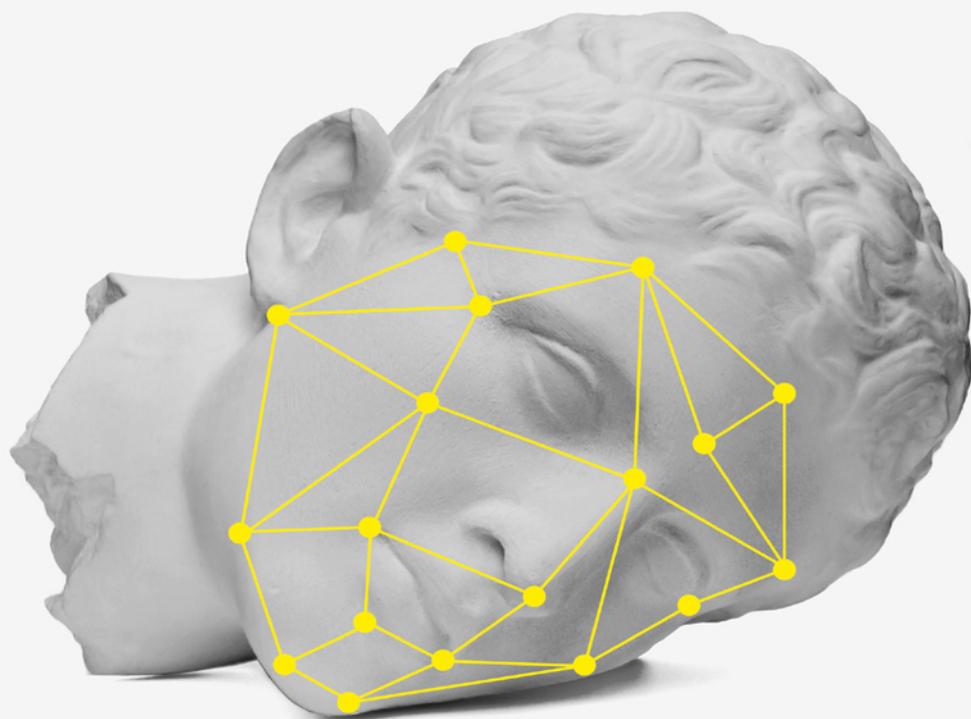


STUART RUSSELL

HUMAN COMPATIBLE



Künstliche Intelligenz

und wie der Mensch die Kontrolle über
superintelligente Maschinen behält





Hinweis des Verlages zum Urheberrecht und Digitalen Rechtemanagement (DRM)

Der Verlag räumt Ihnen mit dem Kauf des ebooks das Recht ein, die Inhalte im Rahmen des geltenden Urheberrechts zu nutzen. Dieses Werk, einschließlich aller seiner Teile, ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und Einspeicherung und Verarbeitung in elektronischen Systemen.

Der Verlag schützt seine ebooks vor Missbrauch des Urheberrechts durch ein digitales Rechtemanagement. Bei Kauf im Webshop des Verlages werden die ebooks mit einem nicht sichtbaren digitalen Wasserzeichen individuell pro Nutzer signiert.

Bei Kauf in anderen ebook-Webshops erfolgt die Signatur durch die Shopbetreiber. Angaben zu diesem DRM finden Sie auf den Seiten der jeweiligen Anbieter.

Neuerscheinungen, Praxistipps, Gratiskapitel,
Einblicke in den Verlagsalltag –
gibt es alles bei uns auf Instagram und Facebook



[instagram.com/mitp_verlag](https://www.instagram.com/mitp_verlag)



[facebook.com/mitp.verlag](https://www.facebook.com/mitp.verlag)

Für Loy, Gordon, Lucy, George und Isaac

Human Compatible

Künstliche Intelligenz

und wie der Mensch die Kontrolle über
superintelligente Maschinen behält

Stuart Russell

*Übersetzung aus dem Englischen
von Guido Lenz*

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <<http://dnb.d-nb.de>> abrufbar.

