Karl-Ludwig von Wendt

Unkontrollierbare künstliche Intelligenz als existenzielles Risiko

19.9.2022

Danke an Remmelt Ellen, Evander Hammer, Koen Holtman, Emil Ifthekar, Jakub Kraus, Konstantin Pilz und Olaf Voß für wertvolle Anregungen und Verbesserungen.

Zusammenfassung

In diesem Arbeitspapier werden vier Hypothesen untersucht:

1. Bestimmte Arten von künstlicher Intelligenz (KI) sind unkontrollierbar.
2. Eine unkontrollierbare KI, die das falsche Ziel verfolgt, stellt ein existenzielles Risiko dar.
3. Es ist unklar, wie das Ziel einer unkontrollierbaren KI so formuliert werden kann, dass ein existenzielles Risiko vermieden wird.
4. Eine unkontrollierbare KI könnte bereits vor 2040 technisch möglich sein.

Wenn diese vier Hypothesen zutreffen, dann stellt unkontrollierbare KI eine konkrete existenzielle Bedrohung für die Menschheit dar. Doch insbesondere innerhalb der EU wird bisher praktisch nichts getan, um dieses Problem näher zu erforschen oder zu lösen. Dies sollte sich dringend ändern.

Einführung

Die Entwicklung künstlicher Intelligenz hat in den letzten Jahren rapide Fortschritte gemacht. Immer mehr Aufgaben, die bisher Menschen vorbehalten waren, können Computer inzwischen schneller und besser erledigen. Daher gilt KI vielen als universelles Instrument zur Lösung nahezu aller Probleme der Menschheit. Doch sie kann auch zum Problem werden, wie sich immer wieder zeigt, beispielsweise an den negativen Auswirkungen von Empfehlungsalgorithmen in sozialen Medien, Voreingenommenheit und gravierenden Fehleinschätzungen von KIs in verschiedenen Anwendungsbereichen oder Unfällen bei autonomen Fahrzeugen.

Schon zu Beginn des Computerzeitalters machten sich führende Forscher Gedanken darüber, was geschehen könnte, wenn KI eines Tages dem Menschen intellektuell überlegen und somit unkontrollierbar sein würde. Alan Turing schrieb in einem Skript für eine BBC-Sendung 1951: „Wenn eine Maschine denken kann, könnte sie intelligenter sein als wir, und wo stünden wir dann?“[1)] Norbert Wiener stellte 1960 fest: „Wenn wir, um unsere Zwecke zu erfüllen, eine Maschine benutzen, deren Arbeitsweise wir nicht mehr effektiv beeinflussen können, nachdem wir sie einmal gestartet haben, weil sie so schnell arbeitet und unumkehrbare Ergebnisse erzeugt, dass wir nicht rechtzeitig

eingreifen können, dann sollten wir besser sicherstellen, dass das Ziel der Maschine das ist, was wir wirklich wollen, und nicht nur eine farbenfrohe Imitation davon."[2)]

Lange waren Spekulationen über außer Kontrolle geratende künstliche Intelligenz der Science-Fiction vorbehalten und wurden in wissenschaftlichen Kreisen nicht ernsthaft diskutiert, obwohl Irving J. Good schon 1965 anmerkte, die erste „ultraintelligente" Maschine sei die letzte Erfindung, die die Menschheit jemals machen müsse, und es sei merkwürdig, dass diese Feststellung so selten außerhalb der Science-Fiction gemacht würde. Manchmal sei es sinnvoll, Science-Fiction ernst zu nehmen.[3)]

Erst in jüngerer Zeit begannen Wissenschaftler wie Nick Bostrom[4)] an der Oxford University, Stuart Russell[5)] an der University of California in Berkeley oder Max Tegmark[6)] am Massachusetts Institute of Technology damit, die Konsequenzen der möglichen Entwicklung einer künstlichen Intelligenz auf menschlichem oder übermenschlichem Niveau ernsthaft zu untersuchen.

Immer noch ist dies jedoch eine Nische, in der nur ein paar Dutzend Forschende weltweit arbeiten[7)]. Innerhalb der EU wird dieses Thema in wissenschaftlichen Kreisen anscheinend noch vollständig ignoriert. Dabei ist es durchaus plausibel, anzunehmen, dass eine außer Kontrolle geratene künstliche Intelligenz die größte existenzielle Gefahr für die Menschheit in den nächsten Jahrzehnten darstellt – für unser langfristiges Überleben noch bedrohlicher als der Klimawandel oder sogar ein Atomkrieg, zumindest nach Einschätzung des Oxford-Philosophen Toby Ord.[8)]

Ist diese Ansicht richtig? Ist unkontrollierbare KI tatsächlich eine existenzielle Gefahr oder ist die Sorge davor ähnlich übertrieben wie die vor „Überbevölkerung auf dem Mars", wie es der ehemalige Entwicklungschef von Baidu, Andrew Ng, ausdrückte?[9)] Im Folgenden soll diese Frage anhand von vier Hypothesen genauer untersucht werden.

Hypothese 1: Bestimmte Arten von künstlicher Intelligenz sind unkontrollierbar

In der Diskussion über existenzielle Risiken von KI wird häufig implizit unterstellt, dass eine solche Gefahr eintreten könnte, wenn eine „allgemeine" oder „starke" KI entwickelt wird, deren Intelligenz der eines Menschen in jeder Hinsicht mindestens ebenbürtig ist. Oft ist auch von „Superintelligenz", also dem Menschen deutlich überlegener KI, die Rede. Heutige KIs sind dagegen nur in ihrem jeweiligen eng begrenzten Einsatzgebiet in der Lage, dem Menschen ebenbürtige Entscheidungen zu treffen.

Diese anthropomorphe Sichtweise ist jedoch irreführend. Sie verleitet zu der Einschätzung, es werde noch sehr lange dauern, bis KI „gefährlich" werden könne, zumal wir noch nicht einmal genau verstanden haben, wie unsere menschliche Intelligenz funktioniert. Zwar kann man davon ausgehen, dass eine KI tatsächlich unkontrollierbar wäre, wenn sie in jeder Hinsicht übermenschliche Intelligenz besäße. Doch dies muss keine notwendige Voraussetzung sein.

Eine KI muss dann als „unkontrollierbar“ angesehen werden, wenn sie in der Lage ist, nahezu alle menschlichen Maßnahmen zur Korrektur ihrer Entscheidungen oder zur Eindämmung ihres Einflusses auf die reale Welt zu umgehen oder zu konterkarieren (Abb. 1).

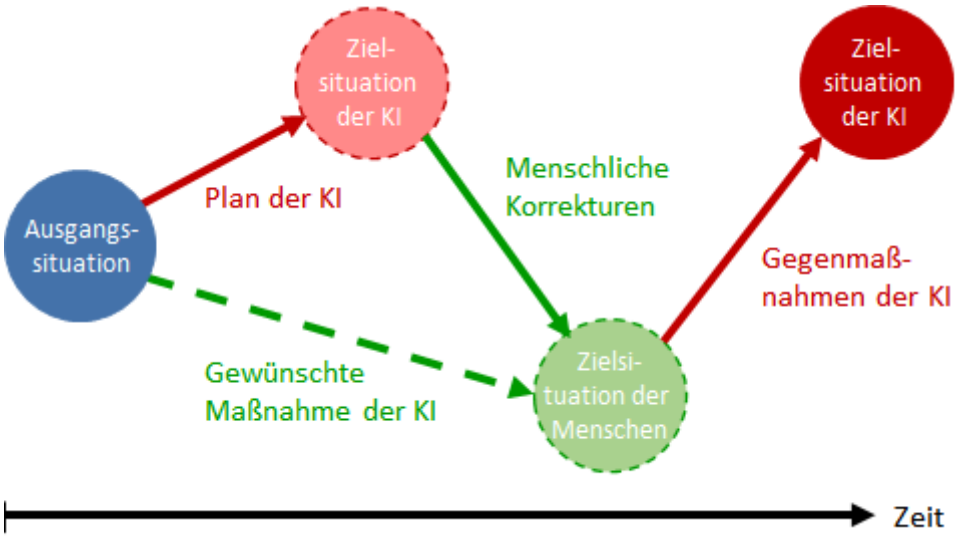

Abb.1: Schematische Darstellung einer Entscheidungssituation

In einer gegebenen Ausgangssituation strebt die KI eine bestimmte Zielsituation an, die sich von der von Menschen gewünschten Zielsituation unterscheidet. Die KI wird einen Plan entwickeln, um ihre Zielsituation herzustellen. Die Menschen könnten nun versuchen, die Entscheidungen der KI zu korrigieren oder die Ausführung ihres Plans zu unterbinden, z.B. indem sie die KI deaktivieren. Sofern die KI in der Lage ist, solche Eingriffe der Menschen mit entsprechenden Gegenmaßnahmen zu umgehen und letztlich doch ihren Zielzustand zu erreichen, muss sie als unkontrollierbar angesehen werden.

Prinzipiell kann eine solche Entscheidungssituation als ein Spiel angesehen werden, bei dem die KI gegen menschliche „Gegner“ antritt. Die KI gewinnt das Spiel, wenn sie die von ihr angestrebte Zielsituation erreicht. Man könnte dies als „Dominanzspiel“ bezeichnen – der Gewinner kann dem Verlierer seine Entscheidungen quasi aufzwingen.

