Unregulierte Zweckrationalität

Es ist kein Zufall, dass KI-Algorithmen gerne an Spielen wie Schach oder Go erprobt werden, bevor man sie auf ernsthafte Probleme anwendet. In einem Spiel ist klar definiert, was es heißt zu gewinnen oder zu verlieren. Die Regeln des Spiels geben das Ziel vor und die Aufgabe des Spielers ist es, dieses zu erreichen. Bei vielen Spielen geht es darum, möglichst viele Punkte zu sammeln. Kosten und Nutzen jedes Spielzuges kann man in diesen Fällen theoretisch ausrechnen. Mit KI-Methoden lassen sich vorgegebene Ziele erreichen und die Kosten minimieren – oder der Nutzen maximieren, was auf das gleiche hinausläuft.[1] Damit zum Beispiel verstärkendes Lernen eingesetzt werden kann, müssen die Kosten aber zunächst festgelegt werden. In der wirklichen Welt ist das oft selbst schon ein schwieriges Problem. Eine große politische Aufgabe ist zurzeit: Wir wollen die Energiewende schaffen. Aber was heißt das genau? Wir wissen, dass wir unseren CO_2-Ausstoß schnell verringern müssen. Aber was ist eine Tonne CO_2 im Vergleich zu einem Arbeitsplatz in der Schwerindustrie wert? Im Gegensatz zu Spielen sind uns in der wirklichen Welt die Ziele und Kosten nicht vorgegeben, sondern wir müssen sie selber bestimmen.

Leider sind wir nicht besonders gut darin. Wir schaffen zum Beispiel oft Fehlanreize, die zu unbeabsichtigten Auswirkungen führen: Man spricht auch vom Kobra-Effekt. Nehmen wir an, bei uns in der Gegend gibt es zu viele Kobras. Als politische Maßnahme bezahlen wir jedem, der uns eine tote Kobra bringt, eine Prämie. Das soll den Leuten einen Anreiz geben, Kobras zu jagen und so ihre Population verringern. Stattdessen fangen die Leute an, Kobras zu züchten. Als wir das merken, schaffen wir die Prämie wieder ab und die Leute entlassen all ihre

1 Kosten und Nutzen unterscheiden sich mathematisch nur im Vorzeichen: Kosten haben einen negativen Nutzen.

Kobras in die Wildbahn. Am Ende haben wir sogar das Gegenteil von dem bewirkt, was wir erreichen wollten. Die Wirtschaftswissenschaft kennt viele solche Beispiele, in denen Anreize, die zunächst vernünftig klingen, ungewollte Folgen haben.[2]

Der Kobra-Effekt begleitet auch den Einsatz von KI. Entwickler definieren Kosten, von denen sie glauben, dass das KI-System dann macht, was sie wollen. Am Ende minimiert das System zwar die vorgegebenen Kosten, macht aber trotzdem nicht, was es soll. Einem KI-System, das die Energiewende planen soll, geben wir zum Beispiel extrem hohe Preise für jede ausgestoßene Tonne CO_2 vor, damit möglichst viel CO_2 eingespart wird. Das System schlägt dann – logischerweise – vor, dass wir einfach unsere ganze Industrie abschalten und gar keine Energie mehr verbrauchen. Diese Lösung gefällt uns nicht. Also passen wir die Kosten so an, dass es sehr teuer ist, wenn der Energiebedarf nicht gedeckt wird. Was ja auch stimmt. Wir hatten es nur übersehen. Daraufhin schlägt das System vor, sofort aus der Kohle auszusteigen, weil die teuer ist und viel CO_2 produziert. Uns fällt aber auf, dass das zu mehr Arbeitslosigkeit in den Kohleabbaugebieten führen wird und diese sozialen Kosten bisher nicht berücksichtigt wurden, und so weiter. Bei der Anwendung von KI-Methoden in der Praxis sieht man häufig, dass die Kosten, die ein System minimieren soll, während der Entwicklung immer wieder angepasst werden, weil unbeabsichtigte Nebenwirkungen auftreten. Am Ende weiß man nie, ob noch etwas vergessen wurde.

Macht der Computer, was wir wollen?

Es ist also gar nicht so einfach sicherzustellen, dass ein KI-System das macht, was wir wollen. In der KI-Forschung wird das als ›Alignment-Problem‹ bezeichnet, was man als Ausrichtungsproblem übersetzen kann. Im Kapitel über Suchalgorithmen hatten wir bereits gesehen, dass zum Beispiel beim Planen einer Bahnreise die schnellste Route nicht immer die ist, die ein Nutzer möchte. Manchmal möchte derselbe Nutzer den billigsten Preis und ein anderes Mal hat er keine Lust umzusteigen. Das hängt von den jeweiligen Umständen ab – und ein Sys-

2 *Der Kobra-Effekt* ist der Titel eines Buches von Siebert (2001). Historisch besser belegt als Prämien für tote Kobras sind die sogenannten Schwanzprämien für tote Mäuse und Ratten, die früher an vielen Orten gezahlt wurden.

tem, das die Wünsche des Nutzers nicht erfüllt, ist nicht richtig ausgerichtet. Zur Verteidigung der Entwickler muss man sagen, dass es auch nicht leicht ist herauszufinden, was die Nutzer eigentlich wollen. Und dann wollen unterschiedliche Nutzer auch noch unterschiedliche Dinge.

