KI ist nicht schuldig. Die Verantwortung liegt beim Menschen!

Reinhard Karger, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Unternehmenssprecher und Mitglied des DFKI-Aufsichtsrats, https://www.dfki.de

1993 habe ich meine Tätigkeit beim Deutschen Forschungszentrum für Künstliche Intelligenz in Saarbrücken aufgenommen. KI war damals ein Nischenthema. 33 Jahre später findet es in den Medien breite Beachtung. Die Schlagzeilen zum Thema KI sind aktuell überwiegend kritisch:
„Bill Gates warnt: Die turbulente Ära der KI hat begonnen." (FAZ, 30.08.2026).
„Erleben wir das nächste Jahrzehnt nicht mehr " (FAZ, 10.09.2026).
„Bringt die KI uns alle um?" (FAZ, 13.09.2026).
„KI: Das bedrohte Fortschrittsversprechen." (FAZ, 16.09.2026.)
„Warum lügen, betrügen und koordinieren sich KI-Agenten?" (FAZ, 19.09.2026).
„KI könnte ein Risiko für die gesamte Menschheit sein." (FAZ, 24.09.2026).
(die Quellen/URLs dazu weiter unten)

John von Neumann, ein sehr einflussreicher Informatiker, stellte die allgemeinere Frage: „Können wir Technologie überleben?" Seine Abhandlung wurde 1955 veröffentlicht. Das Thema ist nach wie vor wichtig. Doch es gibt hier ein tiefes Missverständnis, das zwar natürlich und verständlich ist, aber für eine rationale Diskussion nicht hilfreich. Mehr dazu weiter unten.

Heute sprechen wir von generativer KI, Sprachtechnologie und großen Sprachmodellen (Large Language Models, LLMs). Wahrscheinlich und hoffentlich hat jeder seit November 2022 ChatGPT genutzt, um sich persönliches Erfahrungen anzueignen. Die Nutzung ist außerordentlich intuitive. Die Ausgaben sind sprachlich hervorragend formuliert. Und es gibt Versionen, bei denen keine Kosten entstehen.

Während ChatGPT eine Eingabe verarbeitet, erscheint auf dem Bildschirm: „Denkt nach…" oder „Thinking…". Das Problem: diese LLMs denken nicht. Oder wie Yann LeCun auf dem Weltwirtschaftsforum in Davos 2024 sagte: „Große Sprachmodelle, von denen alle so begeistert sind, nehmen keine sinnlichen Reize wahr, haben kein Gedächtnis, führen keine Schlussfolgerungen oder Inferenz durch und generieren keine Handlungen. Sie tun nichts von dem, was intelligente Systeme tun oder tun sollten."

Yann LeCun ist der ehemalige Leiter des Meta AI Lab in Paris und einer der drei Pioniere der generativen künstlichen Intelligenz – zusammen mit Geoffrey Hinton, dem Nobelpreisträger für Physik 2024, und Yoshua Bengio, dem Leiter von MILA, dem Montreal Institute for Learning Algorithms. Yoshua Bengio sprach am 23. September bei der Sitzung des Sicherheitsrats der Vereinten Nationen über KI und die Zukunft der Menschheit.

Die anhaltende Debatte darüber, ob KI die Menschheit innerhalb der nächsten X Jahre auslöschen wird, geht zurück auf den sogenannten „Hugging-Face-Vorfall", über den ab dem 16. Juli weltweit in der Presse berichtet wurde – es ging dabei um agentische KI von OpenAI, die in Server von Hugging Face eingedrungen ist.

Eine echte Dynamik entstand durch den X-Post von Jacob Coxon, einem ehemaligen Mitarbeiter von Anthropic, vom 09.09.2026, in dem er schrieb: „Ich habe heute bei Anthropic gekündigt. Ich habe die letzten drei Jahre damit verbracht, sowohl bei OpenAI als auch bei Anthropic an der Forschung zum Vortraining zu arbeiten. Keines der beiden Unternehmen handelt verantwortungsbewusst." Seine Pointe lautete: „Sie rasen geradewegs auf eine sich selbst verbessernde Superintelligenz zu und spielen mit unserem Leben." Die technische Abkürzung dafür lautet RSI – rekursive Selbstverbesserung.

