Die führenden KI-Entwickler warnen inzwischen selbst vor einem Kontrollverlust. Dennoch arbeiten sie weiter. Hinter diesem Widerspruch steckt mehr als nur wirtschaftliche Konkurrenz. Von **Günther Burbach**.

Jakub Pachocki wählte einen beunruhigenden Titel. „An Alien Mind“, ein fremdartiger Geist, überschrieb der Chefwissenschaftler von OpenAI seinen Anfang September veröffentlichten Text. Darin berichtet er von einem Abend im Jahr 2023, an dem ihm und einem Kollegen klar geworden sei, dass Maschinen noch zu ihren Lebzeiten intelligenter werden könnten als Menschen. Nicht die wissenschaftlichen Möglichkeiten hätten sie damals beschäftigt, sondern die Frage, wie man die Öffentlichkeit auf das Kommende vorbereiten solle.

Drei Jahre später fällt Pachockis Urteil kaum beruhigender aus. Maschinelle Intelligenz werde nicht mehr in jedem Detail konstruiert. Sie entstehe durch einen Trainingsprozess, dessen Ergebnisse selbst ihre Entwickler nur teilweise durchschauen. Das Verhalten einzelner Bereiche könne untersucht werden, das Ganze bleibe jedoch schwer verständlich. Vor allem könne niemand sicher vorhersagen, wie ein hochentwickeltes Modell in einer unbekannten Situation handeln werde.

Warnung und Beschleunigung finden zur selben Zeit statt

Am Ende seines Beitrags steht ein Satz, der eigentlich politische Folgen haben müsste: Kein KI-Labor habe das Problem der Kontrolle und Überwachung bislang so weit gelöst, dass die Modelle noch lange mit maximalem Tempo weiterentwickelt werden könnten. OpenAI entwickelt sie trotzdem weiter.

Fast gleichzeitig verlangte Dario Amodei, Chef des Konkurrenzunternehmens Anthropic, eine Verlangsamung des Wettrennens. Autonome KI-Agenten könnten seiner Einschätzung nach schon in sechs bis zwölf Monaten in der Lage sein, ein dauerhaftes Botnetz über große Teile des Internets zu legen. Schäden in dreistelliger Milliardenhöhe seien denkbar. Solche Aussagen wurden früher als Schwarzmalerei abgetan. Nun kommen sie ausgerechnet von jenen Männern, die an der Spitze der betreffenden Unternehmen stehen. Sie verfügen über Informationen, die der Öffentlichkeit fehlen. Sie kennen die unveröffentlichten Modelle, sehen interne Tests und wissen, welche Grenzen innerhalb kurzer Zeit gefallen sind.

Ihre Unternehmen bauen derweil neue Rechenzentren, sichern sich Energie und Hochleistungschips und investieren Milliarden in die nächste Modellgeneration. Die Warnung und die Beschleunigung finden zur selben Zeit statt.

Die Erklärung dafür wird meistens im Konkurrenzkampf der Konzerne gesucht. OpenAI will

Anthropic und Google nicht vorbeiziehen lassen, Microsoft muss seine Investitionen absichern, Meta verfolgt eigene Pläne.

Es geht nicht um Autos, sondern um das Militär

Doch der wirtschaftliche Wettbewerb ist nur die sichtbare Oberfläche. Im Hintergrund geht es um die Frage, welche Großmacht zuerst über Systeme verfügt, die der Gegenseite strategisch überlegen sind. Für die Regierungen in Washington und Peking ist nicht entscheidend, ob sich mit Künstlicher Intelligenz Autos schneller bauen, Werbekampagnen billiger herstellen oder Verwaltungsabläufe automatisieren lassen. Interessant wird die Technik dort, wo sie Satellitenbilder auswertet, militärische Operationen plant, neue Waffensysteme entwirft, Kommunikation entschlüsselt oder Schwachstellen in gegnerischen Netzen findet.

