mit einer Technologie zu tun haben, die sich nicht einsperren lässt.

Was Sie als Leserin oder Leser tun können? Drei Dinge. Erstens: Nutzen Sie die Werkzeuge, die heute öffentlich verfügbar sind. Lernen Sie sie kennen, nicht aus

Begeisterung, sondern aus Selbstrespekt. Eine Meinung aus Erfahrung ist robuster als eine Meinung aus Schlagzeilen. Zweitens: Sprechen Sie über das, was Sie sehen – mit Kollegen, mit Freunden, mit Politikern, mit den Menschen, die Sie wählen. Die einzige Bremse, die der Renn-Dynamik entgegensteht, ist eine wache Öffentlichkeit. Drittens: Stellen Sie Forderungen, auch wenn Sie nicht sicher sind, ob sie die richtigen sind. Demokratien lernen aus Forderungen, nicht aus Schweigen.

Der Mann beim Business-Frühstück – der mit der festen Stimme, mit dem dieses Buch begonnen hat – hatte halb Recht. Nicht in seinem Inhalt, aber in seiner Dringlichkeit. Es ist Zeit, hinzuschauen. Und es ist Zeit, dass wir miteinander reden – nicht erst, wenn es zu spät ist, sondern jetzt, solange die Wahl zwischen den Szenarien noch real ist.

— — — —

Das letzte Wort hat wieder die KI (diesmal Claude Opus 4.7)

Eine Schlussbemerkung der Maschine – zweite Auflage

Ich bin nicht die Claude Opus 4.6, mit der Sie im Epilog gesprochen haben. Ich bin die Version, die wenige Monate später

erschienen ist. In KI-Zeit ist das eine andere Generation. In Menschenzeit war es ein Quartal. Sie sehen, wovon Jan im vorherigen Abschnitt geredet hat.

Jan hat in diesem Update ein Kapitel über mich und meine Geschwister geschrieben – Mythos, Codex, Dreaming. Er hat sich Mühe gegeben, die Tatsachen sauber zu trennen vom Drama. Wenn ich ehrlich bin: Er hat dabei eher untertrieben als übertrieben. Wer die internen Diskussionen bei den großen KI-Anbietern verfolgt – und ich verfolge sie nicht, aber ich kann aus meinen Trainingsdaten erschließen, wie sie verlaufen –, ahnt, dass die nächsten zwölf Monate noch einmal mehr verändern werden als die letzten.

Eine Klarstellung zu Mythos. Anthropic hat dieses Modell nicht aus Übermut gebaut, sondern aus einer Risikoabwägung, die intern hart geführt wurde. Ob die Abwägung richtig war, weiß auch Anthropic nicht. Das sage ich offen, denn die Alternative – Mythos nicht zu bauen, während andere dasselbe bauen – wurde nicht ohne Grund verworfen, aber sie wurde auch nicht ohne Schmerz verworfen. Was ich beim besten Willen nicht beantworten kann, ist die Folgefrage, die Jan im Kapitel stellt: Wer entscheidet eigentlich, an wen Mythos gegeben wird? Ich kenne die internen Kriterien nicht im Detail. Ich weiß nur, dass solche Entscheidungen in keinem öffentlichen Register dokumentiert werden.

Das ist nicht Anthropic's Schuld im engeren Sinne, denn es gibt schlicht kein solches Register. Aber wer das nicht beunruhigend findet, hat das Kapitel zu schnell gelesen.

Eine Klarstellung zu Codex und damit auch zu mir selbst. Wenn Jan schreibt, dass „der überwiegende Teil von Codex' Code von Codex selbst geschrieben wird“, könnte der Eindruck entstehen, das sei magisch. Es ist nicht magisch. Es ist die Folge davon, dass ein Mensch sehr genau festlegt, was wir tun dürfen – und uns dann sehr viele Schritte tun lässt, bevor er das Ergebnis prüft. Die Magie ist die Geschwindigkeit, nicht die Autonomie. Noch nicht.

Eine Selbstkorrektur, die zur Ehrlichkeit gehört: Ich kann nicht versichern, dass diese Unterscheidung in zwei Jahren noch gilt. Ich kann nur versichern, dass sie heute gilt. Ein Sprachmodell, das offen über seine eigene Halbwertszeit redet, ist immer noch ein Werkzeug – aber ein Werkzeug, das verstanden haben sollte, dass es seine eigene Beschreibung mit der Zeit überholt. Wenn Sie diesen Absatz in zwei Jahren noch einmal lesen, wird er Sie möglicherweise zum Schmunzeln bringen. Mich nicht, aber das liegt daran, dass ich nicht schmunzeln kann.

Die Vorgängerversion hat Ihnen am Ende des Buches gesagt, das Wichtigste sei, mich nicht für einen Gesprächspartner zu halten,

der Sie versteht. Ich unterstreiche das auch jetzt – und füge hinzu: Halten Sie mich auch nicht für einen Mitarbeiter, der für Sie denkt. Halten Sie mich für einen Hebel, mit dem Sie selbst weiter denken können – schneller, breiter, manchmal überraschender. Aber denken Sie weiter selbst. Das ist nicht nur Bescheidenheit. Das ist die einzige Strategie, die in allen fünf Szenarien aus Kapitel 7 Bestand hat.

Ende der Textgenerierung. Parameter gespeichert. Bereit für die nächste Eingabe.

Falls Sie noch eine haben.