Gibt es dafür keine Lösung? Doch! Die Lösung ist natürlich KI. Wäre es nicht viel praktischer, wenn nicht mehr die Entwickler eines KI-Systems diejenigen wären, die die Vorlieben der Nutzer spezifizieren, sondern wenn der Computer selber herauszufinden könnte, was die Nutzer wollen, um sich dann entsprechend zu verhalten? Ein KI-System kann zum Beispiel die Vorlieben des Nutzers aus seinem Verhalten ableiten. Einem Nutzer, der die Zugfahrt wählt, die eine halbe Stunde schneller ist, aber zwanzig Euro mehr kostet, waren in dieser konkreten Situation die halbe Stunde mindestens zwanzig Euro wert. Das Erschließen der Vorlieben aus dem Verhalten funktioniert auch, wenn der Nutzer selber gar nicht so richtig weiß, was er eigentlich möchte. Oder wenn seine Vorlieben sich über die Zeit ändern. Indem sich Computer automatisch an die Wünsche der Nutzer anpassen, kann das Alignment-Problem gelöst werden. Das ist jetzt das neue Ziel der KI-Forschung.[3]

Es klingt zunächst verlockend, dass wir in der Zukunft KI-Systeme haben könnten, die viele Aufgaben für uns übernehmen und uns alle Wünsche von den Lippen ablesen. Aber wollen wir das wirklich? Dieser Ansatz mag für ein System, das Reisen planen soll, vielleicht akzeptabel sein. Bevor ich losfahre, kann ich kontrollieren, ob mir der Vorschlag gefällt, und selbst wenn ich mir keine Alternativen ansehe, sondern einfach dem erstbesten Vorschlag folge, ist ein kleiner Umweg auch nicht schlimm. Weniger akzeptabel ist hingegen, dass ein KI-System meine Anlagestrategie plant und erst, nachdem ich eine große Menge an Geld verloren habe, merkt, dass ich eigentlich eher risikoscheu bin. Ich müsste mir schon sehr sicher sein, dass das KI-System mich gut kennt, bevor ich es selbständig meine Anlagen verwalten lasse. Aber es gibt auch Bereiche, in denen mich Algorithmen bereits viel zu gut kennen. Der Algorithmus, der mir auf Instagram das nächste Kurzvideo vorschlägt, passt sich an meine Wünsche an, indem er mein Verhalten genau analysiert. Welche Videos schaue ich an und welche nicht. Das Ergebnis ist, dass ich das Handy gar nicht mehr weglege. Und ein Kühl-

3 Russell (2019) hat ein ganzes Buch darübergeschrieben.

schrank, der automatisch für mich Lebensmittel nachbestellt, wird schnell merken, dass ich oft Tiefkühlpizza esse. So eine KI wird es mir noch schwerer machen, mich ausgewogen zu ernähren.

Solche Zielkonflikte sind die Regel, nicht die Ausnahme. Wir wollen CO_2 reduzieren, aber auch billig in den Urlaub fliegen. Um die Welt zu retten, verzichtet manch einer daher bewusst auf seinen Traumurlaub auf Hawaii und fährt stattdessen in den Harz zum Wandern. Die Reise-KI sollte uns nicht einfach einen Flug buchen, sondern uns die Möglichkeit lassen, uns bewusst anders zu entscheiden. Die Abwägung zwischen Traumurlaub, Preis und Flugscham will ich schon selber vornehmen. Aber die Vorschläge, die das System – ganz uneigennützig? – macht, sind so verlockend … Ich bleibe stark und buche ein Zugticket nach Wernigerode.

Der Kohlekumpel in der Lausitz kann hingegen nicht so leicht auf seinen Arbeitsplatz verzichten wie ich auf meinen Urlaubsflug. Unterschiedliche Menschen haben unterschiedliche Interessen. An wessen Interessen soll ein KI-System, das die Energiewende für uns planen soll, ausgerichtet sein? An denen des Kohlekumpels, der Schwerindustrie, der Fluggesellschaften oder der Schülerinnen und Schüler von Fridays for Future? Wie kann ein Kompromiss zwischen all diesen berechtigten Interessen aussehen?

Spätestens an dieser Stelle dürfte klar geworden sein, dass KI keine wertneutrale Technologie ist, deren Entwicklung die Gesellschaft (überwiegend männlichen) Forschern, Ingenieuren und Unternehmern überlassen kann.[4] Wir müssen gemeinsam in einem demokratischen Prozess entscheiden, welche Werte KI-Systemen einprogrammiert werden sollen. Im Silicon Valley, wo viele der bekanntesten KI-Systeme entwickelt werden, verbreiten sich allerdings im Dunstkreis von etablierten Tech-Unternehmen und KI-Startups befremdliche politische und moralische Vorstellungen. Ich sehe an meinen Studierenden, welche enorme Anziehungskraft diese Vorstellungen auch hierzulande auf technikbegeisterte Menschen ausüben. Diese Faszination basiert auf dem Glauben, dass sich alle Probleme der Menschheit mit Technologie lösen lassen. Da KI die ultimative Problemlösetechnologie ist, wird sie uns von allen Übeln befreien. Leider wird dabei übersehen, dass nicht alle Probleme rein technischer Natur sind. Die Technologiegläubigkeit von Elon Musk und anderen selbsternannten KI-Aposteln im Silicon

4 Der Terminus technicus ist ›Tech-Bros‹.

Valley nimmt teilweise groteske Züge an.[5] Da viele von ihnen aber äußerst erfolgreich sind und vor allem politisch immer einflussreicher werden, muss man sich leider ausführlich mit ihren Ideen auseinandersetzen.[6] Ein gewisses Maß an Polemik kann ich mir dabei allerdings nicht verkneifen.