ISBN 978-3-7475-0174-0

1. Auflage 2020

www.mitp.de

E-Mail: mitp-verlag@sigloch.de

Telefon: +49 7953 / 7189 - 079

Telefax: +49 7953 / 7189 - 082

Authorized German translation from the English language edition, entitled HUMAN COMPATIBLE: Artificial Intelligence and the Problem of Control, ISBN 978-0-52555861-3

Copyright © 2019 by Stuart Russell

Original English language edition published by VIKING, an imprint of Penguin Random House LLC, penguinrandomhouse.com. All rights reserved.

© 2020 mitp Verlags GmbH & Co. KG, Frechen

Dieses Werk, einschließlich aller seiner Teile, ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Lektorat: Janina Bahlmann

Sprachkorrektur: Sibylle Feldmann

Fachkorrektur: Manuela Lenzen

Original cover design: Richard Green, adapted by Christian Kalkert

Bildnachweis: © [manx_in_the_world / istockphoto.com](https://www.istockphoto.com)

Satz: III-satz, www.drei-satz.de

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.«

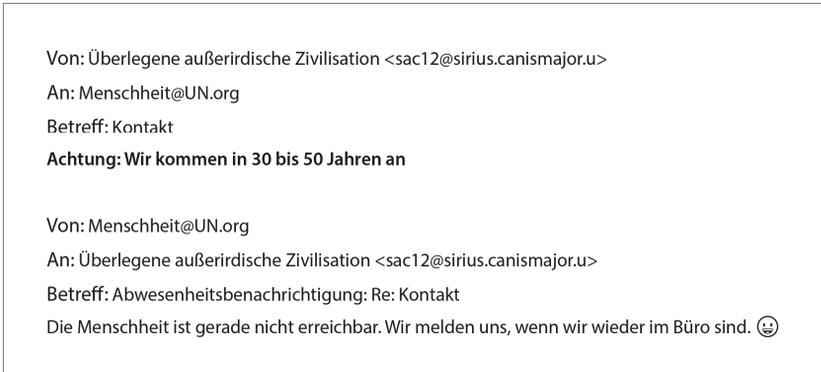


Abb. 1.1: Vermutlich nicht die Antwort, die wir einer überlegenen Zivilisation von Außerirdischen nach deren Kontaktaufnahme senden würden

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]).

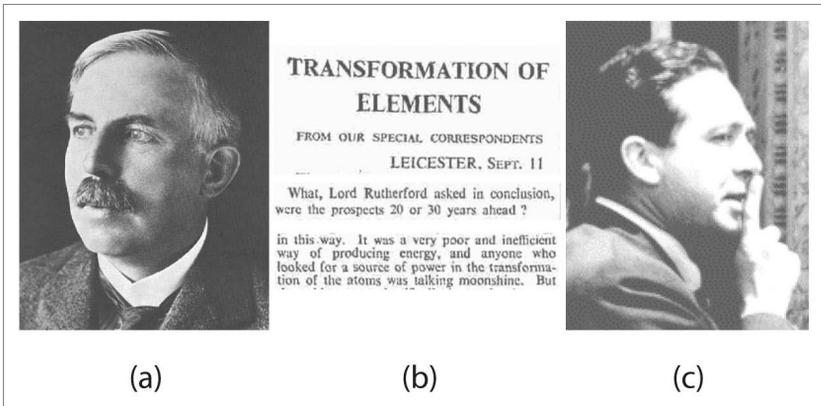


Abb. 1.2: (a) Lord Rutherford, Kernphysiker, (b) Auszug aus einem Bericht der *Times* vom 12. September 1933 über eine Rede, die Rutherford am Vorabend gehalten hatte, (c) Leó Szilárd, Kernphysiker

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 lösbares« 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.

NATÜRLICHE UND KÜNSTLICHE INTELLIGENZ

Wenn wir in eine Sackgasse geraten, drehen wir um und suchen nach der Stelle, an der wir falsch abgebogen sind. Ich habe die Behauptung aufgestellt, dass das Standardmodell der KI, bei dem die Maschinen ein festes Ziel optimieren, eine solche Sackgasse ist. Dabei besteht das Problem nicht darin, dass wir bei der Entwicklung von KI-Systemen *keinen guten Job* machen, sondern vielmehr darin, dass wir ihn möglicherweise *zu gut* machen. Unsere Definition für den Erfolg einer KI an sich ist falsch.

Gehen wir also bis zum Anfang zurück und versuchen wir zu verstehen, wie unser Konzept der Intelligenz überhaupt entstanden ist und wie es auf Maschinen übertragen wurde. Auf diesem Weg finden wir vielleicht eine bessere Definition für ein gutes KI-System.

Intelligenz

Wie funktioniert das Universum? Wie entstand das Leben? Und wo sind eigentlich meine Schlüssel? Das alles sind grundlegende Fragen, über die sich das Nachdenken lohnt. Aber wer stellt diese Fragen? Wie beantworte ich sie? Wie können ein paar Pfund pink-graue Masse namens Gehirn eine Welt unvorstellbarer Weite überhaupt wahrnehmen, verstehen, vorhersagen und manipulieren? Über kurz oder lang widmet sich unser Geist der Erforschung seiner selbst.

Wir versuchen schon seit Jahrtausenden, die Funktionsweise unseres Geistes zu ergründen. Zunächst ging es darum, unsere Neugier zu befriedigen, uns selbst besser zu organisieren, andere zu überzeugen, oder ganz pragmatisch darum, mathematische Probleme zu untersuchen. Doch jeder Schritt auf dem Weg zu einer Erklärung der Funktionsweise unseres Geistes ist auch ein Schritt hin zur Realisierung geistiger Fähigkeiten in einer Maschine, ein Schritt hin zur künstlichen Intelligenz.

Bevor wir verstehen können, wie man Intelligenz erschafft, müssen wir wissen, was Intelligenz eigentlich ist. Die Antwort liefern weder IQ- noch Turing-Tests. Sie liegt stattdessen in einer einfachen Beziehung zwischen dem, was wir wahrnehmen, dem, was wir uns wünschen, und dem, was wir tun. Grob gesagt, ist eine Entität in dem Maße intelligent, wie ihre Handlungen auf Grundlage ihrer Wahrnehmungen zur Erfüllung ihrer Wünsche taugen.

Evolutionäre Ursprünge

Nehmen wir ein simples Bakterium wie *Escherichia coli*. Es verfügt über etwa ein halbes Dutzend Geißeln – lange, haarähnliche Tentakel, die im oder entgegen dem Uhrzeigersinn rotieren. (Allein diese Rotation ist eine erstaunliche Sache, sie zu erläutern, würde aber den Rahmen dieses Kapitels sprengen.) *E. coli* schwimmt in seinem flüssigen Zuhause herum, also zum Beispiel in Ihrem unteren Darm. Es kann die Geißeln zu einer Art Propeller verknüpfen und sich in gerader Richtung fortbewegen. Zwischendurch löst sich diese Verknüpfung und es verharrt taumelnd an Ort und Stelle. Durch diese Kombination aus Schwimmen und Taumeln kann sich *E. coli* herumbewegen und Glukose finden und aufnehmen, statt an einem Ort zu verharren und zu verhungern, wenn die Glukose dort aufgebraucht ist.

Wenn das die ganze Geschichte wäre, würden wir *E. coli* nicht für besonders intelligent halten. Denn das Verhalten des Bakteriums wäre zufällig, es würde keine Entscheidungen treffen, sondern nur ein von der Evolution vorgegebenes Verhaltensprogramm ausführen. Aber es geht weiter: Sobald *E. coli* eine ansteigende Glukosekonzentration wahrnimmt, wird die Taumelphase verkürzt und die Schwimmphase verlängert. Bei abnehmender Glukosekonzentration gilt das Gegenteil. Was es tut (Richtung Glukose schwimmen), ist geeignet, um zu errei-

chen, was es will (mehr Glukose), und zwar aufgrund seiner Wahrnehmung (Anstieg der Glukosekonzentration).

Sie denken jetzt vielleicht: »Aber auch das hat die Evolution programmiert! Wo ist das ein Zeichen für Intelligenz?« Das ist ein gefährliches Argument, denn die Evolution hat auch den grundlegenden Entwurf für Ihr Gehirn in Ihren Genen hinterlegt. Und ich vermute stark, dass das für Sie kein Grund wäre, Ihre eigene Intelligenz infrage zu stellen. Wichtig ist, dass die Evolution in den Genen einen Mechanismus bereitgestellt hat, durch den das Verhalten von *E. coli* von der Wahrnehmung seiner Umgebung abhängt. Die Evolution weiß ja nicht im Voraus, wo sich die Glukose befinden wird oder wo Sie Ihre Schlüssel hingelegt haben. Also hat sie dem Organismus die Fähigkeit zum Suchen mitgegeben.

Natürlich ist *E. coli* kein Genie. Soweit wir wissen, kann es sich nicht daran erinnern, wo es war: Wenn es von A nach B schwimmt und dort keine Glukose findet, kehrt es mit einer gewissen Wahrscheinlichkeit zu Punkt A zurück. Konstruieren wir eine Umgebung, in der jede Glukosespur zu einer Phenolansammlung führt (ein Gift für *E. coli*), folgt das Bakterium diesen Spuren weiterhin. Es lernt nicht. Es hat kein Gehirn, ein paar simple chemische Reaktionen organisieren sein Verhalten.

Das Aufkommen des *Aktionspotenzials* war da ein gewaltiger Sprung nach vorn. Dabei handelt es sich um eine Form elektrischer Erregung, die sich vor etwa einer Milliarde Jahren erstmals in Einzellern entwickelt hat. Spätere Mehrzeller entwickelten spezialisierte Zellen namens *Neuronen*, die das elektrische Aktionspotenzial zur raschen Signalübertragung im Organismus nutzen. »Rasch« bedeutet hier übrigens 120 Meter pro Sekunde oder 430 Kilometer pro Stunde. Die Verbindungen zwischen den Neuronen werden als *Synapsen* bezeichnet. Die Stärke einer synaptischen Verbindung bestimmt, wie groß die von einem zum anderen Neuron fließende elektrische Erregung ist. Tiere lernen, indem sie die Stärke dieser synaptischen Verbindungen verändern.¹ Die Lernfähigkeit ist ein gewaltiger evolutionärer Vorteil, denn so kann sich das Tier an die unterschiedlichsten Umstände anpassen. Gleichzeitig wird dadurch auch das Tempo der Evolution selbst beschleunigt.

Ursprünglich bildeten die Neuronen *Nervennetze*, die im gesamten Organismus verteilt waren. Sie dienten der Koordination von Tätigkei-

ten wie Nahrungsaufnahme und Verdauung oder zur zeitlich abgestimmten Kontraktion von Muskelzellen. Die anmutige Fortbewegung der Qualle ist das Ergebnis eines solchen Nervennetzes. Quallen haben kein Gehirn.

Gehirne entstanden erst später gemeinsam mit komplexen Sinnesorganen wie Augen und Ohren. Einige Hundert Millionen Jahre nachdem die Qualle und ihre Nervennetze entstanden, betraten wir Menschen mit unseren großen Gehirnen die Bühne. Im Gehirn stecken Einhundert Milliarden (10^{11}) Neuronen und eine Billiarde (10^{15}) Synapsen. Im Vergleich zu elektronischen Schaltkreisen »taktet« oder »schaltet« es mit wenigen Millisekunden für jede Zustandsänderung extrem langsam, aber in der Gruppe der biologischen Prozesse nimmt es damit einen Spitzenplatz ein. Das menschliche Gehirn wird von seinen Besitzern häufig als »das komplexeste Ding im Universum« beschrieben. Das ist vermutlich nicht korrekt, aber es liefert eine gute Entschuldigung für unser beschränktes Wissen um seine Funktionsweise. Die Biochemie von Neuronen und Synapsen und auch der anatomische Aufbau des Gehirns sind gut erforscht. Doch was die neuronale Umsetzung auf *kognitiver* Ebene betrifft – Lernen, Wissen, Erinnern, Argumentieren, Planen, Entscheiden usw. –, tapen wir größtenteils im Dunkeln.² (Vielleicht ändert sich das, wenn wir unser Verständnis über die KI erweitern oder exaktere Hilfsmittel zum Messen der Gehirnaktivität entwickeln.) Wenn wir also in den Medien lesen, dass diese oder jene KI-Technik »wie das menschliche Gehirn funktioniert«, besteht der Verdacht, dass es sich um eine bloße Vermutung oder reine Fiktion handelt.

Über das *Bewusstsein* wissen wir eigentlich nichts. Also sage ich dazu auch nichts. Niemand in der KI-Forschung arbeitet daran, Maschinen ein Bewusstsein zu verleihen. Es wüsste auch niemand, wo man ansetzen sollte. Dazu kommt noch, dass Bewusstsein für keine Form von Verhalten Voraussetzung ist. Angenommen, ich bitte Sie zu überprüfen, ob ein Programm eine Bedrohung für die Menschheit darstellt. Sie analysieren den Programmcode und stellen fest, dass bei der Ausführung ein Plan entsteht und umgesetzt wird, der zur Vernichtung der menschlichen Spezies führt, wie ein Schachprogramm eine Strategie entwickeln und umsetzen würde, um den menschlichen Spieler zu besiegen. Was nun, wenn ich Ihnen sage, dass dieser Pro-

grammcode, sobald er ausgeführt wird, auch eine Art Maschinenbewusstsein erschafft? Ändert das Ihre Vorhersage? Wohl kaum. Es macht *absolut keinen Unterschied*.³ Ihre Prognose zum Verhalten bleibt dieselbe, denn sie beruht ja auf dem Code. Wenn in Hollywoodfilmen Maschinen auf geheimnisvolle Weise »aufwachen« und den Menschen hassen, haben die Drehbuchautoren das Thema verfehlt: Es geht um die Fähigkeiten, nicht um das Bewusstsein.

Es gibt einen weiteren wichtigen kognitiven Aspekt des Gehirns, den wir tatsächlich zu verstehen beginnen, nämlich das *Belohnungssystem*. Dieses interne Signalsystem auf Dopaminbasis verbindet positive und negative Reize mit dem Verhalten. Seine Funktionsweise wurde von dem schwedischen Neurowissenschaftler Nils-Åke Hillarp und seinem Team Ende der 1950er-Jahre enthüllt. Das System hält uns an, positive Reize zu suchen und so den Dopaminspiegel zu erhöhen, zum Beispiel in Form von süßen Speisen. Gleichzeitig führt es dazu, dass wir negative Reize wie Hunger und Schmerz vermeiden, da diese den Dopaminspiegel senken. Es erinnert an die Glukosesuche von *E. coli*, ist aber sehr viel komplexer. So sind darin Lernverfahren eingebettet, die bewirken, dass unser Verhalten im Laufe der Zeit immer effektiver zu Belohnungen führt. Auch ein Belohnungsaufschub ist vorgesehen, bei dem die Belohnung nicht sofort, sondern verzögert erfolgt. Dabei kann es zum Beispiel darum gehen, Geld anzusparen, um sich später etwas zu leisten. Wir verstehen das Belohnungssystem unseres Gehirns unter anderem darum, weil es dem *Reinforcement Learning* (dem verstärkenden Lernen) aus der KI ähnelt, das auf einer sehr robusten Theorie fußt.⁴

Vom evolutionären Standpunkt betrachtet, ist das Belohnungssystem ebenso wie die Glukosesuche von *E. coli* ein Mittel zur verbesserten evolutionären Fitness. Organismen, die besser darin sind, Belohnungen zu suchen – leckere Speisen aufspüren, Schmerzen vermeiden, Sexualpartnern finden usw. –, haben wahrscheinlich mehr Erfolg bei der Fortpflanzung und der Weitergabe ihrer Gene. Da ein Organismus extreme Schwierigkeiten bei der Beurteilung der auf lange Sicht hierfür erfolgreichsten Strategien hätte, hat die Evolution uns mit integrierten Wegweisern auf die Sprünge geholfen.

Allerdings sind