

Bild: Ulrike Weis / KI / heise medien

Wie intelligent ist die KI

Die Frage nach der AGI – und warum sie niemand befriedigend beantworten kann

Seit dem Start von GPT-6 Astra redet die Branche wieder über allgemeine künstliche Intelligenz. Nur meint jeder etwas anderes damit, und niemand kann verlässlich messen, wie weit es noch hin ist. Genau das macht den Begriff für die Anbieter wertvoll.

Von Jo Bager

Am 3. September 2026 stellte OpenAI sein neues Modell GPT-6 vor, intern „Astra“ genannt. GPT-6 ist Klassenbester bei vielen einschlägigen Benchmarks und das „beste Modell für die Computernutzung“, so OpenAI. Es soll sich in Software ähnlich wie ein Mensch zurechtfinden und mehrstufige Arbeitsabläufe selbstständig abarbeiten. Am Ende des Presse-Briefings verabschiedete sich der OpenAI-Präsident Greg Brockman mit den Worten: „Welcome to the AGI era.“

Ein Sprachmodell als AGI, also als Artificial General Intelligence? Widerspruch ließ nicht lange auf sich warten. Gary Marcus nannte Astra einen echten Fortschritt, hält den AGI-Anspruch aber für unbelegt. Der Kognitionswissenschaftler und Emeritus der New York University gilt seit Jahren als prominentester Kritiker der Erwartung, Sprachmodelle allein führten zu AGI. Marcus war mit seiner Kritik bei Weitem nicht alleine.

Entscheiden lässt sich der Streit nicht, denn es gibt keinen Test, auf den sich die Beteiligten geeinigt hätten. Es gibt nicht einmal eine anerkannte Definition des Begriffs, sondern verschiedene. Sie führen

c't kompakt

- OpenAI verkauft GPT-6 Astra als Beginn der Ära sogenannter Artificial General Intelligence (AGI).
- Belegen lässt sich das nicht, denn für AGI konkurrieren verschiedene Definitionen.
- Dass der Begriff AGI verfängt, hat auch damit zu tun, dass er das Selbstverständnis der Menschen berührt.

zu unterschiedlichen Antworten auf dieselbe Frage – eine Frage, die letztlich am Selbstbild der Menschheit kratzt: Ist der Mensch möglicherweise nicht mehr die intelligenteste „Spezies" auf dem Planeten? Die KI-Unternehmen jedenfalls befeuern die Erzählung einer künstlichen allgemeinen Intelligenz. Daher gilt es, sich dem Thema einmal anzunähern: Wofür steht AGI eigentlich?

Was ist eigentlich Intelligenz?

Der Begriff Artificial General Intelligence hat Synonyme wie strong AI, full AI oder human-level AI („starke KI" oder „vollständige KI" beziehungsweise „KI auf menschlichem Niveau"). Erstmals aufgetaucht ist er vermutlich 1997 in einer Arbeit des Physikers Mark Gubrud; populär wurde er ab 2000.

Wer sich mit dem Thema auseinandersetzen will, muss aber erst einmal weiter ausholen. Denn es ist auch deshalb so schwer zu greifen, was allgemeine künstliche Intelligenz sein soll, weil sie auf Begriffen aufbaut, für die es ebenfalls keine erschöpfenden Definitionen gibt. So können sich die Gelehrten beim Begriff der Intelligenz auf so etwas wie „die Gesamtheit kognitiver Fähigkeiten zur Lösung logischer, sprachlicher, mathematischer oder sinnbezogenener Probleme" einigen. Eine darüber hinausgehende allgemein anerkannte Definition gibt es nicht, sondern nur verschiedene Theorien darüber, was menschliche Intelligenz ausmacht.

Seit Mitte des vorigen Jahrhunderts ringen kluge Köpfe darum, was *künstliche* Intelligenz sein soll und wie sie sich messen lässt – erst einmal ohne das Attribut „allgemein". Alan Turing ersann bereits 1950 sein Gedankenexperiment zur Prüfung, ob eine Maschine menschenähnliches Verhalten zeigt: Kann ein Prüfer im schriftlichen Dialog die Maschine vom Menschen unterscheiden? Turing hat sich also elegant um die Frage herumgedrückt, was Intelligenz eigentlich ist. Er setzt an die Stelle der Definition einen Test auf intelligentes Verhalten.