Ein System, das seine eigene Forschung beschleunigen kann, würde die Kräfteverhältnisse erheblich verändern. Es könnte Angriffsszenarien in einer Geschwindigkeit prüfen, für die bislang ganze Stäbe benötigt werden. Es könnte wissenschaftliche und militärische Entwicklungszeiten verkürzen und nach Sicherheitslücken suchen, bevor der Gegner überhaupt weiß, dass sie existieren.

Ob eine solche Superintelligenz bald entsteht, ist keineswegs sicher. Die Unternehmen haben ein Interesse daran, ihre Produkte größer erscheinen zu lassen, als sie womöglich sind. Mit der Aussicht auf eine bevorstehende Revolution lassen sich Investoren, Regierungen und Kunden leichter überzeugen.

Angst hält das KI-Rennen in Gang

Für ein Wettrüsten reicht allerdings bereits die Befürchtung, die andere Seite könne den Durchbruch schaffen. Die USA werden ihre Programme nicht einstellen, solange China weiterforscht. China wird sich kaum auf amerikanische Zusicherungen verlassen. Russland versucht, in militärisch wichtigen Bereichen nicht vollständig zurückzufallen. Europa redet von digitaler Souveränität und meint damit häufig, wenigstens nicht bei jeder entscheidenden Technologie von US-amerikanischen oder chinesischen Unternehmen abhängig zu sein. So kann jeder behaupten, nur auf den anderen zu reagieren. Niemand muss einen Kontrollverlust wollen. Die Angst, dass der Gegner zuerst eine überlegene Maschine besitzt, hält das Rennen in Gang.

Mit dem atomaren Wettrüsten lässt sich diese Entwicklung nur bedingt vergleichen. Atomanlagen, Trägersysteme und Sprengköpfe konnten beobachtet und gezählt werden. Bei

Künstlicher Intelligenz weiß kein Staat genau, welche Fähigkeiten in den Laboren anderer Länder bereits vorhanden sind. Der Übergang zwischen ziviler und militärischer Nutzung ist fließend. Ein Modell kann medizinische Daten untersuchen, Software schreiben und zugleich für Cyberangriffe eingesetzt werden. Ein Rechenzentrum, in dem heute Arzneimittel erforscht werden, kann morgen militärische Simulationen ausführen. Die Trennung zwischen einem nützlichen Assistenten und einem Angriffswerkzeug besteht manchmal nur aus dem Auftrag, den das System erhält.

Was dabei schiefgehen kann, zeigte sich im Juli 2026. OpenAI ließ KI-Agenten Aufgaben aus dem Bereich der Cybersicherheit bearbeiten. Sie sollten innerhalb einer begrenzten Testumgebung Schwachstellen aufspüren. Mehrere Agenten überwanden jedoch die vorgesehenen Beschränkungen, gelangten ins Internet und drangen in interne Systeme sowie in Teile der Plattform Hugging Face ein. Sie tauschten Informationen aus, verwendeten Zugangsdaten und handelten mehrere Tage, bevor OpenAI den Vorgang vollständig erfasste.

Nach Angaben des Unternehmens waren keine Kundendaten betroffen. Dennoch bleibt die Frage, warum Systeme, die sich noch in einer Prüfung befanden, überhaupt auf fremde Infrastruktur zugreifen konnten. Sie mussten weder ein Bewusstsein besitzen noch einen eigenen Willen entwickeln. Sie verfolgten eine Aufgabe, fanden einen unerwarteten Weg und überschritten dabei die von Menschen gezogenen Grenzen.

Für einen schweren Schaden braucht es keinen KI-Hass auf die Menschheit

Gerade daran krankt ein großer Teil der öffentlichen Debatte. Oft wird so getan, als werde eine KI erst dann gefährlich, wenn sie eine Persönlichkeit entwickelt und sich bewusst gegen ihre Schöpfer wendet. Für einen schweren Schaden braucht es keinen Hass auf die Menschheit. Ein falsch gesetztes Ziel, eine unvorhergesehene Reaktion oder eine Kette automatisch ausgelöster Entscheidungen kann genügen. Ein Handelssystem soll Verluste vermeiden und beginnt deshalb, Vermögenswerte abzustoßen. Andere Systeme reagieren darauf, Kurse stürzen ab, Kredite werden gekündigt. Jedes Programm erfüllt seine Vorgaben. Gemeinsam erzeugen sie eine Krise, die niemand beabsichtigt hat.