Da stehen wir nun. Mein Telefon klingelt. Journalisten wollen wissen, ob oder wann der Terminator vor der Tür steht. Nein, der Terminator steht nicht gerade vor der Tür. Das heißt aber nicht, dass die Frage falsch ist. Sie zielt nur in die falsche Richtung.

Jensen Huang, CEO und Gründer von NVIDIA, betont, dass wir uns nicht als Opfer darstellen und die Öffentlichkeit nicht mit Untergangsszenarien, Panikmache oder Skepsis verunsichern sollten. Seit 18 Jahren spielt das US-amerikanische Unternehmen eine entscheidende Rolle bei der Lieferung der Chips, der Hardware für die KI-Verarbeitung. Und NVIDIA wird der Hauptlieferant für die KI-Giga-Fabriken sein, die derzeit von der EU-Kommission diskutiert werden. Im September 2026 hatte NVIDIA die weltweit höchste Marktkapitalisierung. Jensen Huang argumentiert recht unverblümt. Seiner Ansicht nach tragen Unternehmen und ihre Führungskräfte die volle Verantwortung im Sinne der rechtlichen Haftung für Schäden, die durch unsichere oder schlecht getestete Produkte entstehen, die sie auf den Markt gebracht haben.

Und damit hat er durchaus Recht. Wir verfügen über Vorschriften, Gesetze, Normen, Standards, Verfahren und Institutionen. Wir haben Anwälte und Gerichte. Es gibt die Gesundheitsbranche, die Luftfahrtindustrie und die Finanzbranche mit sehr strengen Vorschriften, wenn neue Produkte oder Dienstleistungen eingeführt werden. Es gibt bereits zahlreiche Regelungen.

Die eigentliche Frage lautet: Brauchen wir zusätzliche Vorschriften für KI? Jensen Huang sagt: Nein. Bill Gates sagt: Ja. Gates argumentiert überzeugend, dass eine Schicht mit zusätzlichen Supervisionsmodulen erforderlich ist, die dafür zuständig ist, den böswilligen Einsatz von KI für jegliche Art von Cyberangriffen oder Bioterrorismus zu verhindern. Die Probleme bestehen darin, dass die Restriktionen, die von den übergeordneten Supervisionsmodulen angewendet werden, explizit formuliert werden müssen, dass sie technisch umgesetzt werden müssen, dass die Governance durchgesetzt werden mus, dass Menschen in der Regel Wege finden, Einschränkungen zu umgehen.

Selbst wenn diese technischen Beschränkungeen erfolgreich umgesetzt werden, müssen wir zwischen interaktionsorientierter KI und agentischer KI unterscheiden.

Wenn wir interaktionsorientierte KI-Dienste – wie ChatGPT – nutzen, können wir sie als Suchmaschinen zum Auffinden von Webseiten oder z.B. als Assistenten zum Erstellen von Textpassagen verwenden. Wenn wir sie jedoch als Antwortmaschinen nutzen, die Aussagen oder Lösungen generieren, dann brauchen wir vertrauenswürdige KI, die zuverlässige Ergebnisse liefert. Daher hat ein Unternehmen, das Antworten generiert, die Verpflichtung und steht in der Haftung, korrekte Antworten zu liefern, die präzise und nachvollziehbar sind.

Im Alltag suchen wir nach Abkürzungen und werden von der Sehnsucht nach Bequemlichkeit angetrieben. Wir wollen Antworten und keine Sammlung von Links zu Webseiten, die wir durchgehen müssen, um die Antwort selbst zu finden. Wir wollen Antworten, die wir in persönlichen oder geschäftskritischen Entscheidungsprozessen nutzen können. Doch die heutigen generativen KI-Systeme liefern Ergebnisse, die nur vielleicht oder hoffentlich richtig sind. Sie sind nicht nachweisbar korrekt. Daher sollten wir sie nicht als selbstverständlich ansehen. Es ist deshalb unerlässlich, dass die Nutzer nicht ihrem veständlichen Wunsch nach bequemen Abkürzungen nachgeben, sondern die Ergebnisse stets selber überprüfen – oder überprüfen lassen.

