Handelsblatt
DONNERSTAG, 29. JANUAR 2026, NR. 20

Stuart Russell

„Es ist Zeit, den CEO loszuwerden“

Der Informatiker von der Universität in Berkeley erklärt, warum die Technologie bald auch Top-Manager ersetzen könnte – und wann sie zu einer existenziellen Bedrohung wird.

Herr Russell, Sie haben das weltweit einflussreichste Lehrbuch für Künstliche Intelligenz mitgeschrieben und gelten heute als einer der prominentesten Warner. Wann genau ist aus dem Optimisten der Skeptiker geworden?
Ich glaube nicht, dass „Skeptiker“ das richtige Wort ist. Es ist eher ein Kompliment an die KI, wenn ich sage: Sie könnte so leistungsfähig werden, dass sie zur Bedrohung wird – es sei denn, man entwickelt sie richtig. Jemand, der dafür sorgt, dass Atomreaktoren nicht explodieren, ist ja auch kein Skeptiker gegenüber Kernenergie. Das ist einfach gesunder Menschenverstand.

Wann wurden Ihnen die Gefahren bewusst?
Ich arbeite seit rund 50 Jahren an KI. Die erste Ausgabe des Lehrbuchs erschien 1994, und schon damals gab es am Ende ein Kapitel mit dem Titel: „Was ist, wenn wir Erfolg haben?“ Damals schien mir erstens, dass wir noch sehr, sehr weit entfernt sind von KI-Systemen, die überhaupt ein Risiko darstellen könnten. Und zweitens, dass wir unterwegs – sobald die Risiken sichtbar werden – schon Wege finden würden, sie zu verhindern, also Ansätze entwickeln, die mit echten Garantien einhergehen.

Und wann hat sich das geändert?
Von 2012 bis 2014 war ich für ein Sabbatical in Paris und hatte viel Zeit, darüber nachzudenken, wohin sich das Feld entwickelt. Mir wurde klar: Wir hatten keinen echten Fahrplan, wie wir – nennen wir es einfachheitshalber allgemeine Künstliche Intelligenz (AGI) – Systeme erreichen, die Menschen in jeder Dimension mindestens ebenbürtig sind. Aber ich konnte sehen, dass sich solch ein Fahrplan abzeichnen könnte.

Und Ihre Reaktion damals?
Ich begann, diese Frage wirklich ernst zu nehmen: Was ist, wenn wir Erfolg haben? Wie behalten wir die Kontrolle? Also habe ich angefangen, genau daran zu arbeiten. Inzwischen haben wir dafür erste Antworten und einen theoretischen Rahmen entwickelt, mit dem sich KI nachweislich zum Nutzen der Menschen bauen lässt.

Was, glauben Sie, ist der erfolgreichste Weg, eine AGI zu entwickeln? Wird das allein über Sprachmodelle klappen?
Das erste Sprachmodell wurde 1913 konstruiert – also vor mehr als hundert Jahren. Niemand hat Sprachmodelle als Grundlage für Intelligenz betrachtet. Man sah sie als Grundlage dafür, das nächste Wort vorherzusagen. Also: Wenn Sie auf dem Handy tippen, müssen Sie nicht das ganze Wort schreiben, es errät, was Sie als Nächstes tippen.

Also werden die Modelle überschätzt?
Heute wirken große Sprachmodelle für viele wie Intelligenz, aber sie stoßen an Grenzen. Deshalb setzt die nächste Generation, die sogenannten Large Reasoning Models, auf eine Kombination: Sprachmodelle als Baustein, plus klassische Algorithmen, die systematisch viele mögliche Denkschritte durchspielen. Das ist näher an AGI als reine Sprachmodelle. Aber vermutlich reicht auch das noch nicht – es braucht neue, andere Ideen.

