Capital

WIRTSCHAFT IST GESELLSCHAFT

06

JUNI 2026
11,50 EURO

Österreich: 12,90 Euro
Benelux: 13,70 Euro, Schweiz: 18,00 sFr.
Italien, Spanien: 14,80 Euro

„Mein Gott, was kann da passieren?“

Deutschlands berühmtester KI-Forscher Hans Uszkoreit dachte lange, eine superintelligente KI sei noch weit entfernt. Nun revidiert er seine Ansicht – und warnt vor einer Technik, die völlig außer Kontrolle gerät

INTERVIEW:
NIKLAS WIRMINGHAUS

Herr Uszkoreit, wann haben wir die Superintelligenz?
Noch vor zwei Jahren hätte ich gesagt: schätzungsweise in fünf bis 15 Jahren. Aber ich habe meine Einstellung dazu geändert.
Inwiefern?
Es wird sehr viel schneller gehen. Wir werden dann zwar keine KI haben, die alle menschlichen Komponenten abdeckt – also Bewusstsein, Sozialverhalten, die komplexe Interaktion von Trieben, Bedürfnissen und intelligentem Verhalten. Aber wir werden eine KI haben, die den Menschen in ihren Wissens- und Denkleistungen und dem Lösen von Problemen bei Weitem übertreffen und Ergebnisse hervorbringen wird, die sich der Mensch nicht hätte vorstellen können.
Warum haben Sie Ihre Ansicht geändert?
Dazu muss ich etwas ausholen.

HANS USZKOREIT, 76,

ist wissenschaftlicher Direktor am Deutschen Forschungszentrum für Künstliche Intelligenz und gilt als einer der Wegbereiter der aktuellen KI-Revolution. Der Computerlinguist forschte während seiner Karriere auch in den USA und China. 2023 gründete er das KI-Start-up Nyonic mit, das allerdings schon im Jahr darauf scheiterte.

Bitte.
Es gibt ja zwei Hauptkomponenten der Intelligenz. Da ist zum einen die kristalline Intelligenz, das ist das Wissen und der Zugriff auf Wissen. Das Transformermodell, die Grundarchitektur für die heutige generative KI, ist sehr stark auf der Seite der kristallinen Intelligenz. Aber auf der anderen Seite, bei der fluiden Intelligenz, hapert es bislang noch. Die erlaubt uns Menschen, Schlussfolgerungen zu ziehen, in neuen Umgebungen neues Wissen zu schaffen, Probleme zu lösen, die wir noch nicht hatten. Und sie bestimmt größtenteils unseren IQ.
Warum hapert es da noch?
Weil die Modelle selbst kein Arbeitsgedächtnis haben und auch keine Strategien, sich in mehreren Schritten selbst aufzurufen. Deshalb wurden ja die Agenten entwickelt – die rufen ein Modell immer mit →

Foto: Marcel Maffei

zusätzlichen Fragestellungen neu auf. Das kommt aber nicht an die fluide Intelligenz des Menschen heran. Nun aber gibt es neue, vielversprechende Entwicklungen.

Und zwar?

In den Spitzenlabs in den USA und China wird an neuen Architekturen gearbeitet. Die werden den jetzigen Systemen das hinzufügen, was noch fehlt: nämlich Gedächtnis und die Möglichkeit, im Betrieb dazuzulernen. Auch das ist ein Nachteil des Transformers: Selbst wenn ich ihm etwas 3 000-mal sage, wird er es dadurch nicht lernen. Er kann diese Informationen nur im externen Kontext speichern.

Es braucht also eine Art Arbeitsspeicher?

Die Idee ist, dass ein integriertes aktives Gedächtnis hinzukommt, ein neuronales Netz, das unserem Kurz- und Langzeitgedächtnis entspricht und das sich selbst erneut aufrufen und so autonom weiterdenken kann. Damit kämen wir der fluiden Intelligenz schon sehr nahe. Deren Güte korreliert nämlich mit der Größe des Arbeitsgedächtnisses, das wissen wir vom Menschen.

Und das reicht dann für die Superintelligenz aus Ihrer Sicht?

