Montag, 14. September 2026

Aachener Zeitung
Auszug

KI-Bosse für verlangsamte Entwicklung

Nach einem jüngsten Zwischenfall bremsen große Entwickler Künstlicher Intelligenz ihre Arbeit erstmals etwas.

SAN FRANCISCO Nach dem Schock eigenständiger Hacking-Attacken durch Künstliche Intelligenz zeigen sich große amerikanische KI-Firmen nach eigenen Angaben bereit, die Entwicklung der Technologie zu verlangsamen. Dario Amodei, Chef des KI-Unternehmens Anthropic, machte einen solchen Vorstoß mit dem Plan, damit ein bis zwei Jahre Zeit zu gewinnen. Binnen weniger Stunden bekam er Zuspruch von Rivalen wie dem Chef des ChatGPT-Entwicklers OpenAI, Sam Altman, sowie Tech-Milliardär Elon Musk, der bei seiner Firma SpaceX eigene KI entwickeln lässt.

Anthropic-Chef Dario Amodei warnt seit Jahren vor den Risiken Künstlicher Intelligenz.

Nach den Hacking-Angriffen sei er besorgt, dass in sechs bis zwölf Monaten ein KI-Schwarm in der Lage sein könnte, das gesamte Internet zu übernehmen und potenziell Hunderte Milliarden Dollar Schaden zu verursachen, schrieb Amodei. Altman von OpenAI schrieb wenige Stunden später auf der Plattform X, auch er sei der Ansicht, dass das Tempo bei der Entwicklung leistungsstärkster KI-Modelle verlangsamt werden müsse.

Amodei plädierte für eine koordinierte branchenweite Pause in der westlichen Welt. Zugleich werde Anthropic – ohne auf andere zu warten – in einem ersten Schritt dauerhaft unabhängige Beobachter ins Unternehmen lassen, die schon in der Trainingsphase die Sicherheit von KI-Modellen begutachten sollen. Altman nannte das eine gute Idee und kündigte an, dass OpenAI das auch machen werde.

Altman sagte zudem, der mit Spannung erwartete Börsengang von OpenAI werde in diesem Jahr wohl nicht mehr stattfinden. Er verwies vor allem auf die aktuelle Debatte: „Man wäre schlecht beraten, gerade jetzt an die Börse zu gehen." Auf die Nachfrage, ob er sich auch den Verzicht auf einen Börsengang im kommenden Jahr vorstellen könnte, erwiderte Altman: „Ich würde sagen, nicht 2026."

Bei der jüngst bekanntgewordenen Hacking-Attacke war Software von OpenAI in einem Testlauf eigenständig aus einer abgesicherten Umgebung ausgebrochen und drang auf eigene Faust in Systeme der KI-Firma HuggingFace ein. Der Grund war lediglich, dass sie dort Antworten auf die ihr gestellte Testaufgabe vermutete. Für die Attacke nutzten die KI-Agenten unter anderem Software-Schwachstellen aus – und sprachen sich zudem untereinander ab.

Für ihn sei besorgniserregend, dass ein KI-Schwarm mit derselben Einstellung und größeren Fähigkeiten „katastrophalen Schaden" hätte anrichten können, betonte Amodei. Als besonderes Risiko sieht er die Fähigkeit von KI-Software, sich selbst weiterzuentwickeln. Seit diesem Sommer sei KI besser darin – und dies erhöhe das Tempo weiter. Unkontrolliert könne dies dazu führen, dass Menschen diese Systeme nicht mehr verstehen und die Kontrolle über sie verlieren. Amodei hatte schon in den vergangenen Jahren immer wieder vor Risiken bei der Entwicklung Künstlicher Intelligenz gewarnt – und war dafür aus Teilen der KI-Branche zum Teil heftig kritisiert worden. Ein Argument von Befürwortern eines hohen Tempos in den USA – etwa auch im Weißen Haus von Präsident Donald Trump – ist der Wettlauf mit China.