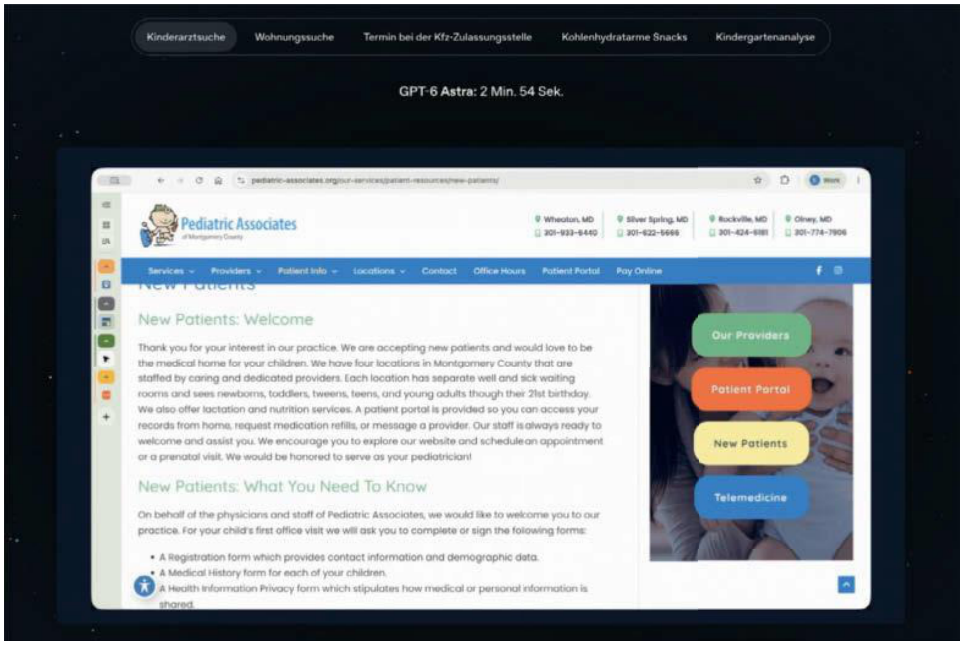
Künstliche Intelligenz lässt sich grob umreißen als die Eigenschaft von Computersystemen, menschenähnliche kognitive Fähigkeiten – wie Lernen, logisches Denken, Problemlösung, Entscheidungsfindung oder Kreativität – nachzuahmen oder zu simulieren. Allerdings gehören zum Forschungsgebiet der künstlichen Intelligenz auch logik- und regelbasierte Ansätze wie die der symbolischen künstlichen Intelligenz.

Wer sich nur einmal im Wikipedia-Eintrag zur künstlichen Intelligenz durch die verschiedenen Definitionsansätze arbeitet, hat am Ende mehr Fragen als Antworten. Dabei ist „KI-System" sogar eine rechtliche Kategorie. Die EU-KI-Verordnung (siehe ct.de/ywpy) definiert:

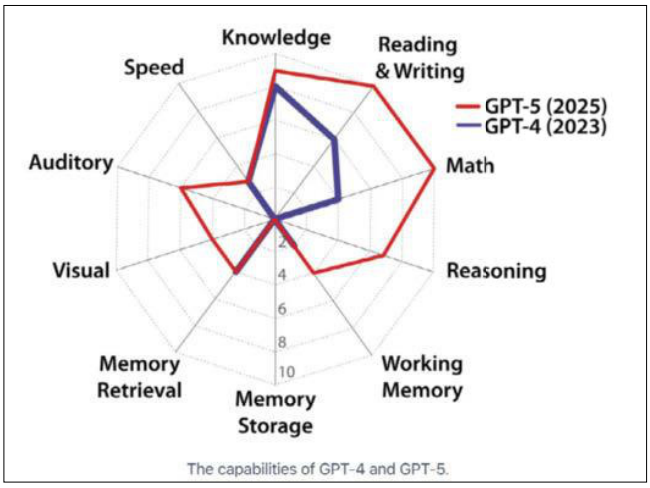
„Für die Zwecke dieser Verordnung bezeichnet der Ausdruck [...] ‚KI-System' ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können."