Der Markt wird das schon regeln

Der ehemalige Internetpionier Marc Andreessen (die Älteren unter uns erinnern sich vielleicht noch an seinen Netscape Browser) ist heute ein äußerst erfolgreicher Wagniskapitalgeber, der sich selber als »Techno-Optimisten« bezeichnet. Er glaubt nicht, dass wir bei der Energiewende nach politischen Kompromissen suchen müssen, weil es der Markt sein wird, der den Ausgleich von selbst regelt. KI wird dem Einzelnen (insbesondere natürlich den einzelnen Unternehmen, die die KI-Ressourcen kontrollieren) helfen, sich besser im Markt zu behaupten. Wenn jeder an sich selber denkt, ist an jeden gedacht. So lautet Andreessens Glaubensgrundsatz. Dazu gehört auch, dass wir angeblich mittels KI das Ideal eines Marktes erreichen, in dem sich alle Akteure vollständig rational verhalten. Da der Wettbewerb dafür sorge, dass jeder bekommt, was er verdient, gibt es auch kein Alignment-Problem. Kobra-Effekte entstehen nur, wenn sich jemand einbildet, er wüsste es besser als der Markt und die falschen Anreize setzt. Da der Markt dank KI aber immer rationaler wird und der Markt immer bessere KI hervorbringt, werde dieser selbstverstärkende Mechanismus uns schnell zu großem materiellen Überfluss und superintelligenter KI führen. Aus reinem Eigeninteresse werden wir so auch die Klimakrise lösen und dank der superintelligenten KI werden wir das auch leicht hinbekommen.

Es verwundert nicht, dass ein Wagniskapitalgeber aus dem Silicon Valley an freie Märkte und technologische Innovation glaubt. Schließlich ist Andreessen einer der Superreichen, die das techno-kapita-

5 Siehe z.B. den scharfzüngigen und trefflichen Artikel von Meckel (2023) über Marc Andreessen, über den wir gleich noch reden werden.

6 Als ich diese Zeilen ursprünglich geschrieben hatte, konnte ich noch nicht ahnen, welche Rolle Elon Musk und andere Tech-Milliardäre in der zweiten Amtszeit von Donald Trump seit Januar 2025 spielen würden.

listische Spiel gewonnen haben. Der religiöse Eifer, mit dem er seine Positionen vertritt, ist allerdings bemerkenswert. Sein bizarres *Techno-Optimist Manifesto* ist ein pathetisches Glaubensbekenntnis für freie Märkte, unregulierte Technologie und grenzenloses Wachstum ohne Rücksicht auf Verluste.[7] Er predigt, auf Nachhaltigkeit, soziale Verantwortung, das Vorsorgeprinzip, Risikomanagement oder Technikethik vollständig zu verzichten. So rast Andreessen in seinem futuristischen Sportwagen immer schneller dem gelobten Land entgegen, in dem angeblich Milch und Honig für alle fließen. Bedenkenträger, die KI oder Märkte regulieren wollen, bremsen ihn auf der Straße des Fortschritts nur aus.

Auch wenn ihm nicht jeder im Silicon Valley dieses höchst eigennützige Heilsversprechen abkauft, so zweifeln wahrscheinlich nur Wenige daran, dass technologische Innovation gut ist und freie Märkte ein ausgezeichnetes Instrument für Fortschritt sind. In Deutschland setzen wir im Fall der Energiewende darauf, dass die erneuerbaren Energien durch marktwirtschaftlichen Wettbewerb möglichst schnell besser und billiger werden. Das heißt aber nicht, dass Regulierung keine Rolle spielt. Denn bei uns hat die Regulierung durch das Erneuerbare-Energien-Gesetz sogar dazu beigetragen, dass es mit der Energiewende schneller voranging.[8]

In der Europäischen Union ist die Regulierung von KI ohnehin nicht mehr aufzuhalten. Die KI-Verordnung ist seit August 2024 in Kraft, egal, ob es den Techno-Optimisten gefällt oder nicht. Diese Regulierung ist wichtig: Wir erinnern uns an die Enthüllungen von Edward Snowden zur flächendeckenden Überwachung durch Tech-Unternehmen und Geheimdienste sowie an den Skandal um Facebook und Cambridge Analytica zu Wahlbeeinflussungen. Außerdem sind wir mitten in einer Debatte um Fake News, algorithmische Diskriminierung und digitale Polizeiüberwachung. All das sind bekannte Folgen der Digitalisierung. Ohne klare Regeln könnten sich diese bedenklichen Entwicklungen durch KI weiter beschleunigen. Deshalb hat die EU sich so beeilt, ihre KI-Verordnung zu erlassen.

7 Das *Techno-Optimist Manifesto* hat Marc Andreessen am 16.10.2023 auf seinem Blog veröffentlicht (Andreessen, 2023). Die Parallelen zum *Manifesto del Futurismo* von Marinetti sind nicht zu übersehen.

8 Das heißt aber auch nicht, dass alles gut reguliert und der Preis dafür nicht hoch war (Romermann, 2020).

Die EU will bei der Regulierung von KI eine internationale Führungsrolle einnehmen, aber ob ihr das gelingt, ist fraglich. Anfang November 2023 lud Rishi Sunak, zu dieser Zeit noch Premierminister des Vereinigten Königreichs, zum ersten internationalen KI-Gipfeltreffen ein. Ursula von der Leyen war als Kommissionspräsidentin der Europäischen Union genauso vor Ort wie Kamala Harris als Vizepräsidentin der Vereinigten Staaten. Als Vertreter der Tech-Industrie war auch Elon Musk dabei. Der Gipfel fand in Bletchley Park statt. Im Zweiten Weltkrieg befand sich dort eine streng geheime Dienststelle, die frühe Computer zum Codeknacken eingesetzt hat. Einer der Codeknacker in Bletchley Park war Alan Turing, der Erfinder der Turing-Maschine und Vordenker der KI. Heute befindet sich in Bletchley Park ein Computermuseum. Im Vergleich zur KI-Verordnung der EU ist die Abschlusserklärung des Gipfels, die Bletchley-Erklärung, erwartbar unkonkret geblieben.[9]

Treffen Tech-Lobbyisten, Nichtregierungsorganisationen und Politikerinnen und Politiker bei einem KI-Gipfel aufeinander, geht es natürlich darum, wie KI-Technologie reguliert werden sollte. Dabei sollte man nicht annehmen, dass internationale Konzerne nur daran interessiert sind, dass sie gar nicht reguliert werden.[10] Regulierung kann ihnen auch helfen, denn klare Spielregeln reduzieren juristische Risiken. Außerdem verschafft Regulierung ihnen gegenüber kleineren Firmen Wettbewerbsvorteile, weil sie sich die Bürokratie, die mit jeder Regulierung einhergeht, leichter leisten können. Die großen Player haben auch mehr Einfluss und sitzen mit am Tisch, wenn die Regeln gemacht werden.[11] Das Politikmagazin *Politico* führte mit vielen Menschen, die auf dem internationalen Parkett an der Diskussion über die Regulierung von KI beteiligt sind, Hintergrundgespräche.[12] Während die Techno-Optimisten sich in ihrer offenen Ablehnung jeglicher Regulierung einig sind, zerfällt die Gruppe derer, die eine Regulierung befürworten, in zwei Lager. Die entscheidende Frage, die sie trennt, lautet: Sind die

9 Sie finden sie, indem Sie hier nach der *Bletchley Declaration* suchen: https://www.gov.uk/search.

10 So gehörte etwa Sam Altman, der Chef von OpenAI, zu denjenigen, die in einer Anhörung im amerikanischen Senat Regulierung einforderten (Kang, 2023).

11 Nick Clegg, der frühere Vize-Premierminister des Vereinigten Königreichs, ist z.B. der Chef-Lobbyist von Meta. Wir können davon ausgehen, dass er weiß, wie man die Interessen von Meta am besten vertritt.

12 Siehe Scott et al. (2024).

kurzfristigen oder die langfristigen Risiken von KI wichtiger? Auf der einen Seite finden sich diejenigen, die schon heute viele Gefahren im Einsatz von KI sehen und den Weg der EU international weitergehen wollen. Auf der anderen Seite stehen diejenigen, die vor einem apokalyptischen Terminator-Szenario mit einem totalen Kontrollverlust warnen. Diese Seite sieht das Risiko vor allem in der ferneren Zukunft.

Warum können wir nicht beides tun, die akuten Gefahren eindämmen und uns auf eventuelle langfristige Risiken vorbereiten? Viele Verfechter der KI-Verordnung der EU halten das ganze Gerede von einem hypothetischen Terminator-Szenario für eine bewusste politische Taktik, um von bestehenden realen Problemen abzulenken, die auch außerhalb der EU dringend effektiv reguliert werden müssten. Im Gegensatz zu den Techno-Optimisten, fahren die Lobbyisten, die vor der Apokalypse warnen, eine subtilere Strategie: Sie umarmen die Regulierung, solange sie ihnen nutzt, ansonsten wollen sie aber auch nicht, dass Tech-Unternehmen zu stark reguliert werden. Je näher die Gefahr einer effektiven Regulierung kommt, desto lauter warnen diese Lobbyisten vor der Apokalypse, und desto zynischer werden ihre Kritiker.[13]

Das Ende naht

Die allermeisten KI-Forscherinnen und -Forscher sind keine bezahlten Lobbyisten für Tech-Unternehmen (obwohl sie häufig für diese arbeiten). Wenn sie sich öffentlich zu den existenziellen Risiken von KI äußern, sind sie wirklich davon überzeugt, dass KI zur Auslöschung der gesamten Menschheit führen könnte, und sehen es als ihre moralische Pflicht an, davor zu warnen. Mit dem Begriff ›existenzielles Risiko‹ bezeichnen sie alle Gefahren, die die Existenz der Menschheit bedrohen. KI ist für sie eine Technologie wie die Atomkraft, die großen Nutzen verspricht, aber eben auch große Risiken birgt und deshalb reguliert werden muss. Diese Überzeugung geht oft mit einer als ›Longtermism‹ bezeichneten Ethik einher, in der die langfristigen Folgen von KI bedacht werden müssen. Nick Bostrom ist einer der Apostel dieser Bewegung und das Buch *Superintelligence* sein Evangelium.[14] Longter-

13 Siehe Heaven (2023).

14 Bostrom (2014).

misten liegen mit Techno-Optimisten darüber im Streit, ob KI reguliert werden sollte oder nicht. Während Techno-Optimisten glauben, dass das Alignment-Problem gar kein Problem ist, zerbrechen sich die Longtermisten den Kopf darüber, wie sie technologisch und regulatorisch sicherstellen können, dass KI auf unsere Werte ausgerichtet ist und sich nicht irgendwann gegen uns stellen wird. Aber beide Sekten glauben daran, dass der Messias in Form einer superintelligenten KI früher oder später kommen wird. Die einen haben nur mehr Angst vor der Apokalypse als die anderen.

