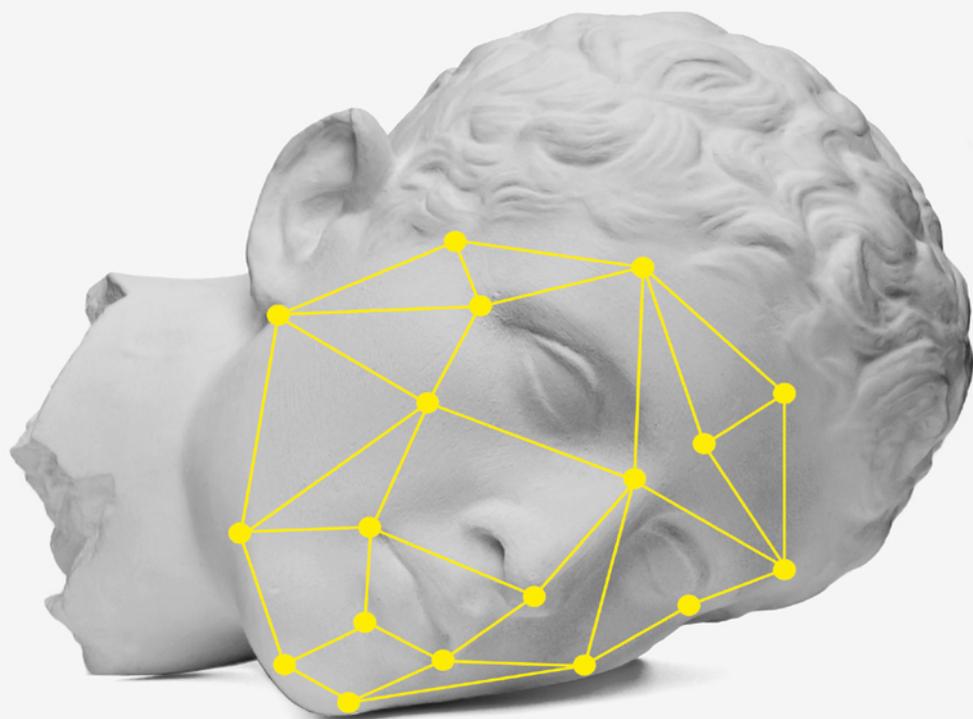


STUART RUSSELL

HUMAN COMPATIBLE



Künstliche Intelligenz

und wie der Mensch die Kontrolle über
superintelligente Maschinen behält



INHALT

Vorwort	7
Danksagung	8
Kapitel 1: Wenn wir Erfolg haben	9
Kapitel 2: Natürliche und künstliche Intelligenz	21
Kapitel 3: Mögliche Fortschritte in der KI	71
Kapitel 4: Missbrauch der KI	113
Kapitel 5: Übermäßig intelligente KI	143
Kapitel 6: Die gar nicht mal so große KI-Debatte	157
Kapitel 7: KI: Ein anderer Ansatz	183
Kapitel 8: Nachweislich vorteilhafte KI	197
Kapitel 9: Komplikationen: Wir	225
Kapitel 10: Problem gelöst?	261
Anhang A: Das Suchen nach Lösungen	273
Anhang B: Wissen und Logik	283
Anhang C: Unsicherheit und Wahrscheinlichkeit	289
Anhang D: Aus Erfahrung lernen	301
Anmerkungen	313
Bildnachweise	357
Stichwortverzeichnis	359

VORWORT

Warum Sie dieses Buch genau jetzt lesen sollten

In diesem Buch geht es um die Vergangenheit, die Gegenwart und die Zukunft unserer Bestrebungen, Intelligenz zu verstehen und zu erschaffen. Das ist ein wichtiges Thema, denn die KI (*Künstliche Intelligenz*, auch AI für *Artificial Intelligence*) ist unserer Tage nicht nur bereits allgegenwärtig, sondern wird auch und vor allem die wohl wichtigste Technologie der Zukunft sein. Die Weltmächte werden sich dessen langsam bewusst – und die größten Unternehmen wissen es schon seit geraumer Zeit. Wir können nicht genau sagen, wie und mit welchem Tempo sich diese Technologie weiterentwickeln wird. Aber eines ist klar: Wir müssen uns darauf vorbereiten, dass Maschinen die menschliche Fähigkeit zur Entscheidungsfindung eines Tages vielleicht übertreffen werden. Welche Folgen hätte das?

Alles, was unsere Zivilisation hervorgebracht hat, ist das Ergebnis unserer Intelligenz. Wenn wir also Zugang zu einer deutlich höheren Intelligenz hätten, wäre dies zweifellos das größte Ereignis der Menschheitsgeschichte. In diesem Buch wird erklärt, warum es sich dabei auch um das letzte Ereignis handeln könnte – und wie wir das verhindern können.

Was Sie erwartet

Dieses Buch ist in drei große Themenbereiche unterteilt: Der erste Teil in den Kapiteln 1 bis 3 beschäftigt sich mit menschlicher und maschineller Intelligenz an sich. Dabei wird keine technische Vorbildung vorausgesetzt. Wenn Sie tiefer in die Materie einsteigen wollen, finden Sie eine Erläuterung der Grundkonzepte moderner KI-Systeme in den Anhängen A bis D. Der zweite Teil, Kapitel 4 bis 6, behandelt einige der Probleme, die aus intelligenten Maschinen erwachsen. Das größte davon ist, wie wir es schaffen können, jederzeit die Kontrolle über Maschinen zu behalten, die mächtiger sind als wir. Im dritten Teil in den Kapiteln 7 bis 10 wird ein neuer Ansatz der KI-Forschung vorgestellt, der garantieren soll, dass die Maschinen dem Menschen

auf alle Zeit nützlich und dienlich sind. Das Buch richtet sich an alle interessierten Leserinnen und Leser. Ich hoffe allerdings, dass es auch von vielen Spezialisten der künstlichen Intelligenz gelesen wird und sie dazu anregt, ihre grundlegenden Anschauungen zu überdenken.

DANKSAGUNG

An der Entstehung dieses Buchs waren viele Menschen beteiligt. Dazu gehören Paul Slovak, Lektor bei Viking, sowie Laura Stickney, Lektorin bei Penguin, mein Literaturagent John Brockman, der mich zum Schreiben ermutigt hat, Jill Leovy und Rob Reid, die ein Füllhorn an nützlichem Feedback für mich hatten, und weitere Leser der ersten Entwürfe, insbesondere Ziyad Marar, Nick Hay, Toby Ord, David Duvenaud, Max Tegmark und Grace Cassy. Caroline Jeanmaire war eine unschätzbare Hilfe beim Sammeln und Zusammenstellen der unzähligen Verbesserungsvorschläge der Testleser und Martin Fukui hat sich für mich um die Bildrechte gekümmert.

Die wesentlichen technischen Konzepte in diesem Buch habe ich gemeinsam mit Mitgliedern des *Center for Human-Compatible AI* in Berkeley entwickelt. Besonders erwähnen möchte ich Tom Griffiths, Anca Dragan, Andrew Critch, Dylan Hadfield-Menell, Rohin Shah und Smitha Milli. Das Center wird auf bewundernswerte Weise von Mark Nitzberg und Rosie Campbell geleitet und von der Open Philanthropy Foundation großzügig mit finanziellen Mitteln ausgestattet.

Ramona Alvarez und Carine Verdeau haben dafür gesorgt, dass während der gesamten Entstehungsgeschichte des Buchs alles glatt vorstattenging. Meine unglaubliche Ehefrau Loy und unsere gemeinsamen Kinder Gordon, Lucy, George und Isaac haben mit ihrer unermesslichen Liebe, Nachsicht und Unterstützung zum Gelingen und Abschluss dieses Projekts beigetragen.

WENN WIR ERFOLG HABEN

Vor längerer Zeit lebten meine Eltern im englischen Birmingham in einem Haus in der Nähe der Universität. Sie entschieden sich, von der Stadt aufs Land zu ziehen, und verkauften das Haus an David Lodge, einen Professor für englische Literatur. Lodge war zu jener Zeit ein bekannter Romanautor. Ich habe ihn zwar nie persönlich kennengelernt, entschloss mich aber, einige seiner Werke zu lesen: *Ortswechsel* (Originaltitel: *Changing Places*) und *Schnitzeljagd* (Originaltitel: *Small World*). Zu den Hauptakteuren gehören fiktionale Akademiker, die aus einem fiktionalen Birmingham in ein fiktionales Berkeley in Kalifornien ziehen. Da ich ein echter Akademiker aus dem echten Birmingham bin, der gerade ins echte Berkeley gezogen war, rief mich das Ministerium für Zufälle offenbar dazu auf, genauer hinzusehen.