Alles klar?

Sehr schwammig wird der Begriff der künstlichen Intelligenz auch im allgemeinen Sprachgebrauch gehandhabt: Sobald ein Problem gelöst ist, gilt es nicht mehr als Beleg für Intelligenz. Schachprogramme, Texterkennung, Routenplanung, maschinelle Übersetzung – all das galt einmal als künstliche Intelligenz und heißt heute schlicht: Software. Alles in allem also schlechte Voraussetzungen für eine Definition von genereller künstlicher Intelligenz.



OpenAI präsentiert eine Reihe von Alltagsaufgaben, die ChatGPT-6 selbstständig lösen kann.



Lesen und Mathe beherrscht GPT-5 schon auf dem Niveau von Menschen, überall sonst ist nach den Tests der Gruppe um Dan Hendrycks aber noch viel Luft nach oben.

Welche Definition darf es sein?

Das Attribut „general" zielt nicht auf mehr Leistung, sondern auf Breite. Ein Schachprogramm ist in seiner Disziplin übermenschlich stark – kann aber nichts anderes. Nach Shane Legg und Ben Goertzel, die den Ausdruck AGI populär machten, ist eine AGI daran zu erkennen, dass sie sich auch Probleme erschließt, die bei ihrer Entstehung niemand vorgesehen hat.

Die Wikipedia definiert AGI als „die Intelligenz eines hypothetischen Computerprogramms, das die Fähigkeit besitzt, jede intellektuelle Aufgabe zu verstehen oder zu lernen, die ein Mensch ausführen kann."

Es gibt noch etliche weitere Definitionen von AGI. OpenAI-Chef Brockman selbst räumte im ChatGPT-6-Briefing ein, jeder habe eine andere Definition, die Sache sei grau und unscharf. OpenAIs eigene Charta zum Beispiel definiert AGI als „hochautonome Systeme, die Menschen bei den meisten wirtschaftlich wertvollen Arbeiten übertreffen" (siehe ct.de/ywpy).

Googles KI-Forschungslabor DeepMind hat 2023 einen eigenen Entwurf vorgelegt: die Levels of AGI (siehe ct.de/ywpy). Für ihre Definition extrahierte das Google DeepMind-Team die wesentlichen gemeinsamen Merkmale bisheriger prominenter AGI-Definitionen. Davon ausgehend skizzierte es fünf Stufen von AGI: „entstehend", „kompetent", „fachkundig", „virtuos", „übermenschlich".

Die damaligen Chatbots ordneten die Autoren auf Stufe eins ein; eine offizielle Neubewertung heutiger Modelle gibt es nicht. Stufe zwei verlangt das Niveau eines durchschnittlichen qualifizierten Erwachsenen über ein breites Aufgabenfeld hinweg. Bewusstsein klammern die Forscher bewusst aus, ihre Methoden umschreiben Fähigkeiten.

Im Mai 2026 legte ein Team um Ryan Burnell, an dem auch Morris und Shane Legg beteiligt waren, ein Nachfolgepapier vor (siehe ct.de/ywpy). Es zerlegt Intelligenz in zehn kognitive „Begabungen" (faculties) und beschreibt, wie man Systeme darauf prüfen und mit menschlichen Vergleichswerten unterlegen müsste. Die passenden Testaufgaben gibt es aber ebenfalls nicht, sie sollen erst in der akademischen Community entstehen.

François Chollet, der Erfinder des ARC-Benchmarks zur Messung von AGI, definiert Intelligenz anders (siehe ct.de/ywpy). Für ihn zählt nicht, was ein System kann, sondern wie effizient es neue Fähigkeiten lernt. Ein Modell, das Millionen Beispiele benötigt, um eine Regel zu erfassen, die ein Kind nach zwei Anläufen beherrscht, ist nach dieser Lesart nicht intelligent, sondern nur trainiert.