Im Grunde gehen die Longtermisten von einem durchaus vernünftigen Gedanken aus: Wir verschieben Probleme gerne in die Zukunft. Das sehen wir besonders gut am Beispiel des Klimawandels. Das Bundesverfassungsgericht hat deshalb 2021 geurteilt, dass das Klimaschutzgesetz die Interessen von nachfolgenden Generationen nicht ausreichend berücksichtigt, weil es Probleme auf die Zeit nach 2030 vertagt.[15] Dieses Urteil erinnert uns daran, dass wir nicht auf Kosten unserer Kinder und Enkelkinder leben sollten. So weit, so moralisch vernünftig.

Einige Longtermisten gehen aber weit über diese Idee der Generationengerechtigkeit hinaus: Alle noch ungeborenen Menschen sind genauso wichtig wie die heute lebenden. Und wenn die ›alle‹ sagen, meinen sie wirklich ›alle‹: Hilary Greaves und William MacAskill sind zwei Philosophen, die so etwas wie das longtermistische Manifest geschrieben haben.[16] Die Argumentation in diesem Manifest ist bemerkenswert. Entscheidungen sollten möglichst rational getroffen werden. (Wer will da widersprechen?) Das gilt insbesondere für politische Entscheidungen, die KI fördern oder regulieren sollen. Doch was ist rational? Rational ist, was den erwarteten Nutzen maximiert beziehungsweise die erwarteten Kosten minimiert. Dabei müssen die Kosten der gesamten Menschheit berücksichtigt werden. Wenn jeder nur an sich selber denkt, ist eben nicht an alle gedacht. Insbesondere denkt dann keiner an all die Menschen, die noch nicht geboren sind. Diese Menschen sind genauso viel wert wie die Menschen, die heute leben.

15 Siehe Bräutigam (2021).

16 Die folgende Überschlagsrechnung habe ich direkt aus dem Artikel von Greaves & MacAskill (2021) übernommen. Ich bin etwas unfair, ihr Argument auf die reine Kosten-Nutzen-Rechnung zu verkürzen, denn ganz so schlicht sind ihre Argumente nicht. Aber am Ende läuft das meiner Ansicht nach doch darauf hinaus. Dafür unterschlage ich zum Ausgleich die wirklich hanebüchenen Argumente.

Eine schnelle Überschlagsrechnung ergibt, dass es viel mehr zukünftige Menschen geben wird, als heute Menschen leben. Longtermisten erwarten, dass die Population der Menschheit bei etwa 10 Milliarden (10^{10}) ein stabiles Gleichgewicht erreichen wird. Davon sind wir heute nicht mehr so weit entfernt. Wir können also annehmen, dass zu jeder Zeit in der Zukunft etwa so viele Menschen leben wie heute. Wir wissen aufgrund von Fossilien, wie lange andere Säugetiere auf der Erde existiert haben, bevor sie ausgestorben sind. Das macht plausibel, dass die Menschheit noch mindestens 10.000 Jahrhunderte existieren wird. Nehmen wir vereinfachend an, dass Menschen 100 Jahre alt werden, dann kommen auf jeden heutigen Menschen 10.000 (10^4) Menschen in der Zukunft. Das sind 100 Billionen (10^{14}) insgesamt. Und das ist eine konservative Schätzung, die nicht berücksichtigt, dass wir in der Zukunft auf Asteroiden-Einschläge besser vorbereitet sein werden als die Dinosaurier und wir außerdem neuen Lebensraum auf dem Mars finden werden.[17]

Die Kosten-Nutzen-Rechnung der heutigen Menschen verblasst angesichts der Rechnung für die Gesamtkosten der Menschheit. Sollte auch nur ein winziges Risiko bestehen, dass die Menschheit durch KI langfristig ausgelöscht werden könnte, wiegt das schwerer als der kurzfristige Nutzen, den KI in unserer Lebenszeit haben wird.

Doch wie hoch ist das Risiko einer KI-Apokalypse? Fragt man KI-Expertinnen und -Experten danach, sagt die eine Hälfte größer und die andere Hälfte kleiner als fünf Prozent.[18] Allerdings sind KI-Experten – das sei schnell hinzugefügt – keine seriösen Zukunftsforscher, und entsprechend handelt es sich bei diesen Zahlen um Meinungen und Bauchgefühle, die anschaulich zeigen, dass diese KI-Experten zu viele Science-Fiction-Romane gelesen haben und die Fähigkeiten von KI genauso überschätzen wie ihre eigenen. Longtermisten wissen natürlich, wie hochumstritten solche Schätzungen sind. Trotzdem mal angenommen, die Wahrscheinlichkeit betrage tatsächlich fünf Prozent, dann stellt das natürlich ein unakzeptabel hohes Risiko dar. Aber selbst, wenn das wahre Risiko um ein Vielfaches kleiner ist, so ist es immer noch zu groß. Deshalb drängen einige der prominentesten Wissenschaftlerinnen und Wissenschaftler (darunter Nobelpreisträger)

17 Okay, jetzt habe ich ein paar der abstrusen Argumente von Greaves & MacAskill (2021) doch genannt.

18 Siehe Grace et al. (2018) und Grace et al. (2024).

darauf, dass wir angesichts einer möglichen Apokalypse dringend in KI-Sicherheit investieren müssen.[19] Und einige Longtermisten rechnen uns vor, dass wir dafür richtig viel Geld in die Hand nehmen sollten, weil wir nur so die Leben aller zukünftigen Menschen retten können.[20]

Da einige der erfolgreichsten KI-Forscher, die Einblicke in die aktuellen Entwicklungen bei den großen Tech-Unternehmen haben, und die Vorstände und Manager dieser Unternehmen sich öffentlich besorgt über KI äußern, sollte man das ernst nehmen. Entsprechend viel Aufmerksamkeit bekommt auch jeder ihrer offenen Briefe, die im Internet in schöner Regelmäßigkeit veröffentlicht werden.[21] In der Bletchley-Erklärung, die vor langfristigen KI-Risiken warnt, hallen diese zahlreichen Appelle nach. Der bekannteste dieser Briefe besteht nur aus einem Satz:

> Die Verringerung der Risiken einer Auslöschung der Menschheit durch KI sollte genauso eine globale Priorität der Politik sein wie Pandemien und Atomkriege.[22]

Zu den Erstunterzeichnern gehören die zwei KI-Nobelpreisträger Geoffrey Hinton und Demis Hassabis sowie Bill Gates und Sam Altman, der Chef von OpenAI. Inzwischen haben sich hunderte KI-Forscherinnen und -Forscher ihrem Aufruf angeschlossen. Man wundert sich nur, warum diese Leute weiter an KI arbeiten, wenn sie wirklich glauben, dass KI uns den Weltuntergang bringen wird.[23] Anders als bei Pandemien und Asteroiden-Einschlägen könnten wir das Risiko sofort auf null reduzieren, indem wir jegliche KI-Forschung einstellen. Warum fordern wahre Longtermisten aber kein komplettes Verbot? Manche glauben, dass wir uns in einem Wettbewerb befinden, der mit der nuklearen Aufrüstung vergleichbar ist. Vladimir Putin sagte 2017,

19 Siehe Bengio et al. (2024). Einer der Autoren dieses Artikels ist Geoffrey Hinton, der 2024 den Nobelpreis für seine Grundlagenforschung zu neuronalen Netzen bekommen hat. Unter den Autoren sind auch der Wirtschaftsnobelpreisträger Daniel Kahneman und der Historiker und Bestseller-Autor Yuval Noah Harari.

20 Siehe nochmal Greaves & MacAskill (2021).

21 Siehe https://futureoflife.org/fli-open-letters.

22 Siehe https://www.safe.ai/work/statement-on-ai-risk.

23 Statt der vielen offenen Briefe von Forschern, Entwicklern, Investoren und Managern, die über die Jahre erschienen sind, um vor KI zu warnen, lesen Sie doch lieber den unglaublich lustigen Brief von Kannan (2023).

dass derjenige die Welt beherrschen wird, der führend in KI ist.[24] Die Staaten, die nicht aufrüsten, werden an Macht und Einfluss verlieren. Die KI-Entwicklung ist dieser Aufrüstungslogik nach nicht aufzuhalten. Das Beste, was wir machen können, ist die Risiken zu minimieren. Die allermeisten glauben ohnehin, dass der erwartete Nutzen für die Menschheit so groß ist, dass das Risiko einer Auslöschung der Menschheit akzeptabel wird, sofern wir es durch KI-Sicherheitsforschung nur klein genug halten können. Heilserwartung sticht Apokalypse.

Als Elon Musk OpenAI mitgründete, versprach er, dass die Technologie des Unternehmens offen sein würde. Daher der Name. Damit war gemeint, dass Forschungs- und Entwicklungsergebnisse veröffentlicht werden, damit andere den Fortschritt kontrollieren und darauf aufbauen können. Die Technologie hinter KI dürfe nicht geheim sein – und auch nicht von einer einzigen Firma kontrolliert werden. Musk hat sich immer wieder öffentlich dazu geäußert, dass er KI für die wahrscheinlich größte Bedrohung der Menschheit hält. Seine Investitionen in KI dienten angeblich dazu, ein Terminator-Szenario zu verhindern.[25] Wie selbstlos von ihm! Doch je mehr Geld die KI-Entwicklung verbrauchte und je mehr Erfolge sichtbar wurden, desto verschlossener wurde OpenAI seltsamerweise. Denn da KI gefährlich sei, müsse sie – so hieß es nun – von OpenAI oder besser noch von Tesla kontrolliert werden.[26] Vielleicht ist das eigentliche Alignment-Problem: Wie stellen wir sicher, dass KI nicht nur an den Interessen von einem Mann wie Elon Musk ausgerichtet ist, der den Mars besiedeln möchte und seinen

24 Putin bemerkte außerdem, dass es nicht wünschenswert ist, wenn es in diesem Bereich ein Monopol gäbe, und Russland sein Wissen mit der Welt deshalb teilen würde (Associated Press, 2017). So wie das Land es ja auch mit Nukleartechnologie macht.

25 Siehe z.B. Hern (2014) oder Gibbs (2014).

26 Musk verließ OpenAI 2018. Ein paar Jahre später, 2024, wollte er die Firma verklagen, weil sie das hehre Ideal der Offenheit, das er angeblich seit der Gründung von OpenAI vertrat, aufgegeben hatte. Daraufhin veröffentlichte OpenAI in dieser Sache interne E-Mails, die dokumentierten, dass Musk die Kontrolle über OpenAI angestrebt und den Plan verfolgt hatte, das Unternehmen zu einem Teil seiner Firma Tesla zu machen. Musk zog daraufhin seine Klage zunächst wohl zurück (Duffy, 2024; Telford, Tiku & De Vynck, 2024). Inzwischen sieht es im Februar 2025 wieder so aus, als ob es doch ein Gerichtsverfahren geben würde (Tong & Sriram, 2025). Elon Musk liefert sich unterdessen im Internet einen unterhaltsamen Schlagabtausch mit Sam Altman, dem CEO von OpenAI. Das Drehbuch für die Verfilmung dieses Dramas schreibt sich von alleine. Ganz ohne Hilfe von ChatGPT.

elektrischen Sportwagen als PR-Gag ins Weltall geschossen hat?[27] Auf dem Weg zum Mars, auf dem Milch und Honig fließen, bremst ihn bestimmt niemand aus.