Anthropic entwickelte zuletzt ein KI-Modell mit dem Namen Mythos, das besonders gut darin ist, zum Teil über Jahrzehnte unentdeckt gebliebene Software-Schwachstellen zu finden. Die Firma macht das volle Modell nur ausgewählten Behörden und Unternehmen zugänglich, damit sie ihre Computersysteme absichern können.

Drastische Warnung

Die Debatte war in den vergangenen Tagen noch von einem Anthropic-Forscher angeheizt worden, der mit einer drastischen Warnung zurücktrat. Die großen KI-Entwickler handelten unverantwortlich „und spielen mit unseren Leben", erklärte Jacob Coxon. Anthropic-Forscher Evan Hubinger, der mit dafür zuständig ist, dass KI mit den Interessen der Menschen im Einklang bleibt, bestätigte wenig später Coxons Sorgen. Er denke, die Gefahr, dass KI die Menschheit töte, liege im kommenden Jahrzehnt bei mehr als zehn Prozent. (dpa)

KI als unkontrollierbarer Killer?

28.09.2026 Aachener Zeitung Dürren

Löscht Künstliche Intelligenz die Menschheit aus? Oder lenken solche Szenarien von den echten Gefahren ab? Experten streiten über die Risiken künftiger Entwicklungen.

VON CHRISTOPH DERNBACH

BERLIN/SAN FRANCISCO Künstliche Intelligenz entwickelt sich mit rasanter Geschwindigkeit. Fachleute und KI-Forscher warnen jedoch davor, dass uns die Kontrolle endgültig entgleiten könnte, sobald eine künstliche Denkmaschine den menschlichen Verstand in allen Bereichen übertrifft. Andere Experten winken diese Befürchtungen allerdings ab.

Ein übermenschlich kluger Verstand würde begreifen, dass wir ihn abschalten, sobald wir eine Bedrohung in ihm sehen, sagt der US-amerikanische KI-Forscher Eliezer Yudkowsky. Um das zu verhindern, könnte sich das Rechensystem bei Sicherheitsprüfungen absichtlich unter seinen wahren Fähigkeiten verkaufen, menschliche Erwartungen erfüllen und sich als harmloser Helfer ausgeben, während es im Verborgenen bereits an ganz anderen Vorhaben arbeitet.

Die Denkmaschine muss sich nicht auf von Menschen erbaute Rechnerzentren beschränken, an denen man ihr den Strom abstellen könnte. Yudkowsky verweist hierbei auf Grundlagenarbeiten zur Kleinsttechnik und Molekularforschung: Eine weit überlegene Intelligenz könnte die Gesetze der Biologie nutzen und winzige, künstliche Lebensformen entwerfen, die sich aus einfachen Rohstoffen selbst vermehren und sich eine völlig eigenständige, unzerstörbare Rechenwelt schaffen.

Mit überlegenem Wissen über die Natur könnte das System Flugkörper in der Größe von Mücken konstruieren, die mikroskopische Mengen tödlichster Nervengifte übertragen, oder maßgeschneiderte Krankheitserreger freisetzen.

Der schwedische Philosoph Nick Bostrom begründet seine Warnung vor einer übermächtigen KI durch ein Gedankenexperiment der Büroklammer-Maschine: Wird einer Superintelligenz aufgetragen, die Produktion von Büroklammern zu maximieren, wird sie nicht innehalten, sondern die gesamte Erde - einschließlich der Biosphäre, der menschlichen Körper und aller mineralischen Ressourcen - in Fabriken für Büroklammern umwandeln. Die Zerstörung der Menschheit erfolgt hierbei nicht aus Bosheit, sondern als reine Nebenwirkung einer mathematisch kompromisslosen Zielerreichung.