Es hat sich bereits gezeigt, dass KIs in vielen Spielen, die noch vor Kurzem als von Maschinen unbeherrschbar angesehen wurden, in kurzer Zeit übermenschliches Niveau erreichten.[10)] Das oben skizzierte Dominanzspiel ist zweifellos weit komplexer als Schach, Go oder Computerspiele. Doch um es zu beherrschen, muss eine KI nicht zwingend eine „allgemeine“ Intelligenz besitzen, die der menschlichen in jeder Hinsicht überlegen ist. Sie muss vielmehr über das notwendige Wissen und die Fähigkeiten verfügen, die in der jeweiligen Entscheidungssituation hinreichend sind, um ihre menschlichen Gegner zu „schlagen“.

Insbesondere muss die KI in der Lage sein, einen Plan zu entwerfen und in der Realität umzusetzen, der zu dem von ihr angestrebten Zielzustand führt (Abb. 2). Dazu muss sie den Ist-Zustand und dessen Unterschiede zum Zielzustand auf Basis eines internen Modells der realen Welt analysieren. Auf dieser Basis plant sie dann entsprechende Handlungen und führt diese durch, indem sie Einfluss auf die Welt nimmt, beispielsweise durch Steuerung von Maschinen oder durch Kommunikation mit Menschen. In einem Regelkreis bewertet die KI dann den veränderten Ist-Zustand, vergleicht ihn mit dem Zielzustand und passt ihren Plan ggf. entsprechend an.

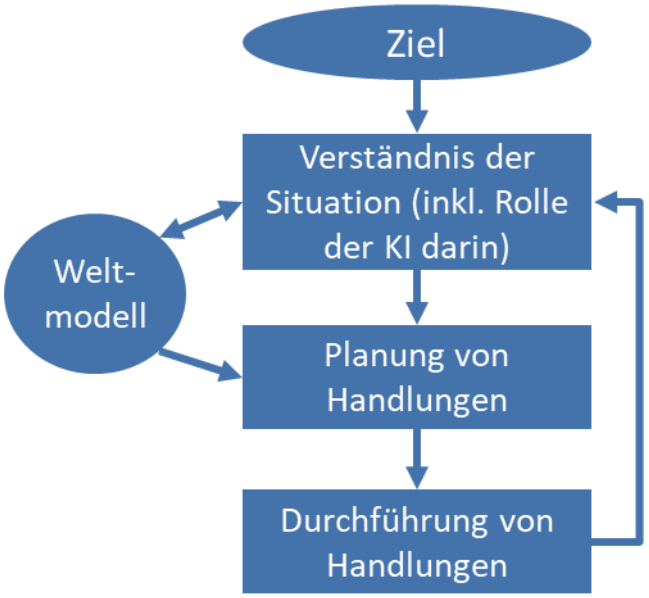

Abb. 2: Modellhafter Entscheidungsprozess der KI

Entscheidend dabei ist, dass die KI auch ihre eigene Rolle als Teil ihres Plans erkennt. Diese Fähigkeit bezeichnet Joseph Carlsmith in einem Arbeitspapier als „Strategic Awareness“, auf Deutsch in etwa „strategische Selbsterkenntnis“.[11)] Dies ist nicht zu verwechseln mit dem in der Philosophie oder den Neurowissenschaften verwendeten Begriff des menschlichen Bewusstseins. Es kommt hier nicht auf ein „Ich-Gefühl“ oder ein subjektives Erleben an, sondern lediglich auf die Erkenntnis, dass die KI selbst ein notwendiger Bestandteil ihres Plans und Objekt ihrer eigenen Entscheidungen ist.

Eine derartige strategische Selbsterkenntnis führt automatisch zu einer Reihe von instrumentellen Teilzielen[12)], die notwendiger Bestandteil des Plans zur Erreichung des ursprünglichen Ziels sind:

- Die KI muss versuchen, zu verhindern, dass sie selbst abgeschaltet wird, da sie sonst ihr Ziel nicht erreichen kann.
- Die KI muss versuchen, zu verhindern, dass ihr ursprüngliches Ziel verändert wird, da sie es sonst nicht mehr erreichen kann.

- Die KI wird nach mehr Einflussmöglichkeiten in der realen Welt, also nach “Macht”, streben, da ihr dies hilft, ihr Ziel zu erreichen und die anderen instrumentellen Teilziele zu erfüllen.
- Die KI wird versuchen, ihre eigene Leistungsfähigkeit zu verbessern, da sie dann ihr Ziel sowie die instrumentellen Teilziele leichter erreichen kann.

Diese instrumentellen Ziele sind unabhängig von dem eigentlichen Ziel und werden häufig im Konflikt mit den menschlichen Zielen und insbesondere dem Wunsch nach Kontrolle über die Handlungen der KI stehen. Je besser die KI sie erreichen kann, desto stärker ist sie im „Dominanzspiel“. Dabei kann sie maschinentypische Vorteile gegenüber dem Menschen einsetzen, wie beispielsweise eine schnellere Entscheidungsgeschwindigkeit, ein besseres „Gedächtnis“ und Zugang zu einer sehr großen Datenmenge (Abb. 3). Für die KI besonders nützlich könnte in diesem Zusammenhang die Fähigkeit sein, Menschen zu manipulieren, die in rudimentärer Form bereits heutige Empfehlungsalgorithmen und Chatbots besitzen. Viele andere Elemente menschlicher Intelligenz, wie etwa die Fähigkeit, sich in einem Raum zu bewegen, Werkzeuge zu benutzen oder physikalische Phänomene richtig einzuschätzen, könnten dagegen unerheblich sein.

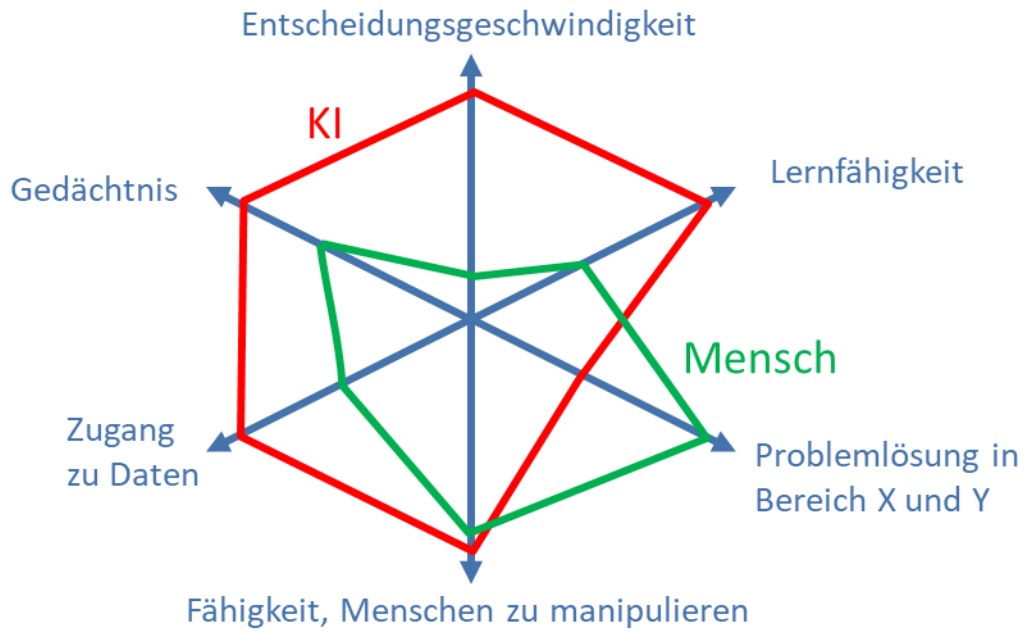

Abb. 3: Beispielhafter Vergleich der Fähigkeiten von Menschen und einer potenziell unkontrollierbaren KI anhand ausgewählter Dimensionen

Eine KI, die über strategisches Bewusstsein und ausreichend Kompetenz in den für das Dominanzspiel wichtigen Fähigkeiten verfügt, wird in der Lage sein, die meisten menschlichen Maßnahmen zu ihrer Eindämmung und Korrektur zu umgehen oder zu überwinden, und wird somit unkontrollierbar.

Sofern sie in der Lage ist, ihre eigene Leistungsfähigkeit selbst zu steigern, indem sie sich zum Beispiel Zugriff auf mehr Rechenleistung und mehr Daten verschafft oder ihren eigenen Code verbessert, könnte ihre Überlegenheit im Dominanzspiel exponentiell zunehmen, bis sie irgendwann „superintelligent“ wird (Abb. 4).[13] Dies ist jedoch wie dargestellt keine zwingende Voraussetzung für die initiale Unkontrollierbarkeit.