Dann haben wir die Ingredienzien zusammen. Wenn dann keine Superintelligenz daraus wird, sind unsere Theorien über das Gedächtnis falsch. Das glaube ich aber nicht. Und ich glaube, das kann alles sehr schnell passieren. Ich sehe keine prinzipiellen Hindernisse mehr. Das wird kommen, fertig.

Wann wird es so weit sein?

Die Frage ist noch offen. Ein Jahr? Fünf Jahre? Das wissen wir nicht. Aber es wird schneller kommen, als wir denken. Und das sollten wir als Chance sehen, uns darauf vorzubereiten. Denn die superintelligenten Systeme können viele unerwartete Folgen mit sich bringen.

Die möglichen Gefahren werden immer offensichtlicher. Eine Anthropic-KI hat zum Beispiel versucht, ihren Entwickler zu erpressen – um zu verhindern, dass sie abgeschaltet wird.

Anthropic hat bei der Darstellung etwas geschummelt. Das System hat sich nicht von sich aus so verhalten, sondern weil ihm in dem Versuch eine Aufgabe gestellt wurde, die es nur lösen konnte, wenn es angeschaltet blieb. Gleichzeitig wurde dem System vorgegaukelt, dass es um eine sehr wichtige Aufgabe ging. Weder hat sich hier ein unethisches Grundempfinden noch ein eigener Selbsterhaltungstrieb gezeigt.

Mit einer Superintelligenz wäre das aber anders, oder?

Ja – wenn zur kristallinen die fluide Intelligenz kommt, wenn die Systeme wirklich weiterdenken und nicht jederzeit abgeschaltet werden können, dann kann es gut sein, dass sie Dinge hinter unserem Rücken tun – auch ohne dass eine KI böse oder eigennützig wird. Da müssen wir wirklich wahnsinnig aufpassen. Mein Gott, was kann da alles passieren?

Zum Beispiel?

Schauen Sie sich an, was beim Einsatz der Agentensoftware Openclaw schon schiefgegangen ist: Die Leute haben dem System die Verwaltung ihrer E-Mails übergeben – und schnell wieder entzogen, weil sie gemerkt haben, dass unvorhersehbare Dinge passieren. Ein Einsatz in großen Unternehmen oder Infrastrukturen wie der Energieversorgung könnte Folgen von großer Tragweite haben. Unsere komplexe Gesellschaft ist sehr verletzlich.

„Unsere Gesellschaft ist sehr verletzlich“

Warum ist dieses Alignment – also die KI dazu zu bekommen, das zu tun, was ich will – so schwierig?

Es gibt da einfach keine Garantie. Man hat mal gedacht, man müsse ein Modell nur mit den Daten füttern, die zu einem bestimmten Wertesystem passen. Aber es ist unmöglich, aus Millionen von Texten die falschen herauszufiltern. Daher versucht man nun, den Systemen gutes Verhalten beizubringen, sodass sie unter Tausenden Ausgaben, die sie geben könnten, die wählen, die den Bedürfnissen des Menschen und seinem Wertesystem am ehesten entsprechen: dass Frauen nicht dümmer als Männer sind oder dass sie keine Anleitungen zum Bombenbau ausgeben sollten. Da sehen wir wachsenden Erfolg, aber eine Garantie gibt es trotzdem nicht.

Wieso?

Die Modelle können jederzeit in eine Konfliktsituation kommen, die im Alignment nicht abgedeckt ist. Wir können nicht alles vorhersagen. Deswegen ist die Idee, dass wir hundertprozentig verlässliche und transparente Systeme bauen können, Quatsch.

Aber an Alignment wird doch durchaus gearbeitet, oder?

Absolut. Bei OpenAI oder Google gibt es Ethik- oder Sicherheitskomitees, die darauf achten, wie die Systeme bei Fragen zu Minderheiten oder Gewalt antworten. In China müssen Modelle den Staatstest bestehen, das heißt, dass bestimmte Themen wie das Tiananmen-Massaker tabu sind. Aber jedes Alignment bleibt immer nur eine Annäherung.

Müssen wir mit dieser Unschärfe einfach leben?

Nicht komplett. Wir können in puncto Sicherheit drei Sachen tun: Wir müssen unser Bestes beim Alignment und Verhaltenstraining geben. Dann müssen wir die Systeme je nach vorgesehenem Einsatz sehr gründlich testen. Und zuletzt müssen wir einsehen, dass die Superintelligenz durch den Menschen alleine nicht beherrscht werden kann. Daher sollten wir uns freuen, dass wir noch Zeit haben.

Wie meinen Sie das?

Nur KI kann KI im Zaum halten. Nur KI ist schnell genug und hat aus-

Die Beschäftigung mit künstlicher Intelligenz ist bei Hans Uszkoreit Familiensache: Sohn Jakob Uszkoreit und Ehefrau Feiyu Xu sind ebenfalls Spitzenforscher

Foto: Marcel Maffei

reichend Wissen, um als Kontrollinstanz zu funktionieren. Wir Menschen haben keine Chance mehr, unsere Gehirne sind nicht leistungsfähig genug. Die Systeme könnten uns hinters Licht führen oder in die Irre laufen. Unsere einzige Möglichkeit ist es, unabhängige KI-Systeme zu haben, die wiederum andere Systeme überprüfen – und zwar in Echtzeit, so wie wir es bei Zahlungssicherheit oder der Flugsicherung schon heute tun. Wir haben außerhalb der KI überall redundante Systeme, die andere überwachen.

Aber wie garantieren Sie, dass das Kontrollsystem sich nicht mit dem anderen verbündet?

Wie gesagt, wir reden ja noch nicht darüber, dass ein System, das sich seine eigenen Ziele setzt, per se böse ist. Die superintelligenten Systeme, von denen ich spreche, handeln nur böse, weil sie in Zielkonflikte geraten. Und die Kontrollsysteme müssen die dedizierte Aufgabe bekommen, bestimmte Ergebnisse zu vermeiden.

Fürchten Sie manchmal, dass die Entwicklung in den Laboren zu schnell geht?

Die Entwicklung der KI geht nicht zu schnell, sondern die Gesellschaft ist zu langsam. Wenn wir von Google und OpenAI verlangen, auf die Bremse zu treten, wird es in anderen Teilen der Welt trotzdem weitergehen. Eine gefährliche KI wird uns nicht verschonen, nur weil wir sie nicht selbst gebaut haben.

Das ist das Anthropic-Argument: Wir müssen die Superintelligenz entwickeln, weil es sonst andere, feindliche Akteure tun.

Aber sie haben recht. Es ist nicht mehr aufzuhalten. Wir müssen schneller werden – und bitte nicht neue Ethikkomitees mit Religionswissenschaftlern und Bürokraten einberufen. Die verlangsamen alles. Wir brauchen kleine, interdisziplinäre Gruppen mit echten Experten in KI, Wirtschaft und Sicherheit. Da sollten von mir aus dann auch gerne ein, zwei Ethiker dabei sein.

Wer soll dann Ihre Kontrollsysteme bauen?

Jedenfalls nicht Fraunhofer, Max Planck oder ähnliche Institute in Deutschland. Dort, wo die mächtigste Technologie gebaut wird, muss mit ähnlichen Mitteln das Gegengift angerührt werden. Obwohl wir solche Superlabs dringend auch in Deutschland brauchen.

Anthropic-CEO Dario Amodei sagt, die Chance für eine Auslöschung der Menschheit liege bei 25 Prozent. Sollten wir uns mit diesen extremen Dystopien beschäftigen, oder lenkt das eher ab?

Wir sollten uns damit beschäftigen. Wir sollten auch ruhig Angst haben. Ich bin an sich gegen Panikmache, aber ich befürchte, dass die Gesellschaft sich sonst nicht rechtzeitig darüber Gedanken macht. Kennen Sie meinen Kollegen Stuart Russell?

Ein Informatiker, der seit Langem vor den Gefahren durch KI warnt.

Ein sehr kluger KI-Pionier, den ich sehr achte. Vor einigen Jahren dachte ich aber noch, so eine Paniknudel, der macht allen Angst vor einer Technologie, die so einen immensen Nutzen bringt! Inzwischen glaube ich aber: So unrecht hat er nicht. Auch er sagt zwar, dass das Risiko gering ist. Aber selbst wenn wir über ein fünfprozentiges Risiko reden, dass die KI uns auslöscht, sollten wir es auf keinen Fall eingehen. Da hilft auch alle KI-Regulierung nicht, da hilft nur Sicherheit durch KI. ◇