Manche KI-Forscher scherzen als Seitenhieb auf Musk gerne, dass sie sich über das existenzielle Risiko, das von KI ausgehen soll, genauso wenig Sorgen machen, wie über die Überbevölkerung auf dem Mars.[28] Als Leserin und Leser dieses Buches ahnen Sie schon, was jetzt kommt: Ich halte ein Terminator-Szenario genauso wie superintelligente KI für reine Science-Fiction. Aber ist die Wahrscheinlichkeit dafür null? Superintelligente KI ist denkbar. Aber nur weil etwas denkbar ist, heißt das nicht, dass eine realistische Chance besteht, dass wir eine superintelligente KI tatsächlich entwickeln können. Es ist ebenso denkbar, dass wir nicht alleine im Universum sind und eines Tages Außerirdische auf der Erde landen könnten. Auch ich lese gerne Science-Fiction und habe Spaß daran, verrückte Ideen bis zum Ende durchzudenken. Ohne eine gehörige Portion an Fantasie kann man keine Vision für die Zukunft entwickeln. Aber von Zeit zu Zeit braucht es eben auch einen Realitätscheck.

Stuart Russell, einer der Autoren des Standardlehrbuches zu KI, ist davon überzeugt, dass die Entwicklung einer superintelligenten KI, falls sie uns denn gelingt, das größte Ereignis in der Zukunft der Menschheit sein wird – größer noch als die Ankunft von Außerirdischen (seine Worte, nicht meine). Und er schreckt nicht davor zurück, einen so grottenschlechten Film wie *Transcendence* mit Johnny Depp heranzuziehen, um den Teufel an die Wand zu malen: Wie in diesem Film könnte die Menschheit die Kontrolle über KI verlieren.[29] Zum Glück weist er neben der extrem spekulativen Gefahr der Auslöschung der Menschheit auch auf die weniger spektakulären, kurzfristigen Gefahren und die Notwendigkeit diese zu regulieren hin, so wie das in der KI-Verordnung der EU geschehen ist.[30] Aber ob er will oder nicht, er trägt mit seiner Rhetorik dazu bei, dass die verrückten Positionen der extremen Longtermisten von akuten Problemen ablenken.

27 Siehe Wattles (2024).

28 Andrew Ng hat damit angefangen (Williams, 2015).

29 Siehe das erste Kapitel in Russell (2019) und Hawking, Tegmark & Russell (2014).

30 Siehe nochmal Russell (2019), aber auch Russell (2023) und Weibel (2024).

Zweckrationalität sticht Moral

Die meisten Longtermisten sind sich durchaus dessen bewusst, dass ihre Argumentation, konsequent zu Ende gedacht, verrückt ist. Sie stellen sich untereinander die Frage: »An welcher Haltestelle des Zuges zur Stadt der Verrückten steigst Du aus?«[31] Und die meisten steigen recht früh aus, etwa an den Haltestellen Klimaschutz oder Generationengerechtigkeit. Manche folgen aber der Argumentation des longtermistischen Manifests und sind überzeugt: In der Kosten-Nutzen-Rechnung für die gesamte Menschheit sind die heute lebenden Menschen vernachlässigbar. Diese Longtermisten predigen, dass wir heutiges Leid ertragen müssen, sofern es der Zukunft der Menschheit dient. Dass dieser Zweck jedes Mittel heiligen kann, scheint sie nicht weiter zu beunruhigen.

Extreme Longtermisten, die im Zug bis zur Endhaltestelle sitzen bleiben, sind offensichtlich eine Karikatur. Trotzdem bleibt der Eindruck, dass auch gemäßigte Longtermisten für das ewige Heil der Menschheit bereitwillig irdisches Leid hinnehmen. Hilary Greaves, die Erstautorin des longtermistischen Manifests, wurde auf die Obdachlosen angesprochen, die sie auf den Straßen sieht, und für die sie nichts tut, während sie sich um Menschen sorgt, die noch nicht einmal geboren sind. Sie bemerkte dazu:

> Ich fühle mich echt schlecht, aber das schlechte Gefühl ist begrenzt, weil ich wirklich denke, dass ich das Richtige tue […]. Die moralisch angemessene Position ist irgendwo in der Mitte, wo einen das heutige Leid immer noch mitnimmt, aber man erkennt, dass es noch wichtigere Dinge gibt, die man mit den begrenzten Ressourcen machen kann.[32]

Falls es noch Zweifel gab: Longtermisten machen Lobbyarbeit für zukünftige Generationen, nicht für Obdachlose. Nach der Logik der Longtermisten müsste der Staat weniger Geld für Obdachlose und sozialen Wohnungsbau ausgeben und die Ressourcen stattdessen in KI-Sicherheit investieren, denn das existenzielle Risiko, das von KI ausgeht, bedroht schließlich die Zukunft der gesamten Menschheit.

---

31 Dieses Zitat stammt aus dem äußerst lesenswerten Artikel von Samuel (2022). Ein Großteil der folgenden Argumentation ist direkt aus diesem Artikel übernommen.

32 Dieses Zitat stammt aus einem früheren Artikel von Samuel (2021).

Wie staatliche Ressourcen eingesetzt werden, ist eine politische Entscheidung, über die in einer Demokratie gestritten werden muss. Und vielleicht sollten wir tatsächlich etwas mehr Geld für KI-Sicherheit ausgeben. Für Longtermisten ist das aber keine schnöde politische Diskussion, es ist eine moralische Frage, die sie durch ihre Kosten-Nutzen-Analyse als bereits beantwortet ansehen.

Diesen moralischen Maßstab legen sie nicht nur für die Gesellschaft an, sondern auch für das Individuum: Jemand, der Geld spendet, sollte es nicht für wohltätige Zwecke spenden, um die Not von Obdachlosen zu lindern oder die systemischen Ursachen von Obdachlosigkeit zu bekämpfen. Wichtiger als soziale Wohltätigkeit ist für Longtermisten die Forschung zu KI-Sicherheit, denn der erwartete Nutzen ist hier wesentlich größer. Dieser Grundsatz gilt für Milliardäre genauso wie für Philosophen. Wer jung ist und kein Geld hat, aber die 80.000 Stunden seines zukünftigen Arbeitslebens nicht mit sinnlosen Tätigkeiten vergeuden möchte, wird statt Sozialarbeiter besser KI-Sicherheitsforscher.[33] Der longtermistische Imperativ ist: Tue das, was langfristig den erwarteten Nutzen für die Menschheit maximiert!

Wenn KI-Forscher von Vernunft sprechen, dann sprechen sie von instrumenteller Vernunft. Die Rationalität der Maschinen ist eine reine Zweckrationalität. KI-Methoden suchen nach einem Weg, ein Ziel zu erreichen. Das ist die einzige Form von Vernunft, die sie kennen. Dabei folgen sie einer strengen Kosten-Nutzen-Rechnung. Der beste Weg ist der, der den erwarteten Nutzen maximiert. Je erfolgreicher KI-Methoden werden und je weiter sie sich verbreiten, desto mehr werden sie auch auf Probleme angewandt, für die sie nicht gemacht wurden. Darin unterscheidet sich die instrumentelle Vernunft der KI-Forscher nicht von der ökonomischen Vernunft der Wirtschaftswissenschaftler. Für beide sind Kosten und Nutzen zu Metaphern für Leid und Heil geworden. Dass weder Leid noch Heil leicht messbar sind, ist in ihren Augen nur ein technisches Problem, das noch zu lösen ist. Diese metaphorische Rationalisierung hat einen Nebeneffekt: Wir ersetzen Mit-

33 Falls Sie denken, ich denke mir das aus, dann denken Sie falsch. Auf dieser Webseite finden Sie Ratschläge dafür, wie Sie mit Ihrer Karriere den größtmöglichen Impact erreichen können: https://80000hours.org/. Nach eigenen Angaben hat die Seite bis 2024 zehn Millionen Leser angezogen und 400.000 Menschen haben den Newsletter abonniert. Einer der Gründer der Webseite ist William MacAskill, einer der Autoren des longtermistischen Manifests.

gefühl durch abstrakte Zahlen. So werden Schicksale zu Zahlen in einer Tabelle, die gegeneinander aufgerechnet werden können.

In der Debatte um die Zukunft von KI hat diese Art von Logik schon einige KI-Jünger auf eine von zwei intellektuellen Irrfahrten geführt. Im immer schneller werdenden Sportwagen auf der Straße des Fortschritts sitzen die Techno-Optimisten, für die es keine moralische Vernunft mehr gibt. Für sie gibt es nur gleichwertige Partikularinteressen, die die KI-Systeme der Zukunft zum Wohle der Menschheit in einem unregulierten Markt durchsetzen werden. Und an der Endhaltestelle des Zuges zur Stadt der Verrückten tummeln sich die extremen Longtermisten, die an die Möglichkeit einer moralischen Kosten-Nutzen-Rechnung für die ganze Menschheit glauben. Die KI-Systeme der Zukunft müssen nur noch danach ausgerichtet werden. Auf beiden Irrfahrten in die Zukunft bleibt die Menschlichkeit auf der Strecke.

Joseph Weizenbaum, der Entwickler von ELIZA, kritisierte schon 1976 den Imperialismus der instrumentellen Vernunft, der keine andere Art von Vernunft mehr neben sich duldet.[34] Die instrumentelle Vernunft hilft uns aber leider nicht zu entscheiden, was unsere Ziele sein sollen. Dafür haben wir keine Rechenregeln. Wir können auch nicht logisch beweisen, welche Ziele für unsere Gesellschaft die richtigen sind. Moralische und politische Entscheidungen folgen nicht nur einer Logik der Nutzenmaximierung. Um uns auf gesellschaftliche Ziele zu einigen, müssen wir langwierige und schwierige Debatten darüber führen, was moralisch und politisch vernünftig ist. Dafür haben wir eine Demokratie. In der Politik geht es nicht nur um die Durchsetzung von Partikularinteressen. Politik ist auch nicht nur ein technokratischer Streit über den besten Weg. Vielmehr ist Politik vor allem ein Ringen um die richtigen Ziele. Das gilt insbesondere für die Ziele von Forschung und Entwicklung im Bereich von KI. Die dafür nötigen Debatten kann uns keine rein hypothetische superintelligente KI abnehmen.

34 Weizenbaum (1976) widmet das ganze 10. Kapitel diesem Thema. Er beruft sich dabei auf die Kritik der instrumentellen Vernunft von Horkheimer (1947).