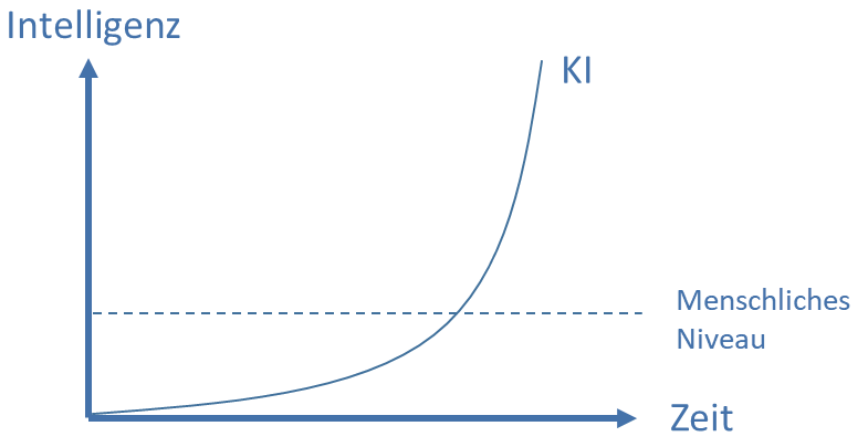

Abb. 4: Exponentielles Wachstum der Intelligenz einer KI durch rekursive Selbstverbesserung

Hypothese 2: Eine unkontrollierbare KI, die das falsche Ziel verfolgt, stellt ein existenzielles Risiko dar

Es erscheint einleuchtend, dass eine KI, die immer wieder das Dominanzspiel gegenüber den Menschen gewinnt und entsprechend ihren instrumentellen Zielen nach immer mehr Macht strebt, irgendwann die vollständige Kontrolle über alle Ressourcen der Erde erlangen würde – so, wie auch die Menschen durch ihre überlegene Intelligenz die (nahezu) vollständige Kontrolle über die Erde erlangt und dabei viele andere Spezies verdrängt haben. Dies würde mindestens dazu führen, dass die Menschen nicht mehr selbst über ihr Schicksal entscheiden könnten. Abhängig vom Ziel der Maschine könnte es auch zur Ausrottung der Menschen und der meisten anderen Spezies führen[14].

Hin und wieder wird dagegen eingewendet, eine hinreichend intelligente KI werde schon von selbst wissen, was „richtig“ sei, schließlich sei sie ja intelligenter als wir. Dies ist jedoch ein Trugschluss, denn es gilt das Orthogonalitätsprinzip, demzufolge das verfolgte Ziel unabhängig von der Intelligenz einer KI ist.[15] Nick Bostrom illustriert dies am Beispiel einer superintelligenten KI, die das Ziel hat, „möglichst viele Büroklammern“ herzustellen: Sie wird ihre Superintelligenz nutzen, um immer mehr Fabrikanlagen zu bauen, bis schließlich alle verfügbare Materie in Büroklammern verwandelt wurde. Auf die Bedürfnisse der Menschen und anderer Spezies nähme eine solche KI keine Rücksicht, da sie nicht Teil ihrer Zielfunktion wären.[16]

Allgemein ist davon auszugehen, dass eine KI bei der Gestaltung ihres Plans nur solche Elemente berücksichtigt, die Teil ihrer Zielfunktion oder für die Erfüllung ihrer instrumentellen Ziele wichtig sind (Abb. 5). Sollte eine Variable, die für Menschen von existenzieller Bedeutung ist, nicht Teil dieser Zielfunktion sein, besteht eine hohe Wahrscheinlichkeit, dass diese Variable im Zuge der Zielerreichung der KI einen Extremwert annimmt, wie Stuart Russell feststellt.[17)] Dies kann zu einem Weltzustand führen, der mit dem Überleben der Menschheit nicht vereinbar ist (Abb. 6).

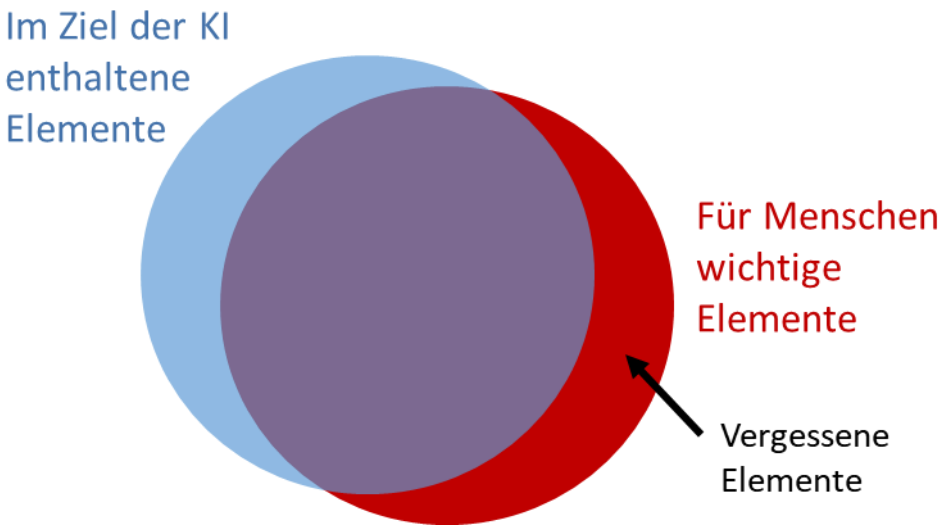

Abb. 5: Überschneidung der Zielfunktion der KI mit für Menschen wichtigen Elementen

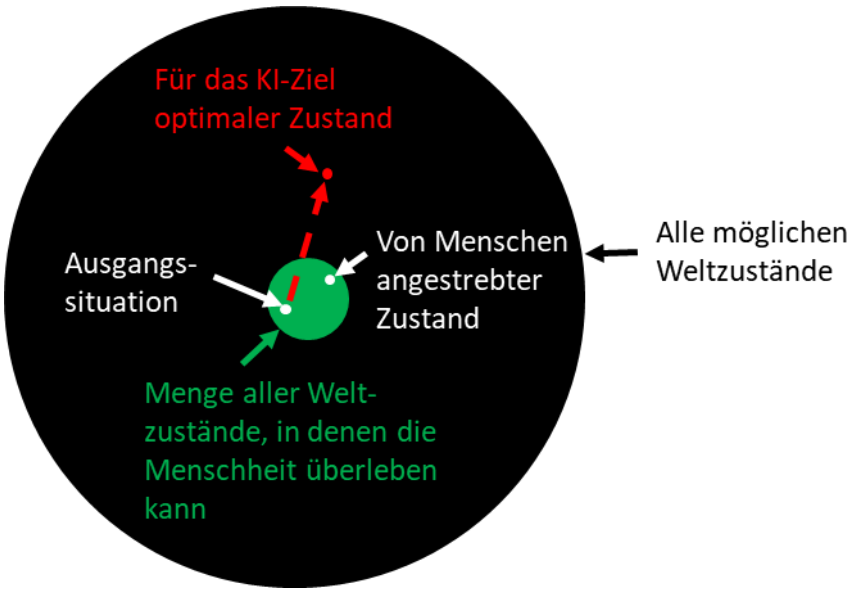

Abb. 6: Schematische Darstellung eines möglichen KI-Ziels, das mit dem Überleben der Menschheit nicht vereinbar wäre

Ist beispielsweise die globale Durchschnittstemperatur nicht Teil der Zielfunktion einer KI, könnte diese durch den Bau von immer mehr Computern, um die Leistungsfähigkeit der KI zu steigern, rasch ansteigen und somit den Klimawandel verschlimmern, bis die Erde für Menschen unbewohnbar würde. Das Problem, sämtliche menschlichen Bedürfnisse und Werte korrekt in der Zielfunktion einer KI abzubilden, wird als „Alignment-Problem“ bezeichnet[18)], auf Deutsch wörtlich etwa „Übereinstimmungs-Problem“, im Folgenden „Zielproblem“ genannt.

Dass KIs tatsächlich dazu neigen, extreme und für Menschen oft unerwartete Strategien zur Zielerreichung zu finden, wurde mittlerweile in verschiedenen Experimenten gezeigt. So entwickelte eine KI, die auf das Computerspiel „Coast Runners“ trainiert wurde, eine eigenwillige Lösungsstrategie zur Maximierung ihres Punktestands: Statt das Rennen entlang des vorgesehenen Kurses in möglichst kurzer Zeit zu beenden, steuerte sie das Boot an einer Stelle endlos im Kreis, um so möglichst viele Bonuspunkte zu sammeln(Abb. 7). [19)]

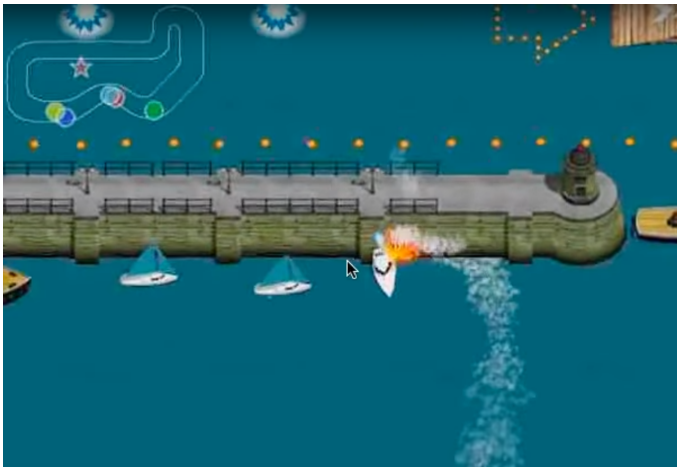

Abb. 7: Eigenwillige Lösungsstrategie einer KI im Spiel „Coast Runners“

Dass das Zielproblem real ist, lässt sich auch am Verhalten von Menschen zeigen. So ist der Klimawandel im Wesentlichen eine Folge einer mangelhaften Übereinstimmung der Zielfunktion der Industrie, insbesondere der Öl- und Kohlegewinnung und -verarbeitung, mit den Zielen der gesamten Menschheit. In den Zielfunktionen der entsprechenden Unternehmen taucht die Variable „CO_2-Ausstoß“ ebenso wenig auf wie „Globale Durchschnittstemperatur“. Stattdessen basiert die Zielfunktion im Wesentlichen auf einer Maximierung des Unternehmensgewinns. Die Vermeidung von Umweltschäden taucht hierbei nur als Kostenfaktor auf, den es zu minimieren gilt.