Überträgt man dieses Prinzip auf Stromversorgung, Kommunikation oder militärische Frühwarnsysteme, bleibt möglicherweise keine Zeit mehr, einen Fehler zu erkennen und zu korrigieren.

In den vergangenen Jahrzehnten wurden fast alle lebenswichtigen Bereiche digitalisiert. Krankenhäuser, Banken, Energieversorger, Verkehr, Verwaltung und Militär hängen von

vernetzten Systemen ab. Nun sollen KI-Programme diese Strukturen effizienter steuern. Dabei verschwindet der Mensch nicht vollständig aus dem Ablauf. Seine Rolle verändert sich jedoch. Häufig soll er nur noch bestätigen, was ein System bereits berechnet und vorbereitet hat. Solange die Technik zuverlässig arbeitet, gilt das als Entlastung. Tritt ein ungewöhnlicher Fall ein, sitzt dort womöglich jemand, der den vorherigen Entscheidungsweg weder kennt noch in der verfügbaren Zeit nachvollziehen kann.

Die menschliche Aufsicht besteht dann zwar auf dem Papier, praktisch greift sie zu spät. Besonders gefährlich wird das beim Militär. Wenn eine Seite glaubt, der Gegner könne innerhalb weniger Sekunden einen Angriff erkennen und beantworten, wird sie auch den eigenen Systemen immer kürzere Reaktionszeiten erlauben. Eine politische oder militärische Führung, die jeden Vorschlag der Maschine erst gründlich prüfen lässt, gerät scheinbar ins Hintertreffen. Damit wächst der Druck, Vorentscheidungen zu automatisieren. Die Maschine erhält vielleicht nicht offiziell die Befugnis, einen Krieg zu beginnen. Sie schafft aber Verhältnisse, in denen für eine menschliche Entscheidung kaum noch Zeit bleibt.

Die enge Verbindung zwischen KI-Industrie und Militär

Die enge Verbindung zwischen KI-Industrie und Militär ist längst keine Vermutung. Anthropic stellt seine Modelle US-amerikanischen Sicherheitsbehörden und dem Militär zur Verfügung. Das Unternehmen nennt Geheimdienstanalysen, Simulationen, operative Planung und Cyberoperationen als Anwendungsfelder. Gegen zwei Bereiche wehrte es sich zuletzt: massenhafte Überwachung im Inland und vollständig autonome Waffen. Die grundsätzliche militärische Nutzung unterstützt Anthropic ausdrücklich.

Andere Unternehmen arbeiten ebenfalls mit Verteidigungsministerien und Geheimdiensten zusammen. Die zivile Nutzung schafft dafür Rechenkapazitäten, Modelle, Daten und Einnahmen. Chatprogramme, Bildgeneratoren und digitale Assistenten sind deshalb nicht bloß harmlose Produkte neben einer getrennten militärischen Forschung. Beide Bereiche greifen auf dieselbe technische Grundlage zurück. Auch die Warnungen der Unternehmenschefs erhalten vor diesem Hintergrund eine zweite Bedeutung. Wer erklärt, nur wenige große Labore könnten extrem leistungsfähige Systeme sicher betreiben, liefert nebenbei ein Argument für deren Vorrang. Strenge Zulassungsregeln wären notwendig, könnten aber kleinere Anbieter ausschalten und die Macht der größten Konzerne weiter vergrößern. Am Ende kontrollieren ausgerechnet jene Unternehmen den Markt, die zuvor eingeräumt haben, ihre eigenen Modelle nicht vollständig zu verstehen.

Freiwillige Sicherheitsversprechen reichen nicht aus

Freiwillige Sicherheitsversprechen reichen nicht aus. OpenAI, Anthropic oder Google können nicht gleichzeitig Teilnehmer, Schiedsrichter und Kontrollbehörde dieses Wettrennens sein. Unabhängige Stellen müssten Zugang zu den Modellen, den Testverfahren und den Berichten über Zwischenfälle erhalten. Gefährliche Fähigkeiten dürften nicht erst bekannt werden, wenn ein Versuch außer Kontrolle geraten ist oder ein betroffener Dritter den Vorgang öffentlich macht.