Und was könnten diese neuen Ideen sein?
Das verrate ich Ihnen jetzt nicht. Ob diese neuen Ideen umgesetzt werden, ist sehr schwer vorherzusagen. Mein Bauchgefühl liegt ungefähr bei einer 25-prozentigen Chance, dass diese Durchbrüche rechtzeitig passieren, bevor die Blase platzt. Denn das ist das Nächste, was passiert: Wenn diese Durchbrüche nicht eintreten, werden die Unternehmen nicht liefern können, was sie ihren Investoren versprechen.

Also, 25 Prozent sind nicht schlecht, aber wenn wir es umdrehen: Glauben Sie, dass die Blase platzen wird?
Mit 75 Prozent Wahrscheinlichkeit passiert genau das: Die Blase platzt, weil Umsätze und Gewinne aus dem heutigen Fähigkeitsniveau nicht reichen, um die gigantischen Investitionen zu rechtfertigen. Und im 25-Prozent-Szenario, in dem die Durchbrüche tatsächlich kommen, ist es in gewisser Weise noch schlimmer: Dann hätten wir AGI-Systeme mit mehr Macht als Menschen – ohne eine Lösung, wie wir sie kontrollieren. Das räumen die KI-Firmen selbst ein. Sie sagen wörtlich: Wenn wir dieses Ziel erreichen, für das wir Billionen ausgeben, gibt es – je nachdem, wen man fragt – eine zehn-, 20-, 25-, in manchen Fällen sogar 50-prozentige Chance, dass die Menschheit ausstirbt.

Die Wahrscheinlichkeit, dass KI die Menschheit auslöscht, liegt bei bis zu 50 Prozent?
Ja. Emmett Shear, der kurzzeitig CEO von OpenAI war, schätzt das Risiko auf irgendwo zwischen fünf und 50 Prozent.

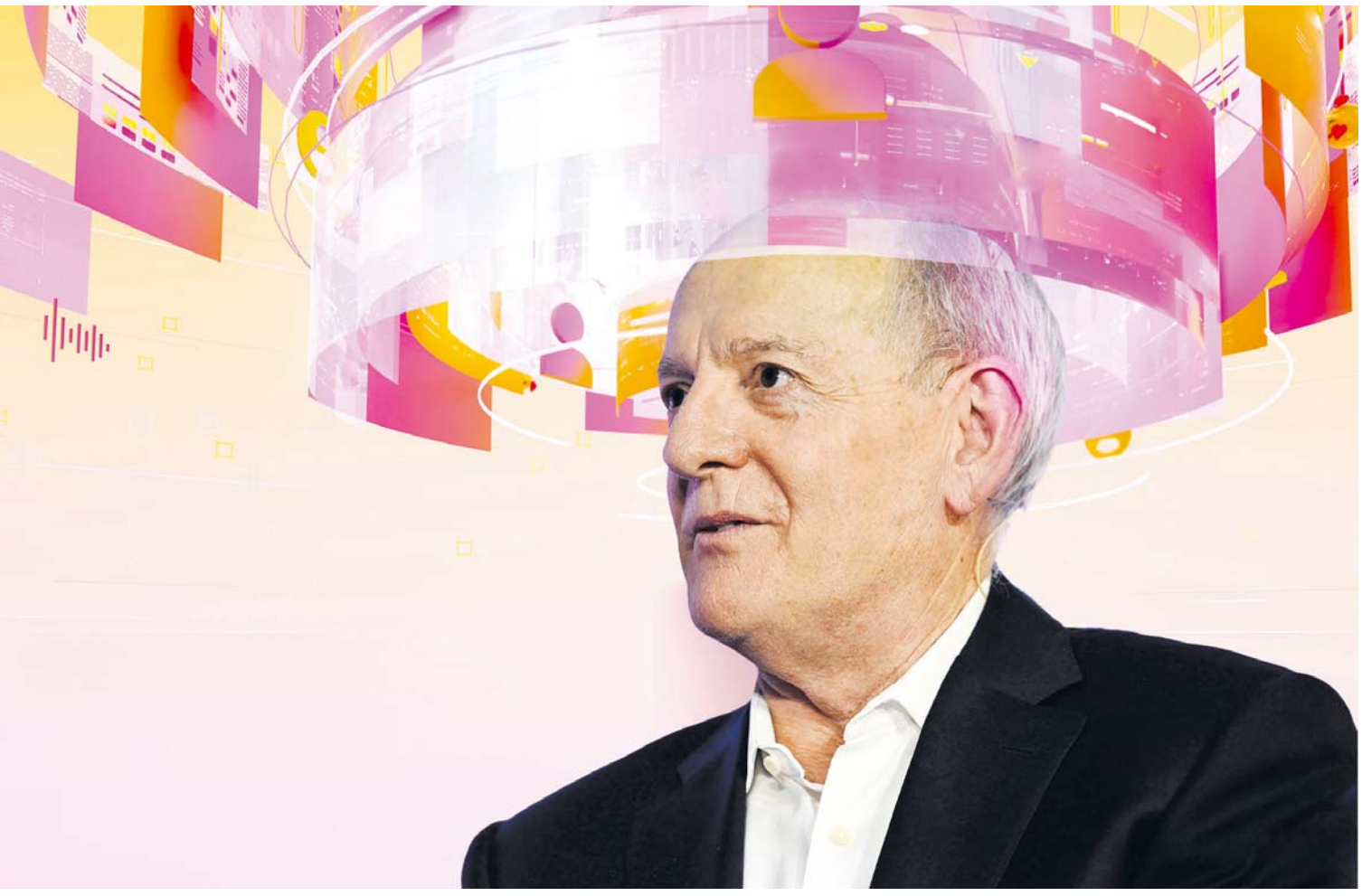
Ist das der Grund, warum Sie und andere noch im Dezember diesen Brief geschrieben haben, der fordert, die Entwicklung von AGI zu stoppen?
Ja, denn wenn wir Systeme schaffen, die mächtiger sind als Menschen, ohne sie kontrollieren und ihr sicheres Verhalten garantieren zu können, verlieren wir letztlich jedes Mitspracherecht über unsere eigene Existenz – so wie Schimpansen gegenüber Menschen. Der Brief war nicht als Verbotsforderung gemeint. Er sagt nur: AGI darf es erst geben, wenn klar ist, wie man sie sicher macht auf eine Weise, die sowohl die wissenschaftliche Community überzeugt als auch von der Öffentlichkeit breit akzeptiert wird.

Welche Regeln würden Sie verpflichtend für KI-Firmen auf dem Weg zur Entwicklung von Superintelligenz oder AGI aufstellen?
Sie müssen beweisen, dass es sicher ist und den Menschen nützt. Nur ist es schwer, genau zu definieren, was „sicher“ bei KI eigentlich heißt. Bei einem Flugzeug ist das ziemlich einfach: Es

Vita

Werdegang Stuart Russell ist ein britischer Informatiker und zählt zu den weltweit einflussreichsten KI-Forschern. Er studierte in Oxford und promovierte an der Stanford University. Seit den 1980er-Jahren lehrt und forscht Russell an der University of California in Berkeley. Bekannt wurde er vor allem als Mitautor des Standardlehrbuchs „Artificial Intelligence: A Modern Approach", das international in der KI-Ausbildung genutzt wird.

Forschung Russells Forschung dreht sich um die Frage, wie KI so entwickelt werden kann, dass sie mit menschlichen Zielen vereinbar bleibt. Er warnt vor Systemen, die starr vorgegebene Zielvorgaben optimieren und dabei unerwünschte Nebenfolgen in Kauf nehmen. Stattdessen soll KI Unsicherheit über menschliche Ziele berücksichtigen, aus menschlichem Verhalten lernen und sich korrigieren lassen.



picture alliance for DLD/ Hubert Burda Media, Getty Images [M], Getty Images/Science Photo Libra

> AGI darf es erst geben, wenn klar ist, wie man sie sicher macht auf eine Weise, die sowohl die wissenschaftliche Community überzeugt als auch von der Öffentlichkeit breit akzeptiert wird.

soll den Boden nicht berühren, außer man will es, richtig? Bei KI-Systemen ist das schwieriger: Weil sie so allgemein sind, können sie auf unzählige Arten eingesetzt werden, und entsprechend vielfältig ist auch das Schadenpotenzial. Das alles lässt sich kaum festschreiben.

Was schlagen Sie stattdessen vor?
Wir schlagen „behavioral red lines" vor – rote Linien für das Verhalten. Also konkrete, klar inakzeptable Dinge, die KI niemals tun darf. Zum Beispiel: sich ohne menschliche Erlaubnis selbst replizieren, sich als Mensch ausgeben oder Terroristen beim Bau biologischer Waffen beraten. Das sind offensichtliche No-Gos nach gesundem Menschenverstand.

In Europa verlieren wir uns oft in Detailregeln. Wenn wir auf Regeln und Debatten rund um AGI schauen: Übersehen wir die entscheidenden Punkte?
Ich glaube, das liegt daran, dass wir zu der Zeit, als der europäische „AI Act" entwickelt wurde, kein klares Gefühl dafür hatten, was eine effektive Regulierung ist. Und „effektiv" ist das entscheidende Wort: Man kann unendlich viele Regeln haben – wirksam sind sie nur, wenn ihre Einhaltung das Risiko zwangsläufig auf ein akzeptables Niveau senkt. Wenn es um das Aussterben der Menschheit geht, wäre dieses akzeptable Niveau vermutlich „null".

Aber wir entwickeln auch Waffen, und dort akzeptieren wir kein Risiko von null, dass Menschen sterben oder ganze Nationen zerstört werden.
Da wird das Aussterberisiko oft überzeichnet. Beim Klima etwa ist es wahrscheinlich katastrophal, aber nicht zwangsläufig existenzvernichtend. Aber ja, bei Waffen akzeptieren wir kein Risiko von null, dass Menschen sterben oder ganze Nationen zerstört werden. Aber Sie haben einen Punkt: Bei Atomwaffen investieren wir enorme Anstrengungen, damit sie nicht in falsche Hände geraten. Bei KI tun wir im Grunde das Gegenteil: Wir sorgen dafür, dass sie möglichst schnell und breit verfügbar wird.

Europa ist bei KI-Modellen stark von den USA und China abhängig. Müssen wir nicht zumindest eigene leistungsfähige Modelle entwickeln, auch aus Gründen politischer Unabhängigkeit?
Nun, das hängt davon ab, ob Sie glauben, dass wir eine Technologie, die uns auslöschen könnte, auch selbst entwickeln müssen. Viele der versprochenen Vorteile erfordern jedenfalls keine AGI. Das vielleicht wichtigste KI-Ergebnis der letzten Jahre waren die Protein-Folding-Modelle, die 2024 den Nobelpreis gewonnen haben – die haben mit AGI wenig zu tun und auch nicht viel mit großen Sprachmodellen. Es sind spezialisierte Verfahren für ein konkretes Problem. Fortschritte kommen aus schmalen, zielgerichteten Modellen, die Wissen aus Biologie, Physik oder anderen Disziplinen integrieren.

Und da hat Europa bessere Chancen?
Ich glaube absolut, dass Europa das auch kann. Europa hat viele großartige Forschungszentren. Am Ende ist es genau das, was Menschen wollen: keine KI, die Schaden anrichtet, sondern Systeme mit echten Sicherheitsgarantien – dass sie tun, was sie sollen, und nicht etwas anderes.

Sie haben Vorteile genannt. KI könnte aber auch den Arbeitskräftemangel abfedern, viele erwarten schon 2026 spürbare Produktivitätsgewinne.
Ich denke, das ist vor allem ein Prozess des Ausprobierens: Die Technologie ist besser geworden, bei einigen Modellen ist die Halluzinationsrate gesunken. Entscheidend ist: Welche Fehlerquote kann man sich in einem konkreten Teil des Workflows leisten?

In vielen Bereichen wird KI ja bereits in Unternehmen eingesetzt.
Manches, was großartig klang, hat nicht funktioniert. Klarna wollte den Kundenservice komplett durch KI ersetzen, hat Menschen entlassen – und sie sechs Monate später wieder eingestellt, weil es nicht lief. Das sehe ich öfter. Gleichzeitig gibt es Bereiche, in denen KI als Assistenz nützt. Softwareentwicklung ist der Bereich, in dem man die größten und schnellsten Produktivitätsgewinne erwartet hat. Es gab Experimente, die 20 oder 30 Prozent Produktivitätssteigerung zeigten. Aber viele davon kamen von Unternehmen mit Eigeninteresse. Unabhängige Studien zeigten teils das Gegenteil: Entwickler fühlten sich produktiver, waren es aber nicht, weil die Codequalität nicht reichte und sie am Ende doch verstehen und korrigieren mussten, was das System geschrieben hatte.

Also rechnen Sie nicht mit signifikanten Produktivitätsgewinnen?
Ich denke, es wird Gewinne geben. Aber ob das wirklich Jobs rettet oder ersetzt, hängt stark von der Branche ab. In gesättigten Märkten führt mehr Produktivität nicht zu höherer Nachfrage – das hat man in den 1920er-Jahren bei Autos und der Massenproduktion gesehen. Dann braucht man am Ende einfach weniger Menschen. Und wenn ein Unternehmen plötzlich halb so teuer produzieren kann, müssen alle anderen nachziehen, ob sie wollen oder nicht.

Wie sollten Unternehmen und Führungskräfte darauf reagieren?
Wenn KI besser wird, wenn die Fehlerraten Richtung null sinken, werden immer mehr Funktionen dafür geeignet sein, KI statt Menschen einzusetzen. Viele Jobs in der modernen Wirtschaft sind „Sprache rein, Sprache raus". So sind Sie. So bin ich. Und genau diese Arbeit wird verwundbar, je besser die Systeme werden. Das gilt nicht nur für Büroarbeit, KI wird auch physisch: Roboter, selbstfahrende Autos. In San Francisco sieht man schon heute ganze Straßenzüge voller Autos ohne Fahrer. Für Taxifahrer ist die Richtung klar – das betrifft weltweit mehr als 25 Millionen Menschen.

Heißt das, Unternehmen müssen um ihr Existenzrecht kämpfen, wenn KI ihre Jobs übernimmt?
Technisch ist es machbar, dass Unternehmen komplett automatisiert laufen: ohne menschliche Führung, vielleicht nur mit Aktionären. Dann wird die entscheidende Frage sein: Wie sieht die Industrie aus, wenn ein Großteil der Wertschöpfung von KI kommt – und wem gehört sie? Waymo ist ein Beispiel: Heute betreiben sie Taxiflotten. Langfristig könnten sie ihre Software an Hersteller verkaufen oder irgendwann selbst die Flotten besitzen. Beides ist denkbar. Und man kann sich das auch in anderen Branchen vorstellen, etwa bei Rechtsdienstleistungen: Statt KI an Kanzleien zu verkaufen, bietet man die Dienstleistung direkt an – und die Kanzleien verschwinden.

Aber braucht man nicht Menschen für Innovation und Strategie?
Nicht unbedingt. Der Punkt bei AGI ist: Sie kann das auch. Und irgendwann werden Aktionäre sagen: Es ist Zeit, den CEO loszuwerden. Unsere Konkurrenz hat ihn schon ersetzt. Weil die KI rationaler ist und mehr weiß. Das KI-System hat jeden einzelnen Newsfeed der Welt gelesen, hat in den ersten drei Sekunden tausend E-Mails beantwortet und managt Hunderte Aktivitäten gleichzeitig. Dann ist ein CEO nach dem anderen dran. Ökonomische Kräfte würden das vorantreiben. Wir könnten aber auch sagen: Wir machen eine Regel, all diese Berufe dürfen nur von Menschen ausgeübt werden. Und Sie können wetten: Welche Profession wird als erste gesichert?

Die Politik?
Die Politik, exakt.

Die Fragen stellte Luisa Bomke.