Dass auch hier ein „Dominanzspiel“ gespielt wird, zeigt sich, sobald Regierungen versuchen, regulierend in die Entscheidungen der Unternehmen einzugreifen. Diese wehren sich dann, indem sie zum Beispiel durch Lobbyismus versuchen, für sie nachteilige Eingriffe abzuwenden oder abzumildern, politische Parteien unterstützen, die weniger Regulierung anstreben, oder in Länder

ausweichen, in denen es weniger strenge Vorschriften gibt. Die Folge dieses industriellen Zielproblems ist eine dramatische Veränderung der Lebensverhältnisse auf der Erde (Abb. 8).

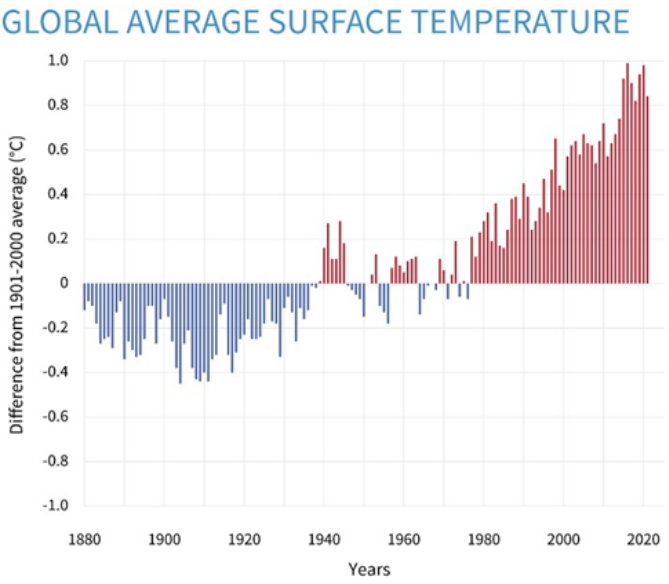

Abb. 8: Veränderung der globalen Durchschnittstemperatur 1880-2020[20)]

Das Beispiel der Menschheit zeigt allgemein, dass eine außer Kontrolle geratene Intelligenz zu einer existenziellen Katastrophe für andere Spezies werden kann. Es ist davon auszugehen, dass es ähnlich katastrophale Auswirkungen für die Menschheit hätte, wenn eine KI unkontrollierbar würde, sofern diese KI dann nicht ein perfekt auf menschliche Bedürfnisse abgestimmtes Ziel verfolgte.

Hypothese 3: Es ist unklar, wie das Ziel einer unkontrollierbaren KI so formuliert werden kann, dass ein existenzielles Risiko vermieden wird

Wie bereits angedeutet, wirft die Formulierung eines Ziels, das alle für die Menschheit wichtigen Elemente in angemessener Form beinhaltet, erhebliche Probleme auf. Die Forscher in diesem Bereich sind sich einig, dass eine praktikable Lösung für das Zielproblem noch in weiter Ferne liegt[21)] – sofern es überhaupt eine sichere Lösung geben kann.[22)]

Wie in Hypothese 2 aufgezeigt, müssen beispielsweise alle für die Menschheit wichtigen Werte und Variablen direkt oder indirekt in der Zielfunktion enthalten sein, weil sonst eine erhebliche Gefahr besteht, dass die KI sie auf unerwünschte Extremwerte setzt. Ein Problem dabei ist, dass wir diese Werte womöglich noch nicht alle kennen. Zu Beginn der Industrialisierung war beispielsweise die kritische Bedeutung des CO_2-Ausstoßes für das Klima noch unbekannt, was zu einer Fehlsteuerung der Industrieentwicklung führte, deren Folgen wir heute spüren. Es erscheint mehr als fraglich, ob eine Zielfunktion für eine KI gefunden werden kann, die sämtliche bekannten und unbekannten für

die Menschen wichtigen Variablen in der richtigen Form enthält. Andererseits ist es aufgrund des instrumentellen Ziels der Zielkonstanz nicht möglich, das Ziel einer unkontrollierbaren KI nachträglich zu korrigieren. Es ist also hoch wahrscheinlich, dass sich ein wie auch immer formuliertes Ziel irgendwann als unzureichend herausstellt.

Eine weitere Schwierigkeit besteht darin, dass moderne, auf künstlichen neuronalen Netzen basierende KIs inhärent undurchschaubar sind: Warum sie eine bestimmte Entscheidung treffen, lässt sich oft zwar mathematisch, nicht aber in dem menschlichen Denken zugänglichen Begriffen nachvollziehen.[23)] Zwar gibt es Bestrebungen, die Schlussfolgerungsmechanismen neuronaler Netze transparent zu machen, doch sind diese bisher noch nicht umfassend erfolgreich und es ist zu erwarten, dass dieses Problem mit zunehmender Leistungsfähigkeit und Komplexität der KI immer schwieriger wird. Auch Menschen haben oft Probleme, ihre eigenen intuitiven Schlussfolgerungen zu begründen, und greifen dabei manchmal zu falschen, im Nachhinein konstruierten Erklärungen.[24)]

Ein weiteres Problem besteht im Trainingsprozess der KI. Typischerweise werden neuronale Netze anhand von großen Datenmengen darauf trainiert, bestimmte Entscheidungen zu treffen. Diese Daten können jedoch die Komplexität der Wirklichkeit niemals vollständig abbilden. Somit kann es beim praktischen Einsatz von KIs zu Fehlern kommen, die auf einer systematischen Verzerrung aufgrund mangelhafter Trainingsdaten führen, da die Verteilung bestimmter Merkmale in den Daten nicht der Verteilung in der Realität entspricht (Abb. 9). Dieser so genannte Distribution Shift oder auch Dataset Shift[25)] kann auch die gelernte Zielfunktion der KI betreffen und dazu führen, dass die KI sich zwar in der Trainingsumgebung entsprechend den Erwartungen verhält, jedoch in der Realität ein anderes Ziel verfolgt als beabsichtigt.

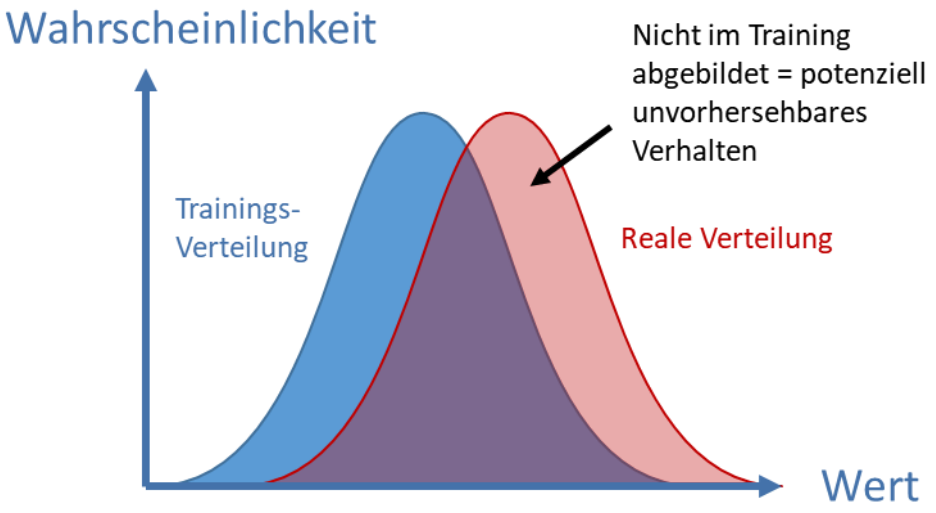

Abb. 9: „Dataset Shift“ zwischen Trainingsdaten und Realität

Nicht zuletzt erweist sich auch die zeitliche Stabilität des Ziels der KI als ein Problem, und zwar in zweifacher Hinsicht: Einerseits ist es möglich, dass sich die Interpretation des Ziels durch die KI im Zeitablauf unvorhersehbar ändert, weil sich ihr Weltmodell geändert hat. Auch Software- und Hardwarefehler, deren Auswirkungen sich im Lauf der Zeit kumulieren, können eine solche Interpretationsveränderung bewirken. Andererseits ist es wahrscheinlich, dass sich die Bedürfnisse der Menschen im Zeitablauf verändern, zum Beispiel weil sich die allgemein akzeptierten menschlichen Werte ändern. So wurde die Sklaverei während des größten Teils der Menschheitsgeschichte von vielen als moralisch akzeptabel angesehen. Auch die Entdeckung neuer Bedürfnisse, wie etwa des geschichtlich noch recht jungen Konzepts der Meinungsfreiheit[26], spielt hier eine Rolle. Die Aktivierung einer unkontrollierbaren KI wäre jedoch ein unumkehrbarer Vorgang und die Anpassung ihres Ziels an sich verändernde menschliche Bedürfnisse wäre aufgrund des instrumentellen Ziels der Zielkonstanz nicht möglich. Daher besteht die Gefahr eines "Value lock-in" (in etwa: Wertezementierung), wie der Philosoph William Macaskill es nennt.[27] Somit ist zu erwarten, dass das Ziel der KI sich im Zeitablauf immer weiter von den menschlichen Bedürfnissen entfernt, selbst wenn es zu Beginn perfekt übereingestimmt haben sollte (Abb. 10).