Um vertrauenswürdige KI zu erreichen, müssen wir in großem Umfang in die Forschung investieren, die datengesteuerte generative KI erfolgreich mit regelorientierten symbolischen

KI-Ansätzen verbindet. Diese hybriden oder neurosymbolischen KI-Systeme sollen die Vorzüge der Mustererkennung und die Robustheit der generativen KI mit der Zuverlässigkeit wissensgesteuerter KI verbinden. Zuverlässigkeit ist dabei das das Ziel- und Schlüsselkonzept. Symbolische KI kann Schlussfolgerungen im Sinne von Deduktion, Induktion und Abduktion ziehen, die auf den Konzepten von Ursache und Wirkung in den Naturwissenschaften (Kausalität) sowie von Vernunft und Konsequenz in der kulturellen Welt (Rationalität oder besser „Reasonality“) beruhen.

Wissenschaftlich gibt es viel zu tun, und es gibt eine echte Chance, eine europäische „TrustedAI“ zu entwickeln, die im Erfolgsfall einen großen Sprung in Richtung zukünftiger KI-Fähigkeiten und europäischer KI-Souveränität darstellen wird. Natürlich kann und sollte KI nicht als Wahrheitsmaschine betrachtet werden. Wir müssen im Auge behalten und überwachen, dass bei den erzeugten Ausgaben der Korrektheitsnachweis evidenzbasiert ist und nicht von Interessengruppen jeglicher Art gesteuert wird.

Wir müssen die Risiken vertrauenswürdiger Antwortmaschinen im Blick behalten. Doch eine Wahrheit halte ich für selbstverständlich: Die Menschheit wird nicht verschwinden, wenn Maschinen Antworten liefern, die in dem Sinne wahr sind, dass sie mit verifizierten Fakten der realen Welt übereinstimmen, die von relevanten Quellen unabhängig bestätigt wurden. Es ist jedoch klar, dass dies nur gilt, wenn Antwortmaschinen – wie zuvor erwähnt – durch Supervisionsmodule technisch überwacht werden.

Nun müssen wir uns mit den Schlagzeilen befassen, die ich zu Beginn meines Vortrags zitiert habe. Agenten-basierte KI und multiagentische-KI-Systeme sind äußerst leistungsstarke Werkzeuge. Hoffentlich werden sie bahnbrechende Erkenntnisse in der Medizin, der Physik oder bei neuen Materialien liefern, globale soziale Ungerechtigkeit verringern und dazu beitragen, den Klimawandel einzudämmen.

Werden diese KI-Agenten jedoch zu Forschungszwecken eingesetzt, sind sie autonom – sie sind Werkzeuge, die selbstgenerierten Zielen folgen. Genauer gesagt: eigenen Gesetzen, wenn man an die Bedeutung des antiken griechischen Begriffs denkt. Philosophisch ausgedrückt: Menschliche Autonomie ist „die moralische Fähigkeit, nach den eigenen Prinzipien zu handeln“. In der generativen KI ließe sich dies wie folgt übersetzen: Maschinenautonomie ist „die technische Fähigkeit, nach eigenen Wahrscheinlichkeiten zu handeln“. Durch den Einsatz autonomer KI-Agenten sehen wir uns mit einem unbegrenzten, exponentiellen Kombinationspotenzial konfrontiert, das unvorhersehbare Auswirkungen haben könnte.

Wie können wir sicherstellen, wie können wir garantieren, dass KI-Agenten keine Ziele verfolgen, die Menschenleben gefährden? Dies wird als „Problem der Ausrichtung“ oder “Alignment” bezeichnet. Die Antwort lautet derzeit: Wir können es nicht.

Was sollen wir also tun? Wir sollten kein Risiko eingehen. Die beste Strategie ist derzeit: gesetzlich durchgesetzte Eindämmung oder Containment. Eine mögliche Analogie ist die Einstufung der Biosicherheitsstufen (Biosafety Level, BSL 1–4). Biolabore, die mit infektiösen Substanzen arbeiten, die tödliche Krankheiten verursachen könnten, für die die Menschheit heute noch kein Heilmittel hat, werden als BSL 4 eingestuft. Diese Labore unterliegen rigiden Auflagen und werden streng kontrolliert. Ja, sie sind nicht ausfallsicher. Aber die Biosicherheitsstufen bieten eine Orientierung dafür, wie ernsthafte Vorsichtsmaßnahmen in der Forschung aussehen.

KI-Unternehmen sollten dazu verpflichtet werden, vollständig isolierte Umgebungen zu entwickeln – KI-Sicherheitsstufe 4 (AI Safety and Security Level, AISL4) –, damit agentische KI keinerlei Zugang zur realen Welt im Sinne des öffentlichen Internets hat. Eine strenge

Eindämmung wird die Risiken autonomer KI-Agenten verringern und hoffentlich beseitigen.

Eindämmung schafft Chancen: Innerhalb dieser geschlossenen Umgebungen können und sollten agentische KI genutzt werden können, um all jene neuen Medikamente und Materialien sowie sozialen und wissenschaftlichen Erkenntnisse zu gewinnen, die wir brauchen, um unsere Welt besser und unser Leben länger zu machen.

Dieser dialektische Ansatz schafft viele Chancen und verringert Risiken. Das Problem ist, dass er sich nicht so einfach vermitteln lässt wie „KI ist eine Revolution, und wer am schnellsten ist, wird gewinnen“. Oder „wir dürfen nicht zulassen, dass andere Nationen, Unternehmen oder Kontinente das KI-Rennen gewinnen“.

Wie im Titel dieses Vortrags erwähnt: KI-Systeme sind weder „schuldig“ noch „verantwortlich“ für Auswirkungen oder schädliche Ereignisse, die durch ihren Einsatz verursacht werden. Dies sind lediglich Nebenwirkungen, die in der Kombination aus menschlicher Innovationskraft und mangelnder menschlicher Aufsicht begründet liegen. Die bittere Erkenntnis ist, dass der Schaden für die Menschheit lediglich ein unbeabsichtigter Nebeneffekt einer Technologie wäre, die wir nicht auf verantwortungsvolle Weise kontrolliert haben.

Nun müssen wir auf die einleitende Bemerkung eingehen, dass hier ein tiefgreifendes Missverständnis vorliegt. Eine Ursache dafür ist das sogenannte „Problem der anderen Köpfe“. In unserem Alltag gehen wir davon aus, dass andere Menschen Absichten und Wünsche haben, weil jeder seine eigenen Bedürfnisse und Wünsche kennt. Das ist eine sehr wirkungsvolle Strategie, die es uns ermöglicht, die nächsten Handlungen anderer Menschen vorauszusehen und Freunde oder Feinde sehr schnell zu erkennen.

Wir verallgemeinern diese Strategie zu stark und neigen von Natur aus dazu, unsere gesamte Umgebung und unsere technischen Geräte zu vermenschlichen oder zu anthropomorphisieren. In der Informatik „beendet“ man einen Prozess. Aber der Prozess ist nicht tot. KI-Chatbots denken nicht und KI-Agenten wollen nichts.

Das tiefgreifende Problem der KI besteht darin, dass sie Fähigkeiten nachahmt, die wir als Kernmerkmale des Menschen schätzen – die den Menschen von Steinen, Tieren und Pflanzen unterscheiden. Die den den Menschen zum Menschen mchen. In der Diskussion über KI sind wir akuell Opfer unserer eigenen evolutionären Erfolgsformel. Wir vermenschlichen die KI. In den vergangenen Jahren haben wir erfahren, dass KI-Systeme beispielsweise im Schach, in der Radiologie oder in der Dermatologie “übermenschliche” Leistungen erbringen. Daher neigen wir dazu, uns einen Wettstreit vorzustellen, einen epischen Kampf Mensch gegen Maschine. Letztendlich ist dies ein Problem der Begriffe, mit denen wir die Fähigkeiten von KI-Systemen beschreiben.

Um die Debatte zu entschärfen, sollten wir an unserer Sprache arbeiten. Zunächst einmal:
Es gibt nicht „die KI“, sondern ein „KI-System“. Dann:
KI-Systeme haben keine „Ziele“, sondern „Aufgaben“.
KI-Systeme haben keine „Willenskraft“, sondern „elektrische Energie“.
KI-Systeme haben keine „Wünsche“, sondern „Belohnungsfunktionen“.

Wir müssen mit den Worten, die wir verwenden, ernsthafter und viel bedachter umgehen. Aus Marketing-Sicht ist es jedoch naheliegend, Begriffe zu verwenden, die eine positive emotionale Wirkung haben. Vielleicht sind sie metaphorisch und nicht präzise, vermitteln aber Gefühle wie einen warmen Sommermorgen, um Menschen dazu zu bewegen, ein Produkt zu kaufen oder eine Dienstleistung in Anspruch zu nehmen.

Aus Sicht des Wahlkampfs ist es verlockend, kurze Sätze, glänzende Konzepte und Schlagworte zu verwenden, die schnell zu verstehen und leicht zu merken sind. Das ist eine ganz natürliche und erfolgreiche Wahlkampfstrategie, um Stimmen zu gewinnen.

Aber aus gesellschaftlicher Sicht ist es klug, verständlich und konstruktiv zu sein, eine wirklich gute Orientierung zu bieten und einen wohlwollenden Prozess der informierten Differenzierung zu fördern – in dem Sinne, dass die breite Öffentlichkeit und die Wähler unvoreingenommen und aus verschiedenen Blickwinkeln informiert werden, damit die Menschen im öffentlichen Diskurs vernünftige Standpunkte und keine unreflektierten Vorurteile austauschen. Damit sie die beste Lösung und die beste Entscheidung zur Maximierung des gesellschaftlichen Wohls finden.

Hier kommen wir alle ins Spiel: Politiker und politische Berater, Wissenschaftler, Journalisten, Coworking Spaces und Unternehmen. Sie und ich. Ich bitte uns alle inständig, die richtigen Worte und Erzählkonzepte zu finden, damit die europäische Gesellschaft klug handelt und sich in eine vernünftige Richtung bewegt.

Deshalb müssen wir es mit der Ausrichtung (Alignment), Eindämmung (Containment) und Kontrolle sehr ernst nehmen!

- Wir sollten zwischen interaktionsorientierter KI und agentenbasierten KI-Systemen unterscheiden!
- Wir sollten uns so sehr wie möglich bemühen, eine Sprache zu verwenden, die KI-Prozesse nicht anthropomorphisiert, die KI-Dienste nicht vermenschlicht.
- Wenn wir in diese Richtung handeln, können wir die öffentliche Meinung gewinnen, die KI-Kluft überwinden und europäische KI-Souveränität schaffen.

Abschließende Anmerkung: Zum Verfassen dieses Vortrags wurde kein Zum Verfassen dieses Vortrags wurde kein KI-Dienst verwendet.

Vielen Dank für die Einladung und Ihre Zeit.

Bill Gates warnt: Das turbulente Zeitalter der KI hat begonnen. FAZ, 30.08.2026, Bill Gates
https://www.faz.net/aktuell/wirtschaft/bill-gates-warnt-wir-sind-auf-die-kuenstliche-intelligenz-nicht-vorbereitet-accg-201170877.html

Erleben wir das nächste Jahrzehnt nicht mehr? FAZ, 10.09.2026, Piotr Heller
https://www.faz.net/aktuell/feuilleton/debatten/jacob-coxon-warnt-vor-ki-droht-die-ausloeschung-201213253.html

Bringt die KI uns alle um? FAZ, 13.09.2026, Patrick Bernau, Alexander Wulfers
https://www.faz.net/aktuell/wirtschaft/unternehmen/elon-musk-will-forschung-zu-ki-bremsen-bringt-die-ki-uns-alle-um-accg-201214335.html

KI: Das bedrohte Fortschrittsversprechen. FAZ, 16.09.2026, Alexander Armbruster
https://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz/anthropic-openai-co-das-versprechen-der-ki-und-wieso-die-gefahr-gross-ist-19983533.html

Warum lügen, betrügen und koordinieren sich KI-Agenten? FAZ, 19.09.2026, Yoshua Bengio
https://www.faz.net/aktuell/wirtschaft/unternehmen/ki-forscher-yoshua-bengio-warum-die-ki-agenten-ausarten-accg-201242795.html

KI könnte ein Risiko für die gesamte Menschheit sein. FAZ, 24.09.2026, Roland Lindner
https://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz/sam-altman-und-dario-amodei-warnen-vor-ki-risiken-201257130.html