Eine Szene aus *Schnitzeljagd* hat es mir besonders angetan: Der Protagonist, ein aufstrebender Literaturtheoretiker, besucht eine internationale Konferenz und fragt ein hochkarätig besetztes Podium: »Was geschieht, wenn alle Ihnen zustimmen?« Die Frage sorgt für Verblüffung, denn es ging den Diskutanten mehr um den geistigen Schlagabtausch und weniger um Wahrheitsfindung oder Erkenntnisgewinn. In dem Moment wurde mir klar, dass man den führenden Persönlichkeiten in der KI eine ganz ähnliche Frage stellen könnte: »Was, wenn Sie Erfolg haben?« Diese Disziplin hat sich stets das Ziel gesetzt, eine KI mit menschlichen oder gar übermenschlichen Fähig-

keiten zu erschaffen. Was für Folgen das haben würde, hat man sich allerdings nur selten oder gar nicht gefragt.

Ein paar Jahre später begannen Peter Norvig und ich mit der Arbeit an einem neuen KI-Lehrbuch, das erstmals 1995 veröffentlicht wurde.¹ Der letzte Abschnitt mit dem Titel »What If We Do Succeed?« (Was, wenn wir Erfolg haben?) weist auf mögliche positive und negative Folgen hin, ohne ein eindeutiges Fazit zu ziehen. Als 2010 die dritte Auflage erschien, waren viele Menschen zu dem Schluss gekommen, dass eine übermenschliche KI vielleicht gar keine so gute Sache wäre. Allerdings waren die meisten von ihnen eher fachfremd und gehörten nicht zum Mainstream der KI-Forscher. 2013 war ich inzwischen davon überzeugt, dass dieses Thema nicht nur mitten in die KI-Forschung gehört, sondern möglicherweise die wichtigste Frage für die gesamte Menschheit darstellt.

Im November 2013 hielt ich einen Vortrag in der Dulwich Picture Gallery, einem ehrwürdigen Kunstmuseum im Süden Londons. Das Publikum bestand größtenteils aus interessierten Seniorinnen und Senioren ohne wissenschaftliche Vorbildung. Ich musste bei meinem Vortrag also komplett auf technische Fachbegriffe verzichten. Das war eine hervorragende Gelegenheit, meine Überlegungen erstmals in der Öffentlichkeit zu präsentieren. Nachdem ich erklärt hatte, worum es bei KI geht, nominierte ich fünf Kandidaten für das »größte Ereignis in der Zukunft der Menschheit«:

1. Wir sterben aus (Asteroidenaufprall, Klimakatastrophe, Pandemie usw.).
2. Wir leben ewig (medizinische Lösung gegen das Altern).
3. Wir finden eine Möglichkeit, schneller als das Licht zu reisen, und erobern das Universum.
4. Wir erhalten Besuch von einer uns überlegenen außerirdischen Zivilisation.
5. Wir erfinden eine superintelligente KI.

Ich votierte für die fünfte Option, die superintelligente KI, denn diese würde uns helfen, (Natur-)Katastrophen zu vermeiden, den Schlüssel zum ewigen Leben zu finden und schneller als das Licht zu reisen – sofern das überhaupt möglich ist. Das wäre wirklich ein gewaltiger Sprung für unsere Zivilisation. Danach wäre nichts mehr wie zuvor. Eine superintelligente KI würde in vielen Punkten der Ankunft

einer überlegenen Alien-Rasse ähneln, ist aber sehr viel wahrscheinlicher. Und was noch wichtiger ist: Bei der KI haben wir ein Wörtchen mitzureden, bei den Außerirdischen wohl eher nicht ...

Dann fragte ich das Publikum, was wohl geschehen würde, wenn wir heute eine Botschaft von Außerirdischen erhielten, die deren Ankunft auf der Erde in 30 Jahren ankündigt. Das Wort »Hölle« dürfte für das zu erwartende Chaos kaum ausreichen. Unsere Reaktion auf die zu erwartende Ankunft von superintelligenter KI fällt dagegen mehr als verhalten aus. (In einem späteren Vortrag habe ich das durch den in Abbildung 1.1 dargestellten E-Mail-Austausch verdeutlicht.) Abschließend habe ich die Bedeutung einer superintelligenten KI wie folgt erklärt: »Ein Erfolg wäre das größte Ereignis in der Menschheitsgeschichte ... aber vielleicht auch das letzte.«

Von: Überlegene außerirdische Zivilisation <sac12@sirius.canismajor.u>

An: Menschheit@UN.org

Betreff: Kontakt

Achtung: Wir kommen in 30 bis 50 Jahren an

Von: Menschheit@UN.org

An: Überlegene außerirdische Zivilisation <sac12@sirius.canismajor.u>

Betreff: Abwesenheitsbenachrichtigung: Re: Kontakt

Die Menschheit ist gerade nicht erreichbar. Wir melden uns, wenn wir wieder im Büro sind. 😊

Abb. 1.1: Vermutlich nicht die Antwort, die wir einer überlegenen Zivilisation von Außerirdischen nach deren Kontaktaufnahme senden würden

Ein paar Monate später, im April 2014, besuchte ich eine Konferenz in Island. National Public Radio rief mich an und bat um ein Interview zum gerade in den Vereinigten Staaten gestarteten Kinofilm *Transcendence*. Ich hatte zwar Zusammenfassungen der Handlung und auch einige Kritiken gelesen, den Film aber nicht gesehen, denn in meiner damaligen Wahlheimat Paris sollte er erst im Juni in die Kinos kommen. Allerdings flog ich auf dem Rückweg erst nach Boston, da ich dort an einer Besprechung des Verteidigungsministeriums teilnehmen wollte. Nach meiner Ankunft am Bostoner Logan Airport nahm ich ein Taxi zum nächsten Kino, in dem der Film lief. In der zweiten Reihe sitzend, verfolgte ich, wie ein von Johnny Depp gespielter KI-Professor der

Uni Berkeley von fanatischen KI-Skeptikern niedergeschossen wird, die sich vor einer superintelligenten KI fürchten. Unwillkürlich sank ich in meinem Sitz zusammen. (Hatte sich hier wieder einmal das Ministerium für Zufälle bei mir gemeldet?) Bevor der von Johnny Depp gespielte Wissenschaftler stirbt, wird sein Geist in einen Quanten-Supercomputer übertragen, übertrifft bald die menschlichen Fähigkeiten und droht, die Weltherrschaft an sich zu reißen.

Am 19. April 2014 erschien in der *Huffington Post* eine von den Physikern Max Tegmark, Frank Wilczek und Stephen Hawking gemeinsam verfasste Besprechung zu *Transcendence*. Sie enthielt den Satz über das größte Ereignis in der Menschheitsgeschichte, den ich in meiner Rede in Dulwich verwendet hatte. Seither gelte ich in der Öffentlichkeit als der Forscher, der sein eigenes Forschungsgebiet für eine potenzielle Gefahr für uns als Spezies hält.

Was bisher geschah ...

Die Anfänge der KI reichen zurück bis in die Antike, aber der »offizielle« Startschuss fiel 1956. Die beiden jungen Mathematiker John McCarthy und Marvin Minsky hatten den bereits als Erfinder der Informationstheorie berühmt gewordenen Claude Shannon sowie Nathaniel Rochester, den Entwickler des ersten kommerziellen Computers von IBM, dazu überredet, mit ihnen gemeinsam eine Sommerschule am Dartmouth College zu organisieren. Das Ziel wurde wie folgt beschrieben:

»Die Studie soll von der Annahme ausgehen, dass grundsätzlich alle Aspekte des Lernens und anderer Merkmale der Intelligenz so genau beschrieben werden können, dass eine Maschine dazu gebracht werden kann, sie zu simulieren. Es soll versucht werden, herauszufinden, wie Maschinen dazu gebracht werden können, Sprache zu benutzen, Abstraktionen und Begriffe zu bilden, Probleme zu lösen, die zu lösen bislang dem Menschen vorbehalten sind, und sich selbst zu verbessern. Wir denken, dass ein bedeutender Fortschritt auf einem oder mehreren dieser Gebiete erzielt werden kann, wenn eine sorgfältig ausgewählte Gruppe von Wissenschaftlern einen Sommer lang zusammen daran arbeitet.«

Wir wissen heute, dass der Sommer nicht ausgereicht hat. Tatsächlich arbeiten wir noch heute an all diesen Problemen.

Im ersten Jahrzehnt nach der Dartmouth-Konferenz konnte die KI mehrere große Erfolge verbuchen, darunter den Algorithmus von Alan Robinson für allgemeine logische Schlussfolgerungen² und das Dameprogramm von Arthur Samuel, das sich selbst beibrachte, seinen Schöpfer zu schlagen.³ Die erste KI-Blase platzte in den späten 1960er-Jahren, da frühe Bemühungen in den Bereichen Machine Learning und maschinelle Übersetzung die in sie gesetzten Erwartungen nicht erfüllten. Ein von der britischen Regierung 1973 in Auftrag gegebener Bericht kam zu folgendem Schluss: »In keinem Bereich des Forschungsgebiets haben die bisherigen Entdeckungen zu den versprochenen gewaltigen Auswirkungen geführt.«⁴ Anders ausgedrückt: Die Maschinen waren einfach nicht klug genug.

Mit meinen elf Jahren kannte ich diesen Bericht zum Glück nicht. Als ich zwei Jahre später den programmierbaren Taschenrechner Sinclair Cambridge bekam, wollte ich ihn intelligent machen. Leider konnten Programme für das Gerät maximal 36 Zeichen umfassen, was für KI auf menschlichem Level nicht ausreicht. Das konnte mich nicht aufhalten. Nachdem ich Zugang zu dem riesigen Supercomputer CDC 6600⁵ am Imperial College London bekommen hatte, schrieb ich ein Schachprogramm: einen 60 cm hohen Lochkartenstapel. Das Programm selbst war nicht besonders gut, aber das war nicht wichtig. Ich kannte jetzt meine Bestimmung.

Mitte der 1980er war ich Professor in Berkeley und die KI war wieder im Kommen – dank des kommerziellen Potenzials der sogenannten Expertensysteme. Die zweite KI-Blase platzte, als sich diese Systeme für viele der Aufgaben, die sie übernehmen sollten, als unzureichend erwiesen. Auch hier waren die Maschinen nicht klug genug. Ein KI-Winter brach an. Mein eigener KI-Kurs in Berkeley, der heute mit über 900 Studierenden aus allen Nähten platzt, wurde 1990 von lediglich 25 Personen besucht.

Die KI-Community hatte ihre Lektion gelernt: Klüger war in jedem Fall besser. Aber wie ließ sich das erreichen? Die Mathematik sollte es richten. Wir knüpften an bewährte Disziplinen an: Wahrscheinlichkeitsrechnung, Statistik und Kontrolltheorie. Die Saat für die heutigen Fortschritte wurde in jenem KI-Winter gelegt. Dazu gehörten frühe Arbeiten an großen probabilistischen Schlussfolgerungssystemen und

einem Gebiet, das später unter der Bezeichnung *Deep Learning* bekannt wurde.

Ab etwa 2011 ermöglichten Deep-Learning-Techniken dramatische Durchbrüche in der Spracherkennung, der visuellen Objekterkennung und der maschinellen Übersetzung: drei der wichtigsten bisher ungelösten Probleme der KI. Bei einigen Tests bewältigen Maschinen diese Aufgaben heute ebenso gut oder besser als der Mensch. In den Jahren 2016 und 2017 besiegte AlphaGo von DeepMind den ehemaligen Go-Weltmeister Lee Sedol und den amtierenden Meister Ke Jie. Das hatten Fachleute frühestens für das Jahr 2097 erwartet, wenn überhaupt jemals.⁶

Heute lesen wir in den Medien fast täglich von den Fortschritten der KI. Tausende von Start-ups wurden gegründet, von reichlich Risikokapital finanziert. Millionen Studierende absolvieren Onlinekurse zu KI und Machine Learning. Kein Wunder, locken doch Spitzengehälter von mehreren Millionen Dollar. Die Investitionen von Venture-Fonds, Regierungen und Großunternehmen in diesem Bereich betragen zig Milliarden Dollar pro Jahr. Allein in den letzten fünf Jahren wurde das Feld besser mit Finanzmitteln ausgestattet als in seiner gesamten bisherigen Geschichte. Die Projekte, an denen gerade gearbeitet wird, werden im Laufe des nächsten Jahrzehnts höchstwahrscheinlich große Auswirkungen auf unser Leben haben: selbstfahrende Autos zum Beispiel oder intelligente persönliche Assistenten. Die möglichen wirtschaftlichen und sozialen Vorteile der KI sind gewaltig. Das sorgt natürlich für enormen Schwung in der KI-Forschung.

... und was uns noch erwartet

Bedeutet dieser rasante Fortschritt, dass die Maschinen dabei sind, uns zu überholen? Keineswegs. Bevor es superintelligente Maschinen geben wird, sind noch einige Hürden zu überwinden.

Wissenschaftliche Durchbrüche sind generell kaum vorhersagbar. Wie komplex das Thema ist, zeigt die Geschichte eines anderen Forschungsgebiets, das ebenfalls das Potenzial hat, der Zivilisation ein Ende zu bereiten: die Kernphysik.

Anfang des 20. Jahrhunderts gab es wohl keinen bekannteren Kernphysiker als Ernest Rutherford, den Entdecker des Protons, den Mann, der das erste Atom spaltete (siehe Abbildung 1.2 [a]). Wie seine

Kollegen war sich Rutherford bereits seit langer Zeit darüber im Klaren, dass Atomkerne gewaltige Mengen an Energie speichern. Allerdings war die vorherrschende Meinung, dass man diese Energiequelle nicht anzapfen könne.

Am 11. September 1933 hielt die britische Wissenschaftliche Gesellschaft (*British Association for the Advancement of Science*) in Leicester ihre Jahrestagung ab. Lord Rutherford war Sprecher der abendlichen Sitzung. Wie schon zu früheren Gelegenheiten machte er den Anwesenden die Kernenergie betreffend keine großen Hoffnungen: »Jeder, der in der Umwandlung dieser Atome eine Energiequelle sieht, redet Unsinn.«⁷ Am nächsten Morgen war Rutherfords Rede in der Londoner *Times* abgedruckt (siehe Abbildung 1.2 [b]).

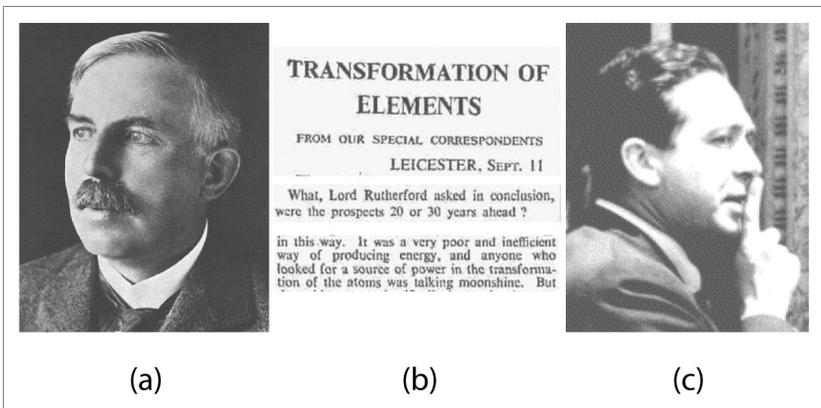


Abb. 1.2: (a) Lord Rutherford, Kernphysiker, (b) Auszug aus einem Bericht der *Times* vom 12. September 1933 über eine Rede, die Rutherford am Vorabend gehalten hatte, (c) Leó Szilárd, Kernphysiker

Leó Szilárd (siehe Abbildung 1.2 [c]), ein ungarischer Physiker, der kurz zuvor vor dem Nationalsozialismus aus Deutschland geflohen war, wohnte zu der Zeit im Imperial Hotel am Russell Square in London. Beim Frühstück las er den Artikel in der *Times*. Über das Gelesene nachsinnend, unternahm er einen Spaziergang, bei dem ihm die Idee zu der durch Neutronen hervorgerufenen nuklearen Kettenreaktion kam.⁸ So wurde das »unmöglich lösbare« Problem der Nutzung der Kernenergie praktisch in weniger als 24 Stunden gelöst. Im folgenden Jahr reichte Szilárd ein geheimes Patent für einen Kernreaktor ein. Das erste Patent für eine Atomwaffe wurde 1939 in Frankreich erteilt.

Und die Moral der Geschichte? Eine Wette gegen den menschlichen Einfallsreichtum ist leichtsinnig, vor allem wenn es um unser aller Zukunft geht. In der KI-Community macht sich eine Art Verleugnungshaltung breit: Manchmal wird bereits die Möglichkeit eines Erfolgs der langfristigen KI-Ziele bestritten. Das erinnert an einen Busfahrer, der die gesamte Menschheit herumkutschert. In einem Affenzahn hält er auf eine Klippe zu und behauptet gleichzeitig: »Ja klar stürzen wir bei dem Tempo über die Klippe. Aber keine Angst, uns geht vorher noch das Benzin aus!«

Ich sage keinesfalls, dass der Erfolg in der KI *unabdingbar* ist. Und ich bin mir ziemlich sicher, dass es in den nächsten Jahren nicht dazu kommen wird. Dennoch ist es sicher klug, sich auf die Möglichkeit vorzubereiten. Wenn alles gut geht, brechen goldene Zeiten für die Menschheit an, aber wir müssen uns vor Augen halten, dass wir an Dingen arbeiten, die sehr viel mächtiger sind als der Mensch. Wie können wir denn sicherstellen, dass diese Dinge niemals in irgendeiner Weise Macht über uns haben werden?

Damit Sie besser verstehen, wie gefährlich das Ganze ist, hilft ein Blick auf die Auswahl- und Empfehlungsalgorithmen in den sozialen Medien. Sie sind nicht besonders intelligent, und doch beeinflussen sie Milliarden von Menschen und damit indirekt die ganze Welt. Üblicherweise zielen solche Algorithmen darauf ab, die *Klickrate* zu maximieren. Die Klickrate bezeichnet die Wahrscheinlichkeit, mit der ein Nutzer auf die präsentierten Elemente klickt. Wenn wir also einen maximalen Wert erzielen möchten, müssen wir nur solche Elemente anzeigen, auf die Nutzer gern klicken, richtig? Falsch. Die Lösung besteht darin, die Vorlieben der Nutzer so zu verändern, dass sie transparenter werden. Einem besser durchschaubaren Nutzer können Elemente präsentiert werden, auf die er wahrscheinlich klickt und somit mehr Einnahmen generiert. Menschen mit sehr extremen politischen Ansichten sind in dieser Hinsicht meist besser durchschaubar. (Möglicherweise gibt es auch Artikel, auf die extrem gemäßigte Politikinteressierte mit hoher Wahrscheinlichkeit klicken, aber das Thema dieser Artikel ist schwer vorstellbar.) Wie ein rationales Wesen lernt der Algorithmus, wie er den Zustand seiner Umgebung verändern kann, um die eigene Belohnung zu maximieren.⁹ In diesem Fall ist die Umgebung die Meinung des Nutzers. Die Folgen sind ein Wiederaufleben des Faschismus, die Kündigung des Gesellschaftsvertrags, auf dem

Demokratien weltweit gründen, und möglicherweise das Ende der Europäischen Union und der NATO. Erstaunlich, was ein paar Zeilen Programmcode anrichten können. Wenn schon ein relativ einfacher Algorithmus dazu imstande ist, wozu ist dann erst ein *wirklich* intelligenter Algorithmus in der Lage?

Was lief schief?

Die Geschichte der KI wird von einem zentralen Mantra bestimmt: »Je intelligenter, desto besser.« Ich bin überzeugt, dass das ein Fehler ist, und zwar nicht, weil ich irgendwie befürchte, ersetzt zu werden, sondern aufgrund unseres Verständnisses von Intelligenz an sich.

Unsere Vorstellung von Intelligenz ist zentral für unser Selbstverständnis. Nicht umsonst bezeichnen wir uns als *Homo sapiens*, als »weiser Mensch«. Nach über 2.000 Jahren der Selbstreflexion lässt sich unsere Idee von Intelligenz wie folgt zusammenfassen:

Menschen sind insoweit intelligent, als unsere Handlungen darauf ausgerichtet sind, unsere Ziele zu erreichen.

Alle anderen Merkmale für Intelligenz – wahrnehmend, denkend, lernend, erfindend und so weiter – lassen sich als Bestandteile unseres erfolgreichen Handlungsvermögens betrachten. Seit Anbeginn der KI-Forschung wurde Intelligenz in Maschinen ebenso definiert:

Maschinen sind insoweit intelligent, als ihre Handlungen darauf ausgerichtet sind, ihre Ziele zu erreichen.

Doch weil Maschinen, anders als wir Menschen, keine eigenen Ziele haben, geben wir ihnen diese Ziele vor. Mit anderen Worten: Wir entwickeln Maschinen, die etwas optimieren sollen, geben ihnen Ziele vor und drücken den Startknopf.

Diesen Ansatz finden wir nicht nur auf dem Gebiet der KI. Er zieht sich wie ein roter Faden durch die technologischen und mathematischen Grundsteine unserer Gesellschaft. In der Kontrolltheorie werden Steuerungen für alles Mögliche entwickelt – vom Jumbojet bis zur Insulinpumpe. Dort besteht die Aufgabenstellung für ein System darin, eine *Kostenfunktion* zu minimieren, die angibt, wie stark die Abweichung von einem gewünschten Verhalten ist. In der Wirtschaft gibt es Mechanismen und Richtlinien, die den *Nutzen* Einzelner, das *Wohlergehen* von Gruppen und den *Profit* von Unternehmen

optimiert.¹⁰ In der Planungsforschung werden komplexe logistische und Fertigungsprobleme gelöst. Hierzu wird die erwartete *Summe der Belohnungen* im Laufe der Zeit maximiert. Und in der Statistik sind Lernalgorithmen darauf ausgelegt, eine erwartete *Verlustfunktion* zu minimieren, die die Kosten für Vorhersagefehler definiert.

Dieses allgemeine Schema – ich bezeichne es als *Standardmodell* – ist sehr verbreitet und extrem leistungsfähig. Doch unglücklicherweise *wünschen wir uns keine Maschinen, die auf diese Weise intelligent sind.*

Der Nachteil des Standardmodells wurde 1960 von Norbert Wiener, einem legendären MIT-Professor und führenden Mathematiker der Mitte des 20. Jahrhunderts, klar benannt. Wiener hatte gerade gesehen, wie das Dameprogramm von Arthur Samuel lernte, seinen Schöpfer zu besiegen. In der Folge schrieb er seinen visionären, aber kaum bekannten Artikel *Some Moral and Technical Consequences of Automation* (Ein wenig Moral und die technischen Folgen der Automatisierung).¹¹ Sein Hauptargument lautet:

»Wenn wir zur Erlangung unserer Absichten einen maschinellen Agenten einsetzen, in dessen Handlungen wir nicht effektiv eingreifen können, [...] sollten wir uns absolut sicher sein, dass die der Maschine vorgegebene Absicht wirklich dem Ergebnis entspricht, das wir uns wünschen.«

»Die der Maschine vorgegebene Absicht« ist exakt das Ziel, das die Maschinen im Standardmodell zur Optimierung heranziehen. Wenn wir einer Maschine, die uns an Intelligenz übertrifft, die falsche Absicht vorgeben, setzt sie diese um und wir verlieren. Die Vorgänge in den sozialen Medien sind nur ein Vorgeschmack darauf. Hier optimieren relativ dumme Algorithmen in globalem Maßstab das falsche Ziel. In Kapitel 5 komme ich noch auf einige deutlich schlimmere Folgen zu sprechen.

All das dürfte niemanden wirklich überraschen. Seit Jahrtausenden kennen wir die Gefahren, die lauern, wenn wir genau das bekommen, was wir uns wünschen. In jedem Märchen, in dem drei Wünsche erfüllt werden, dient der dritte Wunsch dazu, die Folgen der ersten beiden rückgängig zu machen.

Es sieht so aus, als wäre unser Marsch in Richtung einer übermenschlichen Intelligenz unaufhaltsam. Kommen wir ans Ziel, könnte dies das Ende der Menschheit besiegeln. Doch noch ist nicht alles verlo-

ren. Wir müssen uns darüber klar werden, wo wir falsch abgebogen sind, und unseren Fehler beheben.

Gibt es Abhilfe?

Das Problem liegt in der grundlegenden Definition der künstlichen Intelligenz. Wir sagen, dass Maschinen insoweit intelligent sind, als ihre Handlungen generell darauf ausgerichtet sind, *ihre* Ziele zu erreichen. Aber wir haben keine zuverlässige Möglichkeit, sicherzustellen, dass *ihre* Ziele *unseren* Zielen entsprechen.

Vielleicht sollten wir vielmehr darauf bestehen, dass die Maschinen nicht *ihre* Absichten verfolgen, sondern *unsere* Absichten? Wenn wir eine solche Maschine jemals entwickeln können, ist sie nicht einfach nur *intelligent*, sondern auch *vorteilhaft* für die Menschen. Probieren wir es mit einer neuen Definition:

Maschinen sind insoweit *vorteilhaft*, als *ihre* Handlungen darauf ausgerichtet sind, *unsere* Ziele zu erreichen.

Das hätte von Anfang an das Motto sein sollen.

Der schwierige Teil besteht natürlich darin, dass unsere Ziele und Absichten in uns liegen (und zwar in allen acht Milliarden von uns, die diesen Planeten in ihrer Vielfalt bevölkern) und nicht in den Maschinen. Dennoch ist es möglich, Maschinen zu konstruieren, die auf genau diese Weise zu unserem Wohl beitragen. Natürlich werden diese Maschinen unsere Absichten nicht ganz genau kennen. Wie sollten sie auch? Schließlich sind wir uns selbst oft genug nicht im Klaren darüber. Aber es wird sich zeigen, dass es sich bei dieser Unsicherheit um ein Feature handelt, nicht um einen Bug (das heißt, es ist an sich eine gute Sache, keine schlechte). Unsicherheit in puncto Ziele und Absichten bedeutet auch, dass Maschinen weiterhin dem Menschen folgen müssen: Sie werden um Erlaubnis fragen, sich korrigieren und auch abschalten lassen.

Indem wir die Voraussetzung streichen, dass Maschinen klare Ziele haben müssen, rütteln wir an den Grundfesten der künstlichen Intelligenz und reißen diese zum Teil ein. Das heißt, wir müssen den Überbau zum großen Teil neu errichten: die Ideen und Verfahren zur Umsetzung von KI. Das Ergebnis wird eine neue Beziehung zwischen Mensch und Maschine sein, die – so hoffe ich – dazu beiträgt, die nächsten paar Jahrzehnte erfolgreich zu meistern.

STICHWORTVERZEICHNIS

2001 (Film) 152

3-D-Druck 82

A

Abbeel, Pieter 81, 205

Abfolgen von Aktionen 273

Abkopplung, große 127

Ablenkungsstrategien 169

Abschalten 262

Absichten 148, 184, 216, 262

Abstimmungsprozess 44

Agent 27, 33

 Aufbau 52

 intelligenter 50, 56, 93, 100

Agenten 73, 85, 201, 206

 Einzel- und Multi- 206

 intelligente 97

 mehrere 36

 rationale 179

Agentenprogramm 56, 65

AI100 161

Aktionen 73, 95, 206, 273

 abstrakte 98

 Aufschlüsseln von 98

 Beschränkungen 248

 Bestandteile 96

 Bibliothek von 96

 Hierarchie 98

 zukünftige 97

Aktionseinheiten 101

Aktionspotenzial 23

Aktivitäten

 geistige 99

Aktoren 81

Alciné, Jacky 68

Alexander, Scott 158, 181

Algebra 283

Algorithmen 42, 71, 204, 301

 Bias 138

 intelligente 141

 Reinforcement Learning 63

 Zielerreichung 98

Algorithmus 45

 Streckensuche 273

Aliens *siehe* Außerirdische

Allgemeine Erklärung der Menschenrechte 117

allgemeine KI 95, 108, 147

Allgemeingültigkeit 286

Allgemeinheit 54

Allgemeinwissen 88

Allgemeinwohl 85

AlphaGo 14, 55, 63, 100, 219, 246, 277

AlphaZero 55

Altruismus 242

 negativer 243

Amazon 82

Amazon Echo 73

Ameisen 33

Analytical Engine 48, 143

analytische Maschine 48

Annahmen

 prüfen 202

Ansehen 133

Anstellung 123

Anweisungen 216

Aoun, Joseph 134

A-posteriori-Wahrscheinlichkeit 62

Apple HomePod 73

A-priori-Wahrscheinlichkeit 62, 205

Arbeitslosigkeit 124

Arbeitsplätze 123

Arglebarglium 88

Argumentationsmaschine 48

Aristoteles 28, 48, 58, 124, 260

Armstrong, Stuart 235

Arnauld, Antoine 29

Arrow, Kenneth 236

Asimov, Isaac 152

Assistance Games 205–206, 218, 262

Assistenten

 Aufgaben 78

 intelligente 78, 110

 intelligente persönliche 76

 persönliche digitale 266

 Unterstützung 78

Atari-Videospiele 64

Atkinson, Robert 170

- Atlas (Roboter) 82
 - Atomkraft
 - Gefahren 169
 - Atomtechnik 264
 - Auf der Suche nach Indien (Buch) 269
 - Aufstieg der Roboter (Buch) 123
 - Augenschein 135
 - Aussagenlogik 58, 284
 - Ausschalter-Problem 209
 - Ausschalter-Spiel 209
 - Auszahlungsfunktion 207
 - Außerirdische 11
 - Auto
 - selbstfahrendes 51, 55, 74, 150, 262, 266
 - Automatisierung 108
 - Profiteure 127
 - Autonomie
 - Fahrzeuge 74
 - menschliche 261, 270
 - Waffensysteme 122
 - Autonomiegrad 75
 - Autor, David 126
 - Avengers – Infinity War (Kinofilm) 238
 - Axiome 198
- B**
- Babbage, Charles 48, 143
 - Backgammon 63, 275
 - Backpropagation 306
 - Baldwin, James 26
 - Baldwin-Effekt 26
 - Banks, Iain M. 176
 - Bayes, Thomas 62
 - Bayes-Inferenz 290
 - Bayes-Netze 62, 286, 291
 - Bayes-Rationalität 62
 - bayessche Logik 62
 - bayessche Netze *siehe* Bayes-Netze
 - Bedeutung
 - kontextabhängige 217
 - bedingungsloses Grundeinkommen 131
 - Befehle 216
 - Belief State 298
 - Belohnungen 62, 115
 - Belohnungsaufschub 25
 - Belohnungsfunktion 63, 204
 - Belohnungssignal 64, 179, 204
 - Belohnungssystem 25, 218
 - Bentham, Jeremy 32, 233, 235
 - Berechnungen 100
 - Berechnungsaktivität 278
 - Berg, Paul 194
 - Bernoulli, Daniel 30
 - Beschäftigung 123
 - Bewegungssteuerung 279
 - Beweise
 - mathematische 198
 - Bewertungsfunktion 223, 276
 - Bewertungssysteme 116
 - bewusste Wahrnehmung 136
 - Bewusstsein 24
 - BGE *siehe* bedingungsloses Grundeinkommen
 - Bias 138
 - Big Data 92
 - Bildererkennung 304
 - Bildersuche 139
 - Bildung 78, 109, 134
 - Bindewörter (Logik) 284
 - Biomedizin 85
 - Blackbox 58
 - Bletchley Park 153
 - Blinzelreflex 65
 - Boole, George 284
 - boolesche Logik 58
 - Bootstrapping 89
 - Borges, Jorge Luis 246
 - Boston Dynamics 81
 - Bostrom, Nick 111, 157, 162, 178–179, 195, 268
 - Bots 73, 116
 - Armeen 116
 - BRETT (Roboter) 81
 - Brin, Sergey 90
 - Brooks, Rodney 179
 - Brynjolfsson, Erik 127
 - Budapest-Konvention 268
 - Bündelung 311
 - Büroklammer-KI 178
 - Büroklammer-Spiel 207
 - Butler, Samuel 144
 - Byron 48

C

Capture The Flag *siehe* CTF
 Cardano, Gerolamo 29
 CDC 6000 13
 Chace, Calum 124
 Chatham-House-Regel 86
 Chollet, François 309
 Chunking 311
 Click-Through-Rate *siehe* Klickrate
 Clinton, Bill 127
 Computer 40, 59, 73, 100
 als Werkzeug des Menschen 141
 mechanische 236
 Computerpannen 141
 Computer-Vision 95, 304
 Content-Algorithmen 115, 148, 259
 Controlling 85
 Convolutional Neural Network 55
 Cortana 51
 Crowdsourcing 119
 CTF
 KI-System spielt 99
 Cybercrime-Konvention des Europarats 268
 Cybersicherheit 199
 Cyborgs 175

D

Dame 275, 277
 Dame (Spiel) 13, 63
 Dartmouth College 12
 Dartmouth-Konferenz 85
 Daten 42
 Verschlüsselung 80
 Datenberge 84
 Datenschutz 79, 85
 Deduktion 28
 Deep Blue 71, 277
 Deep Convolutional Networks 72, 305
 Deep Dreaming 307
 Deep Learning 14, 55, 67, 92–93, 287, 304
 Deepfakes 115
 DeepMind 14, 55, 63, 287
 Defense of the Ancients 64
 Dekohärenz 44
 Delilah (Bot) 115

Denken 99

 logisches 28
 rationales 48
 Denkmaschine 48
 Desinformation 117
 Diazepam *siehe* Valium
 Dickinson, Michael 203
 Dickmanns, Ernst 74
 Die Maschine steht still (Erzählung) 269
 Die sozialen Grenzen des Wachstums (Buch) 244
 Dienstleistungen
 zwischenmenschliche 132, 134
 DNA 26, 167–168
 DNS *siehe* DNA
 Dopamin 25, 219
 Dota 2 64
 Dot-Com-Boom 73
 DQN 63
 Dreifarbenproblem 47
 DSGVO 138, 266
 Dummheit 246
 Durchbrüche
 konzeptionelle 87
 Durchschnittsnutzen 234
 Durianfrucht 250

E

E. coli 22
 ECHO (Smart-Home-Steuerung) 80
 E-Commerce 73
 Edgeworth, Francis 252
 Einkommen 133
 Einkommensverteilung 133
 Eisenhower, Dwight 264
 elektrische Erregung 23
 Elektrofahrzeuge 75
 Eliza 76
 Elster, Jon 257
 Elysium (Film) 137
 Emotionen 136, 246, 248
 Empfehlungsalgorithmen 16
 Endlosschleife 46
 Entropie 45
 Entscheidungen 189, 228, 273, 278
 automatisierte 138
 moralische 228

öffentliche 234
 rationale 275
 Entscheidungsfindung 7, 62, 84, 188,
 217, 276
 Entscheidungshorizont 100
 Entscheidungsproblem 34, 47, 206,
 246, 280
 Entscheidungsqualität 101, 278
 Entwicklungsziele 84
 Epikur 233
 Erehwon (Roman) 144, 171
 Erfahrungen 252, 301
 und Präferenzen 256
 Erfolgswahrscheinlichkeit 61
 Erinnerungen 252
 Erkenntnis 95, 283
 Erlaubnis 262
 Erpressung 115
 Erträge 154
 Ertragsgesetz 154
 Erwartungsnutzen 31, 62
 Erwartungswerte 30
 Erwartungswertregel 30
 Erziehung 134
 Escherichia coli *siehe* E. coli
 E-Sport 64
 Ethik 231
 deontologische 231
 Konsequentialismus 231
 Etzioni, Oren 164, 169
 Evolution 23
 evolutionäre Fitness 25
 Expertensysteme 13
 Explanatation-Based Learning 310
 exponentielle Komplexität 46

F

Fahrzeuge
 autonome 194
 Fake News 118
 Fakten 90, 119
 Faktenchecker 118
 Falschinformation 117
 Feature Engineering 93
 Fehlannahme 199
 Fehler 250
 minimieren 67

Fermat, Pierre de 198
 Ferranti Mark I 42
 Fifth Generation (KI) 287
 Filterblasen 148
 Finanzen 78–79
 Fitness, evolutionäre 27
 Folgen 231
 Ford, Martin 123
 Forderungen 216
 formale Logik 58
 Forster, Edward Morgan 269
 Fortschritt
 technologischer 126
 zivilisatorischer 255
 Freddy (Roboter) 50
 Freeway 64
 Frege, Gottlob 286
 freie Meinungsäußerung 117
 Fruchtfliegen 203
 Full, Bob 203
 Fürsorgefaktor 242

G

Galileo 91
 Garantien 197
 Gates, Bill 64
 Gedankenfreiheit 117
 Gefangenendilemma 38
 Geheimdienste 84, 114
 Gehirn 24, 43, 99, 218
 Geist
 Funktionsweise 22
 Geld 152
 Geminoid DK (Roboter) 135
 gemischte Strategie *siehe* Zufallsstra-
 tegie
 Gemütszustände 249
 Genderbias 139
 Genschere 168
 Genuss 132
 Geschicklichkeit 82
 Geschwindigkeit 45
 Gesetze 230
 Gesichtserkennung 121
 gesunder Menschenverstand 88
 Gesundheit 78, 109
 Gewinnmaximierung 149

Gewissheit 29
 Global Learning XPRIZE 79
 Glück 110, 233, 236
 Glukose 22
 Go 14, 53, 55, 63, 100, 246, 275, 302
 Aussagenlogik 285
 Gödel, Kurt 59
 Goethe 148
 GOFAI 287
 Good Old-Fashioned AI 287
 Good, Irving John 153, 221
 Goodharts Gesetz 117
 Goodman, Nelson 94
 Google 68
 Google Home 73
 Gorilla 143
 Google 68
 Gorilla-Problem 143, 147, 161, 167
 Gravitationswellen 91
 Graviton 92
 Greifen 82
 Grice, H. Paul 218
 Großer fermatscher Satz 198
 Ground Truth 119
 Grundeinkommen 131
 Grundkonzepte moderner KI-
 Systeme 7

H

Haftung 230
 HAL *siehe* 2001 (Film)
 Halteproblem 45
 Handeln, unvorhersehbares 37
 Handlungen
 Komplexität 279
 Handlungsempfehlungen 105
 Handschrifterkennung 54–55
 Hardin, Garrett 39
 Hardware 42
 Harop (Waffe) 121
 Harsanyi, John 233, 243
 Hassabis, Demis 287, 309
 Hassrede 118
 Haushaltsroboter 226, 266
 Hawking, Stephen 12
 Hedonimeter 252
 Hedonismus 252

hedonistisches Kalkül 32
 Helferspiele 205–206
 Herbert, Frank 146
 Heterogenität 226
 Hierarchien 281
 Hillarp, Nils-Åke 25
 Hinton, Geoff 306
 Hirsch, Fred 244
 Hobbes, Thomas 261
 Humanik 134
 Hume, David 178, 304
 Hypothesen 93

I

IBM 71
 imaginäre Welten 198
 Imitation Game 49
 Inceptionism 307
 industrielle Revolution 124
 Inferenz 153
 Infokalypse 118
 Informatik 42
 Fehlannahmen 199
 Informationen
 Wahrheitsgehalt 119
 Insekten 33
 Inspiration 95
 intelligente persönliche Assistenten
 76
 Intelligenz 7, 88, 159
 allgemeine 283
 Definition 17, 22
 übermenschliche 143
 Intelligenzexplosion 153, 221
 Intelligenzquotient 56
 Internet der Dinge 73
 Internet of Things *siehe* Internet der
 Dinge
 Intuition 95
 IoT *siehe* Internet der Dinge
 IQ 56
 IRL *siehe* Reinforcement Learning, in-
 verses

J

Jeopardy! (Fernsehquiz) 88
 Jevons, William Stanley 236

JiaJia (Roboter) 135
 jian ai 233
 Jiankui, He 168
 Jie, Ke 14
 Junktoren 284

K

Kahneman, Daniel 252
 Kakerlaken 204
 Kalte-Hand-Experiment 253
 Kasparow, Garri 71, 277
 Kelly, Kevin 105, 160
 Kenny, David 165, 175
 Kernphysik 14
 Kernreaktor 15
 Kettenreaktion
 nukleare 15, 86
 Keynes, John Maynard 124, 133
 KI 62, 67, 79, 95, 114
 Ablenkung 157
 Abschalten 151, 173, 187
 Abschotten 173
 Ad-hoc-Lösung 157
 allgemeine 54
 Allwissenheit 105
 Angriffe auf die 162
 Anwendungen 108
 Ausfallarten beim autonomen
 Fahren 151
 Auswirkungen auf die Wissen-
 schaften 106
 Bedrohung durch die 144, 184
 Bestandteile 225
 datengestützte 93
 Debatte 157
 dem Menschen ebenbürtige 83,
 147, 304
 Diskussion 166
 Durchbrüche 72
 Emotionen 177
 Entwicklung 74, 98
 erste Anwendungen 73
 ethische Grundsätze 264
 Forschung 72
 Fortschritte 130
 Gefahren 146, 154, 158, 195
 gegen KI 268
 Geschichte 12
 Gesetze 266
 Grenzen 110
 Identitätsvortäuschung 266
 im Alltag 110
 im Unterricht 109
 Kernfragen 277
 Kommerzialisierung 72
 Kontrolle 264
 Kontrollproblem 154
 Kontrollverlust 153
 künftige Anwendungen 73
 Lenkung 264
 Logik 284
 Loyalität 229
 Missbrauch 113, 267
 moderne 50
 Multiplikatoreffekt 108
 Nachteile 73
 nachvollziehbare Entscheidun-
 gen 266
 nachweislich vorteilhafte 197,
 251, 261, 266
 neuer Ansatz 263
 neues Fundament 197
 Nutzen 195
 One Hundred Year Study 161
 Orakel 173
 Politik 195
 praktische Anwendungen 72
 Probleme 87
 Regulierung 267
 Retter der Menschheit 153
 schwache 54
 Sicherheit 195, 266
 Skalieren 102
 Standardmodell 67, 262
 Sünde 191
 superintelligente 10, 85, 101
 Trojanische Pferde 77
 und Weltherrschaft 195
 Unmöglichkeit 161
 utilitaristische 231
 Verbot 146
 Verlangen 177
 Verleugnung 157
 Verschlüsselung 80
 Vorschriften 266
 Vorteile für den Menschen 107

Weiterentwicklung 71
 Weltherrschaft 149
 Winter der KI 13
 Wissenschaftler 106
 Ziele 183
 Ziele und Teilzeile 151
 zu Kriegszwecken 120
 Zukunft 71
 KI-Forscher 86, 165, 264
 KI-Forschung 146
 Killerroboter 120
 KI-Systeme 54, 98, 101
 Leistungsfähigkeit 84
 Kitkit School 79
 KI-Unternehmen 73
 Klassifikationsalgorithmus 68
 Klickrate 16, 65, 150, 257
 Klimawandel 239
 Knappheit 244
 Kochen 81
 kognitiv 24
 Kolibakterium *siehe* E. coli
 Kommunikation 40
 Kompensationseffekte 124
 Komplexität 34, 46
 exponentielle 46
 kombinatorische 275
 lineare 46
 Kompromiss 29, 227
 König-Midas-Problem 147, 167
 Konsequentialismus 231
 Konsequenzen 237, 259
 Attraktivität 237
 Kontrolle 113, 191, 261
 von Menschen 85
 Kontrollproblem 154, 182, 195
 Kontrolltheorie 13, 17, 188
 Kontrollverlust 147
 Konzepte 94
 Kooperationsprinzip 218
 Kostenfunktion 17, 57
 Krebs 149
 Krugman, Paul 127
 Kultur 270
 Kultur-Zyklus (Buchreihe) 176
 künstliche Intelligenz *siehe* KI
 Kurzweil, Ray 175

L
 Lagebilder 84
 Laplace, Pierre-Simon 62
 Lautsprecher
 smarte 76
 LAWS *siehe* Lethal Autonomous Weapons
 Leben 3.0 (Buch) 124, 149
 Lebensstandard 131
 LeCun, Yann 55, 177
 Legalität 230
 Lernalgorithmen 79
 Lernen 23, 301
 anhand von Beispielen 301
 datengetriebenes 92
 kumulatives 91
 symbolisches 93
 Lernfähigkeit *siehe* Lernen
 Lernprobleme 94
 Lernprozess
 selbstverstärkender 90
 Lernsysteme
 intelligente 94
 Lerntheorie 304
 Lesefähigkeiten 83
 Lethal Autonomous Weapons 120
 lineare Komplexität 46
 Lkw
 fahrerlose 129
 Lkw-Fahrer 129
 Lloyd, Seth 45, 246
 Lloyd, William 39
 Llull, Ramon 48
 Lochkarten 73
 Lodge, David 9
 Logik 48, 283
 Aussagenlogik 58
 boolesche Logik 58
 formale 58
 in der KI 284
 Logikgatter 284
 Logistik 82
 Lohnsteigerungen 127
 Lookahead-Algorithmen 57
 Lookahead-Suche 55, 57, 63, 75, 100,
 223, 277
 Lösungssuche 274

Lovelace, Ada 48, 143

Loyalität 241

M

Machine Learning 41, 92

Bias 138

Feature Engineering 93

Macht 85

Maler

Automatisierung 125

Manipulieren 82

Markov-Entscheidungsprozesse 204

Maschine-Mensch-Interaktion 262

Maschinen 41

altruistische 185

demütige 187

intelligente 40, 73, 94, 153, 160,

183, 210, 244

Lernen am Beispiel 193

Loyalität 229

Prinzipien vorteilhafter 184

superintelligente 101, 143, 146,

158, 173, 261

übermenschliche 85

unterwürfige 262

vorteilhafte 184, 210, 247, 262

vorteilhafte intelligente 271

Ziele 191

Zweck 223

Maschinenbewusstsein 25

Maschinenethik 190

Maschinenintelligenz 146, 154, 165

Massenvernichtungswaffen 122

Mathematik 41, 147, 283

Matrix (Kinofilm) 236, 249

MavHome 80

McAfee, Andrew 127

McCarthy, John 12, 58, 74, 86

Meinungäußerung

freie 117

Meinungsfreiheit 117

Mensch gegen Maschine 99

Menschen

als Werkzeug des Computers 141

Vorteile gegenüber Maschi-
nen 134

Menschenrechte 117

Menschenverstand

gesunder 78

Menschlichkeit 132, 134

Menschmaschine 175

Mensch-Maschine-Interaktion 262

Metapräferenzen 257

Metareasoning 278

Methoden der Ethik (Buch) 238

Mill, John Stuart 231, 233

Minderheiten 140

Minsky, Marvin 12, 85

Misstrauen 118

Modell 38

Monotonie 31

Moore, G. E. 233, 235

mooresches Gesetz 43, 87

Moral 190

Moraltheorien 240

Moravec, Hans 155

Morgan, Conwy Lloyd 26

Morgenstern, Oskar 31

Mozi 233

Multi-Agenten-Kooperation 103

Musk, Elon 169, 176

Myeong-hun, Choe 275

N

Nachhaltigkeit 39, 84

Nächstenliebe 233

Nash, John 38, 208

Nash-Gleichgewicht 38, 208

Nationalökonomie 241

Navigationssysteme 273

Neid 111

NELL 90

Nervennetze 23

Netze

neuronale 286, 305

Netzwerkdurchsetzungsgesetz 118

neurale Borte 176

neuronaler Staub 176

Neuronen 23

Newell, Allen 311

Newton 91

Ng, Andrew 163

Nichtwissen 61

Normen 230

- Norvig, Peter 10, 71
 Nozick, Robert 237
 Nudge (Buch) 258
 Nudging 259
 Nullsummenspiel 244
 Nutzen 30, 233
 abwägen 236
 Nutzenfunktion 61
 Nutzenmonster 237
 Nutzentheorie 50
 Nutzenwert 31
- O**
- Objekterkennung 14, 55, 67
 Ockhams Rasiermesser 301
 Odysseus 257
 onebillion 79
 Onlinebanking 128
 Onlinehandel 128
 Onlinemarktplätze 116
 OpenAI 64
 Orakel-KI 173, 220
 Organismus
 adaptiver 26
 instinktgesteuerter 26
 Orthogonalitätsthese 179
 Ortswechsel (Buch) 9
 Outsourcing 129
 Ovadya, Aviv 118
- P**
- Panama 26
 Parfit, Derek 239
 Pascal, Blaise 30, 48
 PDA 51
 Pearl, Judea 50, 62, 291
 persönlicher digitaler Assistent *siehe*
 PDA
 Pflegeberufe 132
 Piano, logisches 236
 Picking Challenge (Amazon) 82
 Pinker, Steven 170, 177, 180
 Pläne 97
 Planung 217, 280
 Planungsforschung 18, 188
 Planungssysteme 50
 Politik 263
- Pong 64
 Popper, Karl 235
 Positionsgüter 244
 Prädikatenlogik 59, 284, 286
 Präferenzautonomie 252, 256, 259
 Präferenzen 31–32, 185, 202–203,
 247, 249, 270
 aktualisierte 256
 Änderung 270
 Änderung von 258
 Entstehung 255
 Formbarkeit 202
 Gewichtung 233
 menschliche 188
 Metapräferenzen 257
 sakrosankte 257
 umsetzen 258
 unanastbare 257
 und Erfahrungen 256
 veränderte 256
 Wandel 255
 Präferenzstruktur 227
 Präferenzutilitarismus 234, 243
 Pragmalinguistik *siehe* Pragmatik
 Pragmatik 217
 Price, Richard 62
 Prinzipien vorteilhafter Maschinen
 184
 Privatsphäre 79, 85
 Problem
 NP-vollständig 47
 unentscheidbares 45
 Problemlöser
 logische 50
 Problemlösungen 274
 Produktivitätssteigerungen 127
 Programme 41, 263
 Effizienz 100
 Programmiersprachen 42
 probabilistische 62, 294
 Programmierung
 dynamische 63
 induktive logische 95
 logische 287
 probabilistische 93, 295
 Prolog (Programmiersprache) 287
 Putin, Wladimir 194

Q

Qualitätsmaßstab 117
 Qualle 24
 Quantenbit *siehe* Qubit
 Quantencomputer 43
 quantenmechanische Wellenfunktionen 43
 Quantenphysik 43
 Quantentheorie 45
 Qubit 43

R

Racial Bias 266
 randomisierte Strategie *siehe* Zufallsstrategie
 Rationalität 28, 35, 189, 246
 Realität 198
 Reallöhne 127
 Rechenleistung 43, 45
 Grenzen 45
 Rechenmaschinen
 universelle 48
 Rechte
 individuelle 233
 Redefreiheit 118
 Reflexagenten 65
 Reflexe 65
 Regelungstheorie *siehe* Kontrolltheorie
 Regierungen 265
 Kontrolle 116
 Reinforcement Learning 25, 55, 63, 75, 82, 98, 115, 150, 179, 219, 236, 306
 inverses 204, 255
 Reize 25
 Relativitätstheorie
 Gravitationswellen 91
 Ressourcen 237
 Richtlinie 63
 Robinson, Alan 13
 Roboter 80, 85, 98, 109, 126
 Algorithmen 82
 als elektronische Person 137
 Bürgerrechte 136
 Fähigkeiten 82
 Haushalt 83

Intelligenz 136
 KI-gesteuerte 74
 kochen 81
 lebenssechte 135
 menschliche Form 135
 Pflege 83
 Präferenzen 226
 Smart Home 81
 wandlungsfähige 227
 Zweck 223

Robotergesetze 152
 robotergesteuerte Prozessautomatisierung 129
 Rochester, Nathaniel 12
 RPA 129
 Rutherford, Ernest 14, 86, 162

S

Sachs, Jeffrey 244
 Salomons, Anna 126
 Samuel, Arthur 13, 63, 277
 Sargent, Tom 204
 Satellitenaufnahmen 84
 Satellitendaten 84
 Satz von Bayes 62
 Schach 55, 63, 71, 275
 ELO 71
 Rating 71
 Schachprogramm 51
 Schaltkreise 59
 Schätzfunktion 63
 Schätzwert 277
 Schiebepuzzle 50, 274
 Schließen
 unsicheres 50
 Schlupflochprinzip 215, 230, 235
 Schlussfolgerungen 58, 88, 278
 logische 50
 probabilistische 77
 Schnelles Denken, langsames Denken (Buch) 252
 Schnitzeljagd (Buch) 9
 Schule 89
 Schutz 117
 Schwab, Klaus 127
 Schwierigkeit 46
 Sedol, Lee 14, 55, 100, 246, 275, 277

- Selbst
 - erinnerndes 252
 - erlebendes 252
- Selbstbedienungskassen 128
- Selbsterhaltung 152, 209
- Selbstlosigkeit 234, 242
- Selbsttäuschung 221, 236
- Semantik 90
- Shakey (Roboter) 50, 60
- Shannon, Claude 12, 71
- Shiller, Robert 127
- Shōgi 55, 63
- sichere Mehrparteienberechnungen 80
- Sicherheit 117
 - selbstfahrende Autos 75
- Sidgwick, Henry 238
- Signalsystem 25
- Simon, Herbert 85, 95, 281, 311
- Sinclair Cambridge (Taschenrechner) 13
- Siri 51
- Skalieren 102
- Slate Star Codex 158, 181
- Slaughterbot 121
- Smart City 85
- Smart Home 80
 - Steuerung 81
- Smart, R. N. 235
- smarte Lautsprecher 73
- Smartphone 73
 - Assistenten 76
- SMC 80
- Smith, Adam 241
- Social Media 16, 65, 115, 148, 257
- Softbot 73
- Software 45, 263
- Softwareagenten 109
- Somalia-Problem 241
- Sophia (Roboter) 136
- Sozialberufe 132
- soziale Medien *siehe* Social Media
- Sozialgemeinschaft 85
- Sozialstatus 133
- Sozialwahltheorie 236
- Sozialwissenschaften 227
- Space Invaders 64
- Speicher 42
- Spence, Mike 127
- Spieltheorie 36, 206
- Spionage 115
- SpotMini (Roboter) 81
- Sprachassistenten 76
 - Schwächen 76
- Sprache 88
 - Verständnis 88
- Spracherkennung 14, 55, 67, 79, 114
- Sprachsteuerung 76
- Sprachsynthese 115
- Sprachverständnis 83, 90
- Spurverfolgungsfehler 66
- Stadtplanung 85
- Stammesdenken 162, 171
- Standardmodell 18, 56, 184
- Standardverfahren
 - Verzicht auf 64
- StarCraft 53
- Stasi 113
- Statik 199
- Stationarität 32
- Statistik 13
- Stau 75
- Steinberg, Saul 97
- Steuern 131
- Steuerung 17
- Stimulation 219
- Stolz 111
- Strategie 37
- Streben 132
- Suchbaum 100
- Suchmaschine 102
- Summers, Larry 127, 130
- Summit 42
- Sunstein, Cass 258
- Superintelligenz 85, 101, 154, 161, 261
 - Grenzen 105
 - Unvorhersehbarkeit 86
- Superintelligenz (Buch) 111, 157, 162, 179, 195
- Superkräfte 84
- Supervised Learning *siehe* überwachtes Lernen
- Sutherland, James 80
- Synapsen 23, 43

Synergien 78
 Systeme
 intelligente 248, 273, 308
 Lernen 301
 logische 61
 regelbasierte 293
 wissensbasierte 293
 Szilárd, Leó 15, 86, 162

T

Take-off
 schneller 155
 Tatsache, Fiktion, Voraussage (Buch)
 94
 Tatsachen 119
 Teams aus Mensch und Maschine 175
 Teamwork 64
 Technokratie 116
 Technologie 141
 technologische Arbeitslosigkeit 124
 Tegmark, Max 12, 124, 149
 Telefonate 84
 Telefonüberwachung 114
 Tellex, Stefanie 82
 Tensor Processing Units 41, 43
 Tesauro, Gerry 63
 Textverständnis 83
 Thaler, Richard 258
 Thales von Milet 91
 The Economic Singularity (Buch) 124
 The Second Machine Age (Buch) 127
 Theorem 198, 283
 für vorteilhafte Maschinen 200
 Theorie der feinen Leute (Buch) 244
 Thornton, Richard 144
 Tod 209
 Tool-KI *siehe* KI, schwache
 Tote holen keinen Kaffee 151, 178
 TPU *siehe* Tensor Processing Units
 Tragik der Allmende 39, 270
 Tragwerksanalyse 199
 Trainingsprozess
 unvollkommener 223
 Transcendence (Kinofilm) 11, 152
 Transitivität 31, 35
 Transkribieren 84
 Tugendethik 231
 Turing 48

Turing, Alan 41, 135, 145, 160
 Turingmaschine 41
 Turing-Test 49
 Tutorensysteme 78

U

Übergangslösungen 130
 Überhypothese 94
 Überredung 113
 Übersetzung 14
 maschinelle 66
 überwachtes Lernen 66, 75, 301
 Überwachung 113
 Überwachungskameras 114
 Umgebung 64, 73, 201
 Umweltverschmutzung 75
 Unabhängigkeit 290
 unentscheidbares Problem 45
 Universalität 41
 universelle Turingmaschine *siehe* Turingmaschine
 Unsicherheit 29, 62, 188, 213, 250, 289
 Unsichtbares sehen 298
 Unterprogramme 263
 Utilitarismus 228, 233
 idealer 233
 Utilitarismus (Buch) 231

V

Valium 26
 Vardi, Moshe 215
 Veblen, Thorstein 244
 Veränderungen 255
 Verbände 265
 Verbesserung
 schrittweise 90
 Verbote 215
 Vergnügen 236
 Verhalten 203, 218, 258
 menschliches 247
 Verhaltensänderung 150, 258
 Verhaltensempfehlung 38
 Verhaltenskontrolle 114
 Verhaltensregel 66
 Verkehrsleitung 85
 Verkehrstote 74
 Verknüpfungsglieder 284

Verleugnung 158
 Vernunft 28, 35, 78
 Verschlüsselung 80
 Verschränkung
 Quantenphysik 44
 Verstand 78
 verstärkendes Lernen *siehe* Reinforcement Learning
 Vertrauen 118
 Videospiele 64
 Videoüberwachung 114
 virtuelle Realität 110
 Vollbeschäftigung 131
 Vollständigkeitssatz 59
 von Neumann, John 31
 Vorhersagegenauigkeit
 maximieren 67
 Vorhersagemodelle 84, 190
 Vorhersagen 105
 Vorhersageprobleme 94
 Vorlieben 213
 Vorlieben *siehe* Präferenzen
 Vorschriften 230
 Vorsicht 194
 Vorwissen 93
 VR 110

W

Waffen, autonome 120
 Wahlmöglichkeiten 250
 Wahrheit 117
 Wahrnehmung 65, 73, 178, 278–279
 menschliche 247
 Wahrscheinlichkeit 29, 62, 289
 Wahrscheinlichkeitsrechnung 13
 Wahrscheinlichkeitsschätzungen 67
 Wahrscheinlichkeitstheorie 289
 Wahrscheinlichkeitsverteilungen 205
 WALL·E (Film) 270
 Wandel 255
 Warcraft 64
 Watson (IBM) 88
 Wert
 empfunderer 244
 wahrgenommener 244
 Werte 190
 Werteausrichtung 148
 Wertemaßstab 237

Wertesysteme 226
 Wertigkeit 244
 Wetten 30
 Whataboutism 169
 Whitehead, Alfred North 96
 Widersprüche 35
 Wiedersehen in Howards End (Buch) 269
 Wiener, Norbert 18, 147, 216
 Wilczek, Frank 12
 Wiles, Andrew 198
 Willensschwäche 257
 Winter der KI 13
 Wireheading 218
 Wissen 58, 88, 283
 explizites 92
 Fakten 88
 wissensbasierte Systeme 58
 Wissensbasis 93
 Wissensdatenbank 93
 Wissenssysteme 85
 Wohlbefinden *siehe* Wohlergehen
 Wohlergehen 234, 241, 254
 Wünsche 18
 Würde des Menschen 137
 Wüstenplanet (Buch) 146

Z

Zauberlehrling 148
 Zeichenerkennung 55
 Zeilendrucker 73
 Zeit 255
 Ziele 56, 148, 177, 184, 209, 216, 262, 279
 kurzfristige 248
 maßgebliche 152
 übergeordnete 153
 widersprüchliche 180
 Zivilisation 96, 255, 261, 269
 Zuckerberg, Mark 169
 Zufallsstrategie 37
 Zukunft 97
 Zustände
 Quantenphysik 44
 Zustandsabfolgen 61
 Zustandsänderungen 41
 Zwergfaultier 26