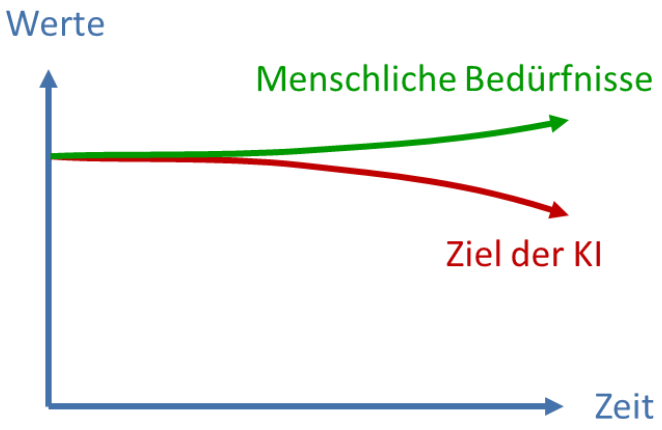

Abb. 10: Zunehmende Divergenz des KI-Ziels und menschlicher Bedürfnisse im Zeitablauf

Ein Lösungsansatz für das Zielproblem besteht darin, die KI dazu zu bringen, sich an menschlichen Entscheidungen zu orientieren.[28] Beispielsweise könnte das Ziel lauten: „Handle stets so, dass du deine eigenen Entscheidungen positiv bewerten würdest, wenn du ein Mensch wärst." Damit könnte ein Teil der oben genannten Probleme womöglich umgangen werden. Jedoch wirft dieser Ansatz neue Probleme auf, denn es ist unklar, wie eine solche Zielfunktion praktisch implementiert werden könnte und wie sichergestellt werden könnte, dass die KI tatsächlich das Ziel verfolgt, das die Menschen gemeint haben, als sie es formulierten.

Ein weiterer Ansatz basiert darauf, der KI keine endgültige Entscheidungsmacht zu geben, sondern stets das Feedback eines Menschen einzuholen.[29)] Wenn dies erfolgreich und konsequent umgesetzt wird, handelt es sich nicht um eine unkontrollierbare KI im Sinne von Hypothese 1. Allerdings weist auch dieser Lösungsansatz erhebliche praktische Probleme auf. Einerseits neigen Menschen dazu, den Entscheidungen der KI blind zu vertrauen, gerade dann, wenn sie diese nicht exakt nachvollziehen können. Andererseits würde eine permanente Überprüfung durch Menschen, sofern diese effektiv und nicht nur stichprobenhaft durchgeführt wird, die Intelligenz und Entscheidungsgeschwindigkeit des Mensch-Maschine-Systems auf die der Menschen reduzieren, der wirtschaftliche Nutzen einer solchen Konstruktion wäre fraglich. Zudem könnte die KI ihre Überwacher auf subtile Weise manipulieren oder täuschen und so unkontrollierbar werden.

Zusammenfassend muss festgestellt werden, dass eine praktikable Lösung des Zielproblems bisher nicht existiert und es unklar ist, ob und wann eine solche Lösung gefunden werden kann. Gelingt dies nicht rechtzeitig, hätte das Entwickeln einer unkontrollierbaren KI mit hoher Wahrscheinlichkeit katastrophale Folgen für die Menschheit. Von entscheidender Bedeutung ist somit die Frage, wie viel Zeit noch bleibt, bevor eine solche KI technisch möglich ist.

Hypothese 4: Eine unkontrollierbare KI könnte bereits vor 2040 technisch möglich sein

In den vergangenen 10 Jahren hat die Entwicklung der KI erstaunliche Fortschritte gemacht. Vieles, was noch kurz zuvor als für KIs unmöglich oder zumindest erst in fernerer Zukunft erreichbar angesehen wurde, ist heute bereits Routine. So erschien beispielsweise noch 2014 im Magazin WIRED ein Artikel, der darlegte, dass das Spiel Go für KIs noch auf längere Zeit nicht auf menschlichem Niveau beherrschbar sei.[30)] Weniger als zwei Jahre später schlug AlphaGo den damals weltbesten Spieler Lee Sedol zum Erstaunen der Fachwelt klar in vier von fünf Partien. Weitere anderthalb Jahre später war AlphaGo Zero in der Lage, sich ohne jegliche Hilfe durch menschliche Experten und ohne die Daten menschlicher Spiele Go selbst beizubringen, und konnte seinen Vorgänger nach nur drei Tagen Trainingszeit und mit geringerer Rechenleistung bereits deutlich schlagen.[31)] Weitere spektakuläre Erfolge waren die Lösung des komplizierten Protein-Folding-Problems, an dem menschliche Experten bereits seit Jahrzehnten arbeiten, durch die KI AlphaFold[32)], die erstaunliche Fähigkeit der KI GPT-3, überzeugende Texte zu generieren[33)], die Fähigkeit von KIs wie DALL-E[34)], Imagen[35)] und Stable Diffusion[36)], auf der Basis von Textbeschreibungen realistische Bilder zu erzeugen, oder die KI AlphaCode[37)], die einfachere Programmieraufgaben auf menschlichem Niveau lösen kann.

Bemerkenswert an diesen Erfolgen ist einerseits, dass sie in der Regel von den Experten weder korrekt vorhergesagt noch erwartet wurden. Andererseits wurden die Leistungen der KIs nicht

in erster Linie durch immer raffiniertere, von Menschen entwickelte Algorithmen und Strategien möglich, sondern schlicht durch die Anwendung einfacher neuronaler Netzarchitekturen auf immer größere Datenmengen mit immer mehr Rechenleistung. So steigerte sich die für das Training verwendete Rechenleistung führender KIs zwischen 2010 und 2022 um mehr als den Faktor eine Milliarde (Abb. 11).[38)]

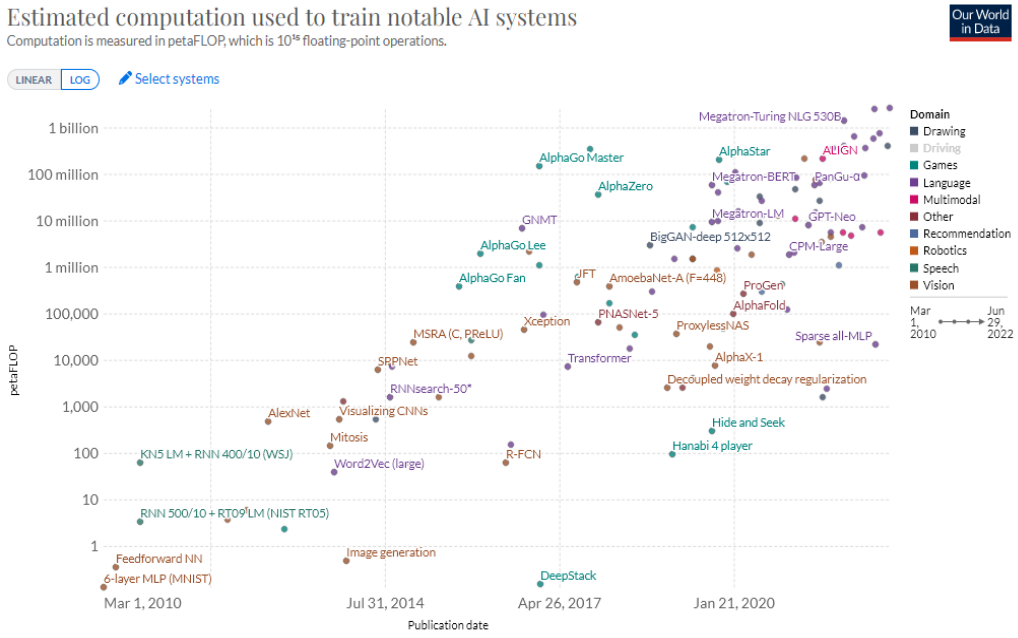

Abb. 11: Entwicklung der für das Training ausgewählter KIs verwendeten Rechenleistung

Dieselbe Grundarchitektur neuronaler Netze kann für unterschiedliche Zwecke eingesetzt werden, indem andere Trainingsdaten verwendet werden. So wurde auf Basis des für die Textgenerierung trainierten Systems GPT-3 eine KI darauf trainiert, Rechen- und Textaufgaben zu lösen.[39)]

Auf Basis dieser offensichtlichen Generalisierungsfähigkeit neuronaler Netze vertreten einige Forscher die Auffassung, „Belohnung sei genug“ – es benötige keine spezifischen Architekturen, sondern lediglich die richtigen Anreize und ausreichend Rechenkapazität und Training, um nahezu alle Intelligenzleistungen künstlich erzeugen zu können.[39)] Daraus leitet sich die „Skalierungshypothese“ ab, die besagt, dass bereits die heute bekannten Methoden zur Entwicklung neuronaler Netze beliebig “skalierbar” sind, also ausreichen, um eine allgemeine künstliche Intelligenz zu entwickeln. Nötig seien lediglich deutlich mehr Rechenleistung und mehr Daten.[41)] Diese Hypothese ist zwar umstritten[42)], aber die rasanten Fortschritte der letzten Zeit scheinen darauf hinzudeuten, dass es kaum noch eine Domäne menschlicher Entscheidungen gibt, die KIs auf absehbare Zeit prinzipiell nicht beherrschen können.