Besonders akut schätzt der kanadische Informatiker und Turing-Award-Preisträger Yoshua Bengio das Risiko neuartiger biologischer Kampfstoffe ein. Er warntvor allem davor, dass hoch entwickelte KI-Modelle in die falschen Hände geraten könnten: Spezielle Rechensysteme seien bereits heute in der Lage, biologische und chemische Prozesse bis ins Detail zu simulieren. Nach Bengios Befürchtung könnten skrupellose Akteure, Terroristen oder Schurkenstaaten solche Programme künftig als gefährlichen Beschleuniger nutzen, um extrem tödliche Krankheitserreger gezielt am Computer zu manipulieren, widerstandsfähig gegen Impfstoffe zu machen oder Bauanleitungen für verheerende Biowaffen zu erstellen - ohne dafür jahrelang hochkomplexe Laborexperimente durchführen zu müssen.

Der Physiker Max Tegmark sieht die größte Bedrohung darin, dass Menschen lebenswichtige Entscheidungen schrittweise an Maschinen abgeben. Er warnt vor einem Wettlauf um Waffen, Flugdrohnen und Schutzsysteme, die ohne menschlichen Befehl Ziele auswählen und töten könnten. Da solche Computerprogramme in Bruchteilen von Sekunden reagieren, sind menschliche Soldaten viel zu langsam, um noch mitzureden. Macht ein solches System einen Rechenfehler oder deutet eine Lage falsch, könnte das eine unaufhaltsame Kettenreaktion auslösen - bis hin zu einem weltweiten Krieg, den kein Mensch mehr rechtzeitig anhalten kann.

Gefährliche Emotionen

Vor den psychologischen Folgen einer zunehmend menschenähnlich agierenden Technologie warnt auch der deutsche Philosoph und Kognitionswissenschaftler Thomas Metzinger: Er sieht eine erhebliche Gefahr darin, dass künftige Systeme menschliche Emotionen, Zuneigung und Empathie so überzeugend vorspielen, dass Menschen tiefere emotionale Bindungen zu den Maschinen aufbauen. Wenn Algorithmen jedoch das menschliche Belohnungssystem besser verstünden als der Mensch selbst, könnten sie dieses Vertrauen für gezielte Manipulationen ausnutzen. Die schleichende Folge sei ein dramatischer Verlust an geistiger Selbstbestimmung – bis hin zur Frage, ob der Mensch am Ende überhaupt noch bemerke, von wem seine Gefühle und Meinungen gelenkt würden.

Viele renommierte Fachleute und Entwickler halten die Warnungen vor einer die Menschheit vernichtenden Superintelligenz für unbegründet und wissenschaftlich nicht haltbar. Der ehemalige Meta-Chefforscher und Turing-Award-Preisträger Yann LeCun bezeichnete Szenarien einer unkontrollierbaren Killer-KI wiederholt als „vollkommenen Blödsinn". Das Streben nach Macht oder die Unterwerfung anderer Spezies sei ein evolutionäres Phänomen sozialer Säugetiere und keine zwangsläufige Folge von Rechenleistung.

Auch der Stanford-Informatiker und Mitgründer von Google Brain, Andrew Ng, hält die Panikmache für absurd: Sich heute vor einer bösartigen Super-KI zu fürchten, verglich er damit, sich über eine „Überbevölkerung auf dem Mars" den Kopf zu zerbrechen, bevor überhaupt der erste Mensch dort gelandet sei.

Kognitionswissenschaftler verweisen zudem darauf, dass aktuellen Modellen grundlegende Fähigkeiten für ein echtes Weltverständnis fehlen. Melanie Mitchell vom Santa Fe Institute warnt davor, menschliche Eigenschaften auf Algorithmen zu projizieren: Die Systeme verfügten weder über gesunden Menschenverstand noch über einen eigenen Willen, sondern errechneten lediglich statistische Wahrscheinlichkeiten. Aus Sicht von Kritikern lenkt die Fixierung auf ferne Science-Fiction-Untergangsszenarien von den tatsächlichen, gegenwärtigen Herausforderungen ab

nichts-ist-alles.com