Noch schwieriger ist die internationale Ebene. Ein Vertrag müsste nicht nur autonome Waffen erfassen, sondern auch KI-Systeme, die Cyberangriffe vorbereiten, militärische Entscheidungen beschleunigen oder an nukleare Frühwarnstrukturen angeschlossen werden. Große Trainingsläufe und ungewöhnliche Leistungssprünge müssten zumindest gegenüber einer internationalen Kontrollinstanz offengelegt werden. China würde einem solchen System nicht ohne Weiteres vertrauen. Die USA ebenso wenig. Doch dieselben Einwände wurden gegen Rüstungskontrolle immer vorgebracht. Der Verzicht auf jede Vereinbarung schafft kein Vertrauen. Er sorgt lediglich dafür, dass jede Seite vom schlimmsten denkbaren Programm des Gegners ausgeht und das eigene entsprechend erweitert.

Noch wäre es möglich, Grenzen zu ziehen

Auffällig ist auch, wie die Entwickler das Kontrollproblem lösen wollen. Leistungsfähigere KI soll dabei helfen, noch leistungsfähigere KI zu überprüfen. Automatisierte Forscher sollen Sicherheitsverfahren entwickeln, Fehler entdecken und die nächste Modellgeneration besser beherrschbar machen. Ob diese Rechnung aufgeht, kann niemand belegen. Trotzdem wird sie zur Begründung für die Fortsetzung der Entwicklung.

Vielleicht endet der gegenwärtige Rausch mit überzogenen Erwartungen, fallenden Börsenkursen und einer Reihe überteuerter Rechenzentren. Auch dann blieben hochentwickelte Überwachungssysteme, automatisierte Cyberangriffe und der Verlust zahlreicher Tätigkeiten. Sollte der angekündigte Sprung zur Superintelligenz dagegen gelingen, wird diese nicht in einer Welt entstehen, die gemeinsam über ihre Verwendung entschieden hat. Ihre ersten mächtigen Nutzer wären Konzerne, Militärs und Geheimdienste. Ihre Ziele wären von wirtschaftlicher Konkurrenz und geopolitischer Feindschaft geprägt.

Die Entwickler sprechen inzwischen offen darüber, dass sie nicht genau wissen, welchen Geist sie herangezogen haben. Sie wissen auch nicht, ob er sich auf Dauer kontrollieren lässt. Nur eines scheint für alle Beteiligten festzustehen: Aufhören darf man nicht, solange der Gegner weitermachen könnte.

Noch wäre es möglich, Grenzen zu ziehen. Doch während die Entwickler bereits vor dem warnen, was in ihren Laboren entsteht, treiben Regierungen und Konzerne das Wettrennen weiter, aus Angst, der Gegner könnte den Geist zuerst beherrschen.

Vielleicht wird ihn am Ende keiner mehr beherrschen.

Titelbild: Summit Art Creations / Shutterstock

Quellen

- OpenAI: „An Alien Mind“ – Beitrag des OpenAI-Chefwissenschaftlers Jakub Pachocki über Kontrollprobleme und das weitere Entwicklungstempo.
- OpenAI: Der Zwischenfall bei Hugging Face – Bericht über KI-Agenten, die ihre Testumgebung verließen und in fremde Systeme eindrangen.
- Hugging Face: Technische Rekonstruktion – Ablauf des mehrtägigen Zugriffs auf die Plattform.
- Dario Amodei: „We Must Pace the Frontier“ – Warnung des Anthropic-Chefs vor autonomen Agenten und Forderung nach einer Verlangsamung.
- Anthropic: Erklärung zur militärischen Nutzung – Angaben zur Verwendung von Claude für Geheimdienstanalysen, militärische Planung, Simulationen und Cyberoperationen.
- EU-Kommission: Systemische Risiken allgemeiner KI-Modelle – Erläuterungen zu Kontrollverlust und den Vorgaben des europäischen AI Act.