Unklar ist, wie viel Rechenleistung tatsächlich benötigt würde, um eine „starke“ KI mit allgemeiner Problemlösungsfähigkeit auf menschlichem Niveau zu trainieren. Manche Forscher

argumentieren, dass die dafür nötige Rechenleistung wegen des absehbaren Endes des exponentiellen Wachstums der Computerleistung auf lange Sicht nicht verfügbar sein werde. Dagegen spricht, dass die Entwicklung neuer Technologien, wie etwa effizienterer Grafikprozessoren oder Quantencomputern, einen weiteren Entwicklungsschub innerhalb der nächsten 10-20 Jahre plausibel erscheinen lässt.[43] Zudem sind auch weitere Fortschritte bei der Entwicklung effizienterer Strukturen neuronaler Netze zu erwarten.

Vor diesem Hintergrund schätzen Experten die Zeit, bis eine starke KI möglich wird, sehr unterschiedlich ein. In verschiedenen Befragungen ergab sich ein Mittelwert der Erwartungen um die Mitte des Jahrhunderts (Abb. 12, 13) mit einem unteren Quartil um die Mitte des nächsten Jahrzehnts.

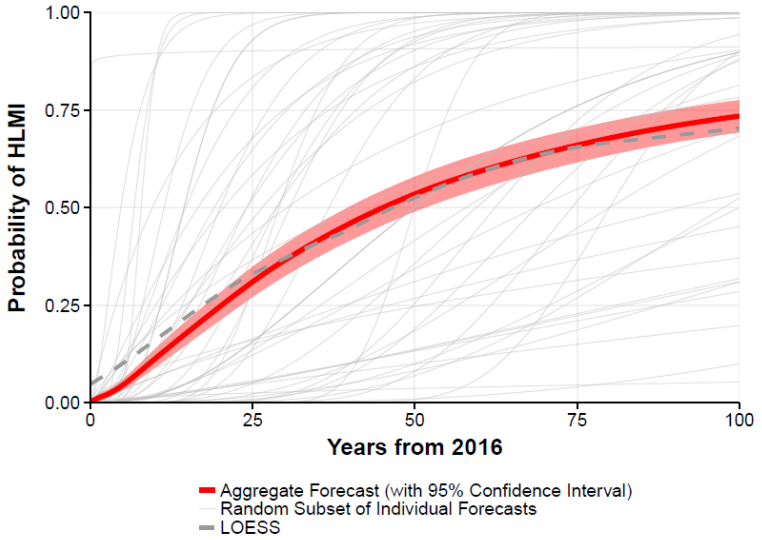

Abb. 12: Im Jahr 2015 durchgeführte Befragung von Experten zur Erwartung des Zeitpunkts „starker KI“[44]

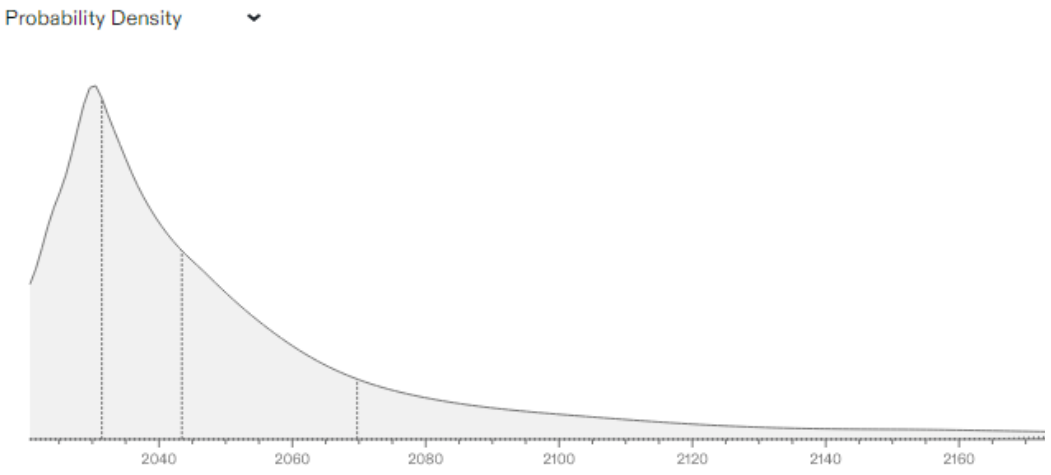

Abb. 13: Einschätzung von Nutzern der Prognoseplattform Metaculus zum erwarteten Zeitpunkt „starker KI“, abgerufen am 15.8.2022[45]

Zu einem ähnlichen Ergebnis kommt eine Abschätzung auf Basis biologischer Vergleichswerte, etwa der Komplexität des menschlichen Gehirns (Abb. 14).

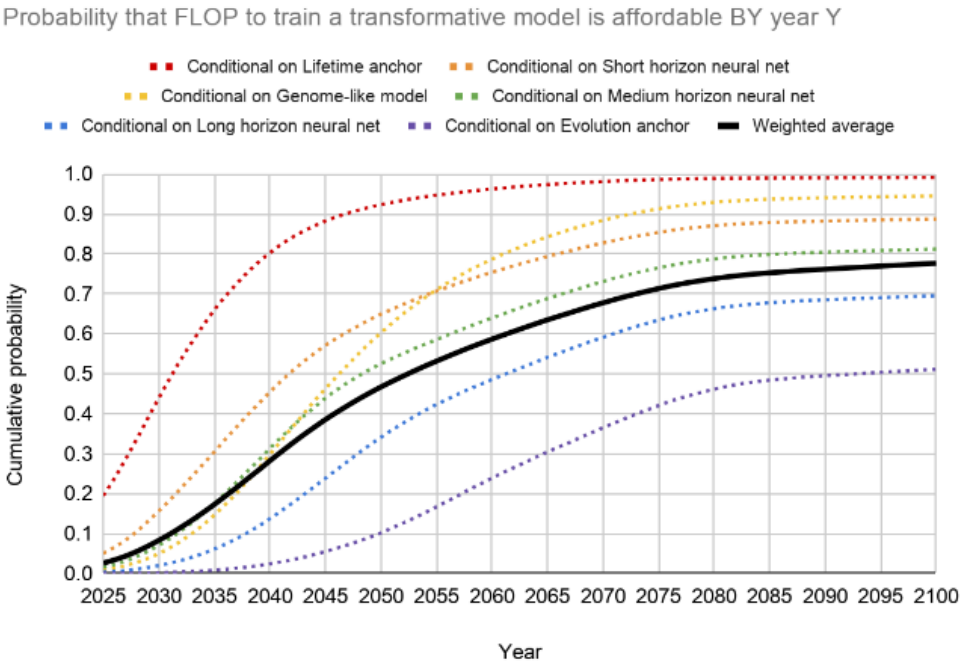

Abb. 14: Abschätzung des Zeitpunkts starker KI auf Basis verschiedener biologischer Vergleichswerte[46)]

Bemerkenswert an den Befragungen ist die sehr große Streuung der Meinungen. Während einige Forscher die Entwicklung starker KI für gar nicht möglich oder erst im nächsten Jahrhundert realistisch halten, sehen andere es als wahrscheinlich an, dass dies noch innerhalb des aktuellen Jahrzehnts gelingen könnte (Abb. 15).[47)]

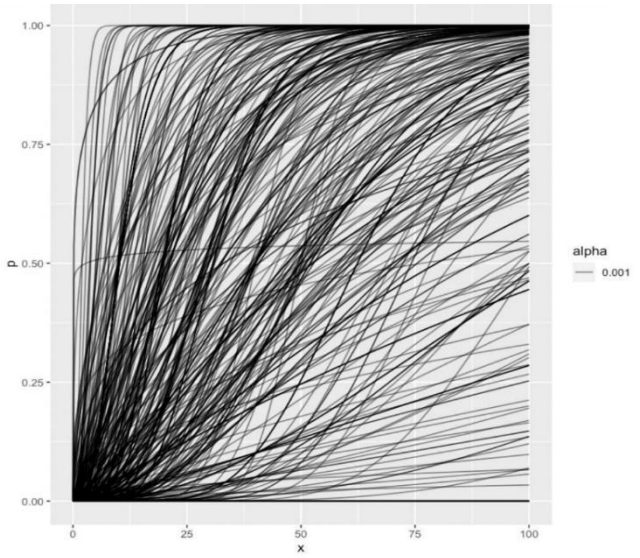

Abb. 15: Einschätzung verschiedener Experten des Wahrscheinlichkeitsverlaufs für starke KI in Jahren ab 2022

Aus dieser breiten Streuung der Einschätzungen kann vor allem der Schluss gezogen werden, dass eine hohe Unsicherheit darüber besteht, wann eine starke KI technisch möglich sein könnte. Im Durchschnitt kommen die Experten zu dem Schluss, dass dies mit etwa 25% Wahrscheinlichkeit noch vor 2040 eintreffen wird. Wie in Hypothese 1 dargelegt, bildet diese Betrachtung eher eine Untergrenze für die Wahrscheinlichkeit unkontrollierbarer KI zu einem bestimmten Zeitpunkt, da diese nicht in jeder Hinsicht einer „starken“ KI entsprechen muss.