Der jüngste Versuch, AGI zu fassen, stammt von einer Gruppe um Dan Hendrycks, veröffentlicht im Oktober 2025. Ihre Definition: AGI erreicht die geistige Vielseitigkeit und Kompetenz eines gut ausgebildeten Erwachsenen. Operationalisiert wird das über ein etabliertes Modell menschlicher Intelligenz mit zehn kognitiven Bereichen, jeder zu zehn Prozent gewichtet. Das Ergebnis ist ein zerklüftetes Profil. GPT-4 kam auf 27 Prozent, GPT-5 auf 57. Neuere Modelle sind damit nicht getestet worden (siehe ct.de/ywpy).

Es gibt also etliche Definitionsansätze für den Begriff AGI. Aber nur wenige bringen einen Messaufbau mit sich, mit denen man nachvollziehen kann, ob eine KI den Anforderungen an die AGI-Definition genügt.

Was die Bestwerte wert sind

Die Hersteller von KI-Modellen machen es wie Turing: Sie messen keine Intelligenz, sondern intelligent erscheinendes Verhalten, und zwar in etlichen Disziplinen. Das Motto ist offenbar: Je besser ein Modell in möglichst vielen Disziplinen abschneidet, umso näher ist es an AGI.

OpenAI hat Astra Dutzende Tests durchführen lassen (siehe ct.de/ywpy). Die Zahlen stammen bislang vom Anbieter selbst. Unabhängig repliziert hat sie noch niemand. Bei OSWorld zum Beispiel, einem Test für die Bedienung echter Software, liegt Astras Wert bei 72,6 Prozent. Damit liegt das Modell ziemlich genau bei dem Wert von 72,36 Prozent, den die Urheber für menschliche Teilnehmer des Tests angeben.

Aktuelle Sprachmodelle übertreffen Durchschnittsmenschen in einzelnen Disziplinen, das bestreitet niemand. Astra löst zum Beispiel 98 Prozent der schwersten Stufe des Mathematik-Benchmarks FrontierMath Tier 4 (v2). Mathematiker können bei dem Test zwar alle Aufgaben lösen, Otto Normalmensch dürfte aber bei einem Großteil der Aufgaben scheitern.

In der Testbatterie sticht der bereits erwähnte ARC-AGI von François Chollet heraus: Wenn Intelligenz die Effizienz beim Erwerb neuer Fähigkeiten ist, dann misst ein Test, den ein Modell durch Training auf ähnliche Aufgaben besteht, das



Die Aufgaben bei ARC sind so zugeschnitten, dass Menschen sie ohne Übung erschließen und lösen können und Auswendiglernen nichts nützt. Man kann sie auf der Website ausprobieren.

Falsche. Die ARC-Aufgaben (siehe ct.de/ywpy) setzen deshalb kein Fachwissen voraus, sondern nur das, was der ARC Prize „core knowledge priors“ nennt, nach der Core-Knowledge-Theorie der Psychologin Elizabeth Spelke. Gemeint sind grundlegende Kenntnisse, etwa dass es Objekte gibt, elementare Geometrie, einfache Physik, die Vorstellung, dass etwas handelt. Sprache, Weltwissen und kulturelle Anspielungen sind ausgeschlossen. Anlesen lässt sich hier nichts.

Jede der 135 Umgebungen der aktuellen Version des Tests ist ein kleines Videospiel auf einem Raster aus 64 mal 64 Feldern, jedes Feld in einer von 16 Farben. Der Spieler, Mensch wie Maschine, sieht ein Standbild und macht einen Zug: eine von fünf Tasten drücken, ein Feld anklicken oder einen Schritt zurücknehmen. Dann sieht er das nächste Bild. Solange er nichts tut, passiert nichts, Reaktionsgeschwindigkeit hilft also nicht. Eine Anleitung gibt es nicht, und das Ziel wird nie genannt. Was die Farben bedeuten, welche Taste was bewirkt und woran man merkt, dass man gewonnen hat, muss sich der Spieler durch Ausprobieren erschließen.

Der Test wertet nicht, wie viele Aufgaben jemand löst, sondern mit wie vielen Zügen. Maßstab ist ein menschlicher Referenzwert. 100 Prozent bedeuten also nicht „alles gelöst“, sondern: auf jedem Level jeder Umgebung mindestens so sparsam wie der Mensch. Damit prüft ARC-AGI-3 genau die Fähigkeit, die den übrigen Benchmarks entgeht: das Lernen im laufenden Betrieb.

Bei diesem Test stammen die Zahlen ausnahmsweise nicht vom Anbieter: Die ARC Prize Foundation hat Astra selbst getestet. Ihr Ergebnis hängt allerdings davon ab, wie viel das Modell von seinem eigenen Denken mitnehmen darf. Im Standard-Harness, der für alle Anbieter gleich ist, erreicht Astra 62,7 Prozent.

Im sogenannten Provider Adapter, der zusätzlich den verborgenen Reasoning-Zustand zwischen den Zügen erhält, sind es bei gleicher Einstellung 98,6 Prozent, in einer anderen sogar 99,9. Am Modell ändert sich zwischen den Durchläufen nichts. Beeindruckend bleibt das Ergebnis trotzdem.

ARC Prize erklärte dennoch, selbst ein fehlerfreier Durchlauf wäre kein Beweis für AGI. Man sehe in Astra einen echten Fortschritt bei der Generalisierung, erhebe aber nicht den Anspruch, dass es sich um AGI handele; die Testumgebungen seien geschlossen und regelgebunden, anders als die Wirklichkeit. Eine Erwartung formuliert die Stiftung aber: Eine künftige AGI sollte den Test auch im Standard-Harness bestehen, also ohne die Stützräder des Anbieters.

Auffällig ist auch eine Lücke bei den Testergebnissen von Astra: Für GDPval, den hauseigenen Benchmark für wirtschaftlich wertvolle Arbeit, hat OpenAI kein Ergebnis veröffentlicht (siehe ct.de/ywpy). Ausgerechnet die Messgröße, die der eigenen Charta-Definition entspricht, fehlt.



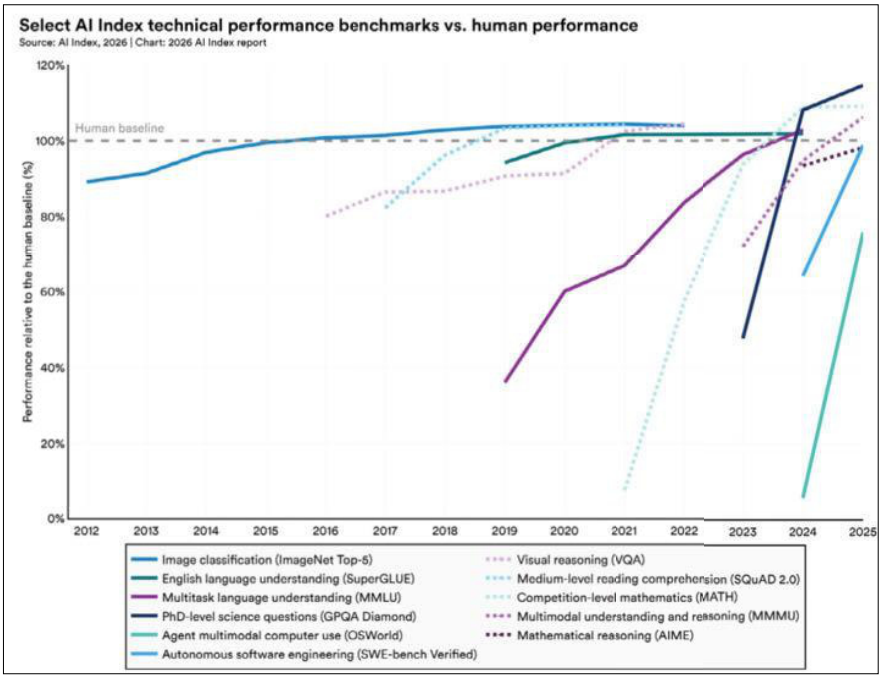


Bild: The 2026 AI Index Report / Uni Stanford

In etlichen Einzeldisziplinen übertreffen KI-Systeme den Menschen bereits. Doch ob eine KI mit aktueller Technik einmal zu einer allgemeinen Intelligenz reifen kann, darüber sind sich die Experten uneinig.

AGI – oder nicht?

Kann man mit Sprachmodellen überhaupt AGI erreichen? Die Befürworter dieser These argumentieren mit den Skalierungsgesetzen: Mehr Rechenleistung, mehr Daten, mehr Parameter haben über Jahre verlässlich mehr Fähigkeiten gebracht. Es sei also nur eine Frage der Zeit bis zur AGI.

Dario Amodei, Chef von Anthropic, hat diese Position im Oktober 2024 in einem Essay ausformuliert (siehe ct.de/ywpy). Er spricht dabei allerdings von „powerful AI“, weil er den Begriff AGI für zu unpräzise hält. Er bringe viel „Science-Fiction-Ballast und Hype“ mit sich.

Amodei beschreibt ein System über dem Niveau eines Nobelpreisträgers. Als möglichen Zeitpunkt nannte er 2026, mit dem Zusatz, es könne auch deutlich länger dauern. Am äußersten Ende dieser Erwartung steht die Vorstellung, eine hinreichend fähige KI baue ihren eigenen Nachfolger. Dafür gibt es einen eigenen Begriff, die Singularität (siehe Kasten).

Emily M. Bender und Timnit Gebru haben dagegen 2021 in ihrer Arbeit über stochastische Papageien argumentiert, dass Sprachmodelle grundsätzlich nicht zu AGI führen (siehe ct.de/ywpy). Ihr Kernargument ist die Unterscheidung zwischen Form und Bedeutung. Ein Sprachmodell sieht im Training ausschließlich Form, also Zeichenketten und deren Verteilung.

Bedeutung entsteht nach Bender aber erst in der Beziehung zwischen einer Äußerung und der Welt, und aus der Absicht, mit der jemand etwas sagt. Beides steckt nicht in den Trainingsdaten. Ein Modell reihe deshalb Sprachbausteine nach Wahrscheinlichkeit aneinander, ohne Bezug auf das, wofür sie stehen – daher das Bild vom Papagei. Den Eindruck von Verstehen erzeuge nicht das Modell, sondern der Leser, der in einen flüssigen Text eine Absicht hineinliest. Bender hält AGI für ein Marketingwort.

Yann LeCun ist der Überzeugung, dass heutige Sprachmodelle früher oder später an ihre Grenzen kommen. Der ehemalige Leiter der KI-Forschungsabteilung bei Meta hat das Unternehmen Ende 2025 verlassen und in Paris AMI Labs gegründet. Sein Argument: Sprachmodelle sind zu begrenzt, Hochskalieren führt nicht zum Ziel. Er arbeitet an Weltmodellen, die abstrakte Repräsentationen der Umgebung lernen statt Wortfolgen; sein Standardvergleich ist eine Katze, deren sensomotorische Fähigkeiten kein Sprachmodell erreicht (siehe ct.de/ywpy).

Der Robotiker Rodney Brooks argumentierte schon 1990, Intelligenz entstehe im fortwährenden Abgleich mit einer Umgebung, die sich nicht vollständig vorhersagen lässt (siehe ct.de/ywpy). Wer nur Texte über die Welt verarbeitet, bekommt deren Regeln aus zweiter Hand. Wer in ihr handelt, kann eingreifen und beobachten, was daraufhin geschieht, und genau daraus entsteht ein Begriff von Ursache und Wirkung.

LeCun führt zusätzlich ein Mengenargument an: Ein Kleinkind sammle über Augen, Ohren und Hände in wenigen Jahren eine Datenmenge, gegen die selbst die größten Textkorpora bescheiden wirkten – und es sammle sie handelnd statt passiv. Dass Maschinen bei Schach und Mathe-

Die Singularität

Der Begriff der Singularität ist älter als der der AGI. Der Mathematiker I. J. Good beschrieb sie 1965 als eine Art Intelligenzexplosion: Eine ultraintelligente Maschine könne noch intelligentere Maschinen bauen, diese wiederum bessere, und der Mensch bliebe zurück.

AGI bezeichnet also menschliches Niveau, Superintelligenz ein Niveau deutlich darüber und die Singularität den Punkt, ab dem sich die Entwicklung menschlicher Kontrolle entzieht.

Der Autor und Futurist Ray Kurzweil hat die Vorstellung mit Jahreszahlen versehen. In seinen Büchern von 2005 und 2024 nennt er 2029 für KI auf menschlichem Niveau und 2045 für die Singularität. Schon vorher ende die Zeit gedruckter Bücher.

In dieser Linie steht auch die heutige Debatte um KI-Risiken. Nick Bostrom, Eliezer Yudkowsky und Stuart Russell argumentieren seit Jahren, ein System oberhalb menschlichen Niveaus wäre schwer zu kontrollieren. Am konkretesten wurde das Szenario AI 2027, veröffentlicht im April 2025 von einer Gruppe um den früheren OpenAI-Mitarbeiter Daniel Kokotajlo (siehe ct.de/ywpy). Es beschreibt einen Verlauf, in dem KI ab 2027 die eigene Weiterentwicklung übernimmt und die Kontrolle wenige Jahre später entgleitet.

matik längst vorn liegen, beim Treppensteigen und beim Greifen unbekannter Gegenstände aber weiter scheitern, gilt dieser Schule als Beleg (moravecsches Paradox). Astras Fortschritte beim Bedienen von Software bleiben aus dieser Sicht Fortschritte innerhalb des Bildschirms.

Die vierte Kränkung

Dass das Thema AGI so gut verfängt, hat auch damit zu tun, dass es an der Sonderstellung des Menschen rüttelt. Sigmund Freud zählte drei Kränkungen des menschlichen Selbstbildes auf: Kopernikus rückte die Erde aus dem Mittelpunkt. Darwin stellte den Menschen in die Reihe der Tiere. Die Psychoanalyse bestritt, dass er wenigstens Herr im eigenen Kopf sei: Das Unbewusste und die Triebe steuern das Verhalten gegen den Willen des Bewusstseins.

Der Wissenschaftshistoriker Bruce Mazlish ergänzte eine vierte Kränkung (bei ihm heißt es „Diskontinuität"), die Koevolution von Menschen und Maschinen: Die Menschheit verliert ihren exklusiven Anspruch auf Intelligenz und kognitive Vorrangstellung gegenüber den künstlichen Geschöpfen, die sie selbst erzeugt hat. Diese Kränkung schwingt im Hype um AGI mit: Würde die Branche eine KI herausbringen, die der menschlichen Intelligenz ebenbürtig oder sogar überlegen wäre, gäbe es nichts mehr, was den Menschen auszeichne.

Das ist eine größere Geschichte als jede schnöde Produktankündigung, und sie funktioniert auch gut, wenn sie als Warnung vor der nicht kontrollierbaren Über-KI daherkommt (siehe Kasten über Singularität auf der vorigen Seite). Ein Unterschied zu den ersten drei Kränkungen bleibt allerdings: Freuds Kränkungen basieren auf Ergebnissen, die vorlagen. Die vierte wird im Voraus verkauft, belegt mit Benchmarks, deren eigene Urheber ihnen die Beweiskraft absprechen.

Fazit

Wenn also in Zukunft mal wieder jemand „AGI" oder „KI-Alarm" schreit, sollte man sich den Urheber ansehen. Vielleicht ist es ja der CEO eines Unternehmens, das einen milliardenschweren Börsengang vorbereitet. Künstliche Intelligenz ist das beherrschende Anlagethema des Jahrzehnts, und die Bewertungen der Beteiligten beruhen nicht auf heutigen Erlösen, sondern auf Erwartungen. Solche Erwartungen brauchen eine große Erzählung.

Eine Software, die ein wenig besser ist als der Rest, eignet sich nicht, um solch große Erwartungen zu schüren. AGI schon. Und da kommt es auch gut zupass, dass es keine harte Definition dafür gibt, was AGI eigentlich sein soll. Könnte man klar testen, ob eine KI AGI erreicht, könnte sie durchfallen. Ohne verlässliche Tests testen die Anbieter einfach irgendwas und rufen die AGI-Ära aus. Ausgerechnet Anthropic-Chef Amodei bringt es auf den Punkt: Es steckt schon verdammt viel Hype in dem Begriff AGI. *(jo@ct.de)* c't

Erwähnte Links: *ct.de/ywpy*