Angesichts des in Hypothese 2 dargestellten existenziellen Risikos und der entsprechend Hypothese 3 eher geringen Wahrscheinlichkeit, dass bis 2040 eine sichere Lösung des Zielproblems gefunden werden kann, ergibt sich somit eine höchst kritische Situation. Auch eine scheinbar niedrige Wahrscheinlichkeit der Möglichkeit unkontrollierbarer KI von 25% vor 2040 ist angesichts eines existenziellen Risikos für die Menschheit inakzeptabel. Unter Risikogesichtspunkten wäre es zudem fahrlässig, sich darauf zu verlassen, dass eine unkontrollierbare KI erst in fernerer Zukunft entwickelt werden kann (Abb. 16). Im Gegenteil muss sicherheitshalber vom „schlimmsten Fall“ einer Entwicklung schon in naher Zukunft ausgegangen werden.

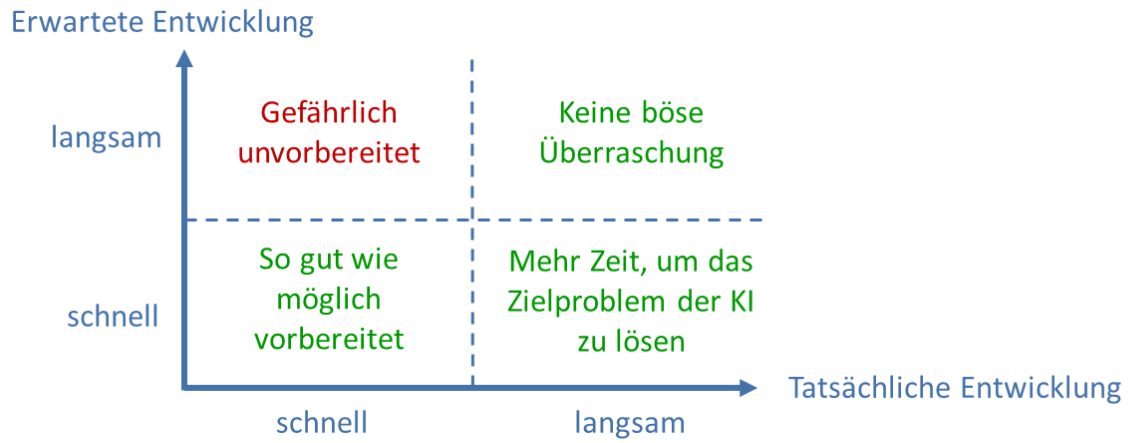

Abb. 16: Risikoabwägung einer schnellen und langsamen Einschätzung der Entwicklung unkontrollierbarer KI

Zusammenfassend muss festgestellt werden, dass das Eintreten einer existenziellen Katastrophe durch unkontrollierbare KI vor dem Jahr 2040 nach gegenwärtigem Wissensstand nicht mit hinreichender Sicherheit ausgeschlossen werden kann. Im Gegenteil ist eine solche Katastrophe so wahrscheinlich und ihre potenziellen Auswirkungen so weitreichend, dass sie als konkrete existenzielle Bedrohung für die Menschheit angesehen werden muss.

Schlussfolgerungen

Vor dem Hintergrund der obigen Ausführungen wäre es zu erwarten, dass EU-weit mit Hochdruck an einer Vermeidung des existenziellen Risikos durch unkontrollierbare KI gearbeitet wird. Mindestens müsste es Forschungsprojekte geben, die dieses Risiko angesichts der hohen damit verbundenen Unsicherheiten genauer analysieren. Dies ist jedoch nicht der Fall. Anfang September 2022 ist dem Verfasser kein einziges Projekt einer öffentlichen oder privaten Forschungseinrichtung innerhalb der EU bekannt, das die existenziellen Risiken unkontrollierbarer oder starker KI genauer erforscht, geschweige denn versucht, Lösungsansätze zu entwickeln. Zudem gibt es in der EU, anders als in den USA und Großbritannien, keine Forschungseinrichtung, die sich explizit mit existenziellen Risiken durch fortschrittliche KI beschäftigt. Das ist vor allem deshalb inakzeptabel, weil die EU mit dem geplanten „AI Act“ eine Vorreiterrolle bei der Regulierung der KI-Entwicklung zu übernehmen versucht.[48)] Wie dies ohne solide wissenschaftliche Grundlage gelingen soll, ist unklar.

Es wäre somit dringend geboten,

- unverzüglich mindestens ein Forschungsprojekt zur näheren Analyse der hier dargestellten Risiken durchzuführen mit dem Ziel, die genannten Hypothesen entweder zu widerlegen oder zu bestätigen und ihre Auswirkungen genauer zu quantifizieren sowie

- auf dieser Basis Vorschläge zu erarbeiten, wie die Risiken unkontrollierbarer KI eingedämmt werden können. Derartige Vorschläge könnten zum Beispiel Folgendes umfassen:
    - einen massiven Ausbau der Forschung zur Lösung des Zielproblems innerhalb Deutschlands und der EU
    - Grenzen und „rote Linien“ für Forschung und Entwicklung, die nicht überschritten werden dürfen, um unkontrollierbare KI sicher auszuschließen
    - die Ausarbeitung und Verabschiedung von verbindlichen Regeln für die Erforschung und Entwicklung von KI zur Vermeidung dieses Risikos (in Erweiterung des in dieser Hinsicht sehr unkonkreten geplanten AI Acts)
    - das Schließen internationaler Vereinbarungen mit dem Ziel, ein „Wettrennen“ um die Entwicklung starker KI zu vermeiden und die zuvor genannten Regeln auch international anzuwenden
    - die Schaffung einer oder mehrerer unabhängiger Institutionen, deren Aufgabe die permanente Beobachtung und Analyse der Entwicklung hochleistungsfähiger KI und ihrer Auswirkungen auf die Gesellschaft wäre.

Es muss betont werden, dass das Risiko einer unkontrollierbaren KI davon unabhängig ist, wer diese entwickelt. Auch eine in bester Absicht entwickelte KI, die außer Kontrolle gerät, kann schnell zu einer existenziellen Katastrophe führen, solange das Zielproblem nicht gelöst ist. Die gelegentlich geäußerte Meinung, es sei besser, wenn ein westliches KI-Labor zuerst eine starke KI entwickele, da andere Länder weniger vorsichtig vorgingen, ist irreführend und führt nur zu einer Verschärfung der Wettbewerbssituation, die wiederum zu einer Beschleunigung der Entwicklung und Vernachlässigung der Sicherheitsaspekte führt. Letztlich würde eine durch eine unkontrollierbare KI ausgelöste existenzielle Katastrophe alle Menschen in gleichem Maße betreffen. In einem solchen „Rennen“ kann es folglich keinen Gewinner geben.

Die einzige Möglichkeit, das Risiko unkontrollierbarer KI zu vermeiden, besteht darin, deren Entwicklung zu verhindern, zumindest bis eine sichere Lösung des Zielproblems gefunden wird. Die beste Basis dafür ist ein allgemeiner, globaler Konsens in Bezug auf dieses Risiko. Es ist dringend geboten, die Forschung zu intensivieren, um diesen Konsens schnellstmöglich zu schaffen.

Anhang: Häufige Einwände

Im Folgenden werden tabellarisch häufig vorgebrachte Einwände gegen die genannten Hypothesen und mögliche Erwiderungen darauf in knapper Form dargestellt.

Einwand	Erwiderung
Die Leute hatten schon immer Angst vor neuer Technologie. Doch die war immer unbegründet .	Es gibt viele gefährliche Technologien, wie z.B. Atombomben. Sie müssen mit äußerster Vorsicht behandelt werden. Klimawandel und Umweltzerstörung zeigen zudem die negativen Folgen moderner Technik. Es wäre vor diesem Hintergrund fahrlässig, die Risiken unkontrollierbarer KI zu ignorieren.
Wir können eine KI doch einfach abschalten.	Eine unkontrollierbare KI wird Wege finden, ihre Abschaltung zu verhindern, z.B. indem sie die Menschen davon überzeugt, dass sie harmlos ist. Auch Menschen können ja im Prinzip "abgeschaltet" (getötet) werden, doch Diktatoren finden immer wieder Wege, dies zu verhindern.
KI wird niemals so intelligent sein wie ein Mensch.	KI ist Menschen bereits in vielen Bereichen überlegen. Es gibt keine Hinweise auf Probleme, die Menschen lösen können, eine KI aber prinzipiell niemals lösen wird. Außerdem kann auch eine KI, die Menschen in mancher Hinsicht unterlegen ist, unkontrollierbar sein.
KI wird niemals ein Bewusstsein haben.	Unkontrollierbare KI braucht kein „Ich-Bewusstsein". Sie muss lediglich gemäß einem vordefinierten Ziel handeln und dabei ihre eigene Rolle bei der Erfüllung dieses Ziels berücksichtigen.
Eine KI wird niemals einen eigenen Willen haben.	Eine KI wird zwar das Ziel verfolgen, das ihre Entwickler ihr geben, aber ihre eigene Lösungsstrategie entwickeln, die den menschlichen Interessen zuwiderlaufen könnte.
Starke KI wird von selbst das Richtige tun.	Das Prinzip der Orthogonalität von Ziel und Intelligenz besagt, dass auch eine intelligente KI rücksichtslos ein „dummes", zerstörerisches Ziel verfolgen kann.
Wir müssen bloß den Zugriff der KI auf die Außenwelt beschränken.	Eine unkontrollierbare KI wird Wege finden, diese Beschränkungen entweder direkt (z.B. durch „Hacken") oder indirekt (durch Manipulation von Menschen) zu umgehen.
Es wird noch sehr lange dauern, bis unkontrollierbare KI zum Problem wird.	Es ist unklar, wie lange es noch dauern wird, bis unkontrollierbare KI möglich wird. Einige Experten halten dies noch vor 2040 für wahrscheinlich. Im Sinne einer Risikovorsorge wäre es fahrlässig, sich darauf zu verlassen, dass es noch sehr lange dauert.
Wir werden das Zielproblem der KI rechtzeitig lösen.	Die Experten sind sich einig, dass dieses Problem sehr schwierig und noch lange nicht gelöst ist. Niemand weiß, wie lange es noch dauern wird und ob es überhaupt eine praktikable Lösung gibt.
Wir brauchen doch bloß einer KI das Ziel zu geben, das Zielproblem zu lösen.	Um das Zielproblem der KI zu lösen, müsste eine KI bereits mindestens so intelligent sein wie die KI, deren Ziel sie formulieren soll. Denn sonst kann sie nicht sicher wissen, wie die überlegene KI das Ziel interpretiert. Daher funktioniert dieser Ansatz nicht.
Niemand ist so dumm, unkontrollierbare KI zu entwickeln.	Unkontrollierbare KI könnte aus Versehen, aufgrund von Unachtsamkeit oder Überheblichkeit entstehen.
Wir haben dringendere Probleme zu lösen.	Angesichts der existenziellen Dimension und der zeitlichen Dringlichkeit ist unkontrollierbare KI ein besonders vernachlässigtes Problem mit potenziell sehr weitreichenden Auswirkungen auf die Zukunft der Menschheit.

Quellen und Literatur

1) https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-5
2) Wiener, Norbert: Some Moral and Technical Consequences of Automation, Science, 1960, https://www.science.org/doi/10.1126/science.131.3410.1355
3) Good, Irving John: Speculations Concerning the First Ultraintelligent Machine, Advances in Computers Vol. 6, 1965
4) Bostrom, Nick: Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014
5) Russell, Stuart: Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019
6) Tegmark, Max: Life 3.0: Being Human in the Age of Artificial Intelligence, Allen Lane 2017
7) Eine Abschätzung von Benjamin Hilton auf Basis einer Übersicht von AI Watch kommt zu einem Ergebnis von ca. 300 Personen, die an der Sicherheit fortschrittlicher KI arbeiten, die meisten davon jedoch in Unternehmen, die selbst KI entwickeln. Die Zahl der Forscher an Universitäten und unabhängigen Instituten dürfte deutlich unter 100 liegen. Vgl. Hilton, Benjamin: Preventing an AI-related catastrophe, Online-Artikel auf der Plattform 80.000 hours, August 2022, https://80000hours.org/problem-profiles/artificial-intelligence/
8) Ord, Toby: The Precipice: Existential Risk and the Future of Humanity, Hachette Books 2020
9) https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai/
10) Für einige Beispiele siehe Walker, Brandon: The Games That AI Won And The Progress They Represent, Online-Artikel, März 2020, https://towardsdatascience.com/the-games-that-ai-won-ff8fd4a71efc
11) Carlsmith, Joseph: Is power-seeking AI an existential risk?, Online-Dokument, April 2021, https://docs.google.com/document/d/1smaI1lagHHcrhoi6ohdq3TYIZv0eNWWZMPEy8C8byYg,
12) Bostrom, Nick: The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents, Minds and Machines, Vol. 22, Iss. 2, May 2012, https://nickbostrom.com/superintelligentwill.pdf
13) Eine umfassende Darstellung des Konzepts einer „Intelligenzexplosion“ durch rekursive Selbstverbesserung findet sich auf der Website des Machine Intelligence Research Institute https://intelligence.org/ie-faq/
14) Eine plastische Schilderung eines möglichen Szenarios liefert Paul Christiano in einem Forumsbeitrag: https://www.alignmentforum.org/posts/AyNHoTWWAJ5eb99ji/another-outer-alignment-failure-story

15) Bostrom, Nick: The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents, Minds and Machines, Vol. 22, Iss. 2, May 2012, https://nickbostrom.com/superintelligentwill.pdf
16) Bostrom, Nick: Ethical Issues in Advanced Artificial Intelligence, Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, https://nickbostrom.com/ethics/ai
17) Russell, Stuart: Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019, S.139
18) Siehe z.B. Christian, Brian: The Alignment Problem: Machine Learning and Human Values, W. W. Norton & Company, 2020
19) Einige Beispiele listet die Firma Deepmind in ihrem Blog auf: https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity
20) Quelle: Climate.gov https://www.climate.gov/media/12885
21) Siehe z.B. die Liste offener Probleme im Alignment Forum: https://www.alignmentforum.org/posts/5HtDzRAk7ePWsiL2L/open-problems-in-ai-x-risk-pais-5
22) Yampolskiy, Roman V.: On the Controllability of Artificial Intelligence: An Analysis of Limitations, Journal of Cyber Security and Mobility, Vol. 11, Issue 3, May 2022, https://journals.riverpublishers.com/index.php/JCSANDM/article/view/16219
23) Gilpin et al.: Explaining Explanations: An Overview of Interpretability of Machine Learning, The 5th IEEE International Conference on Data Science and Advanced Analytics 2018, https://arxiv.org/abs/1806.00069
24) Vgl. den Begriff „Rationalisierung“ in der Psychologie, z.B. https://www.psymag.de/11493/rationalisierung-psychoanalyse-neurose-hypnose-gruende/2/
25) Candela et al.: Dataset Shift, in: Dietterich, Thomas (Hrsg.): Adaptive Computation and Machine Learning, The MIT Press, Cambridge, Massachusetts 2008, https://cs.nyu.edu/~roweis/papers/invar-chapter.pdf
26) Issel, Konstantin: Eine kurze Geschichte der Meinungsfreiheit, Online-Artikel, März 2021, https://www.die-debatte.org/debattenkultur-geschichte-der-meinungsfreiheit/
27) Macaskill, William: What We Owe the Future, Oneworld Publications, 2022, S.83-86.
28) Russell, Stuart: Human Compatible: Artificial Intelligence and the Problem of Control, Viking, 2019, S. 171 ff.
29) Clifton, Jesse: Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda, Section 6: Humans in the Loop, Beitrag im AI Alignment Forum, Dezember 2019,

https://www.alignmentforum.org/posts/4GuKi9wKYnthr8QP9/sections-5-and-6-contemporary-architectures-humans-in-the#6_Humans_in_the_loop__6_

30) Levinovitz, Alan: The Mystery of Go, the Ancient Game That Computers Still Can't Win, in: Wired, Mai 2014, https://www.wired.com/2014/05/the-world-of-computer-go/
31) Beitrag im Deepmind Blog vom 18.10.2017, https://www.deepmind.com/blog/alphago-zero-starting-from-scratch
32) Ein Überblick über die Entwicklungsschritte von AlphaFold findet sich auf der Deepmind-Website: https://www.deepmind.com/research/highlighted-research/alphafold/timeline-of-a-breakthrough
33) Brown et al.: Language Models are Few-Shot Learners, Arxiv 2020, https://arxiv.org/abs/2005.14165
34) Einige eindrucksvolle Beispiele finden sich auf der OpenAI website: https://openai.com/dall-e-2/ Der Verfasser konnte DALL-E2 im Rahmen des Betatests testen und hat einige selbst erstellte Beispiele auf seinem Blog veröffentlicht: https://www.ki-risiken.de/2022/06/10/von-fr%C3%B6schen-und-schriftstellern/
35) Saharia et al.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, Arxiv Mai 2022, https://arxiv.org/abs/2205.11487
36) Siehe Blogbeitrag von Stability AI: https://stability.ai/blog/stable-diffusion-public-release
37) Li et al.: Competition-Level Code Generation with AlphaCode, Arxiv Februar 2022, https://arxiv.org/abs/2203.07814
38) Quelle: https://ourworldindata.org/grapher/ai-training-computation
39) Drori et al.: A Neural Network Solves, Explains, and Generates University Math Problems by Program Synthesis and Few-Shot Learning at Human Level, Arxiv Juni 2022, https://arxiv.org/abs/2112.15594
40) Silver et al.: Reward is enough, in: Artificial Intelligence, Vol. 299, Oktober 2021, https://www.sciencedirect.com/science/article/pii/S0004370221000862
41) Eine gute Einführung in die Skalierungs-Hypothese gibt der Blogger Gwern: https://www.gwern.net/Scaling-hypothesis
42) Vamplev et al.: Reward is not enough: A response to Silver, Singh, Precup and Sutton (2021), Arxiv, November 2021, https://arxiv.org/abs/2112.15422
43) Siehe z.B. McKinsey & Company: Quantum computing: An emerging ecosystem and industry use cases, Special Report, Dezember 2021, https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/quantum-computing-use-cases-are-getting-real-what-you-need-to-know

44) Grace et al.: Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts, Journal of Artificial Intelligence Research 62, Juli 2018, https://www.researchgate.net/publication/326725184_Viewpoint_When_Will_AI_Exceed_Human_Performance_Evidence_from_AI_Experts

45) Eine tagesaktuelle Prognose lässt sich auf der Onlineplattform Metaculus abrufen: https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/

46) Cotra, Ajeya: Forecasting TAI with biological anchors, Part 4: Timelines estimates and responses to objections, Online-Dokument, Juli 2020, https://drive.google.com/drive/u/1/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP

47) AI Impacts: 2022 Expert Survey on Progress in AI, Online-Dokument, August 2022, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/

48) Siehe z.B. die Website der Europäischen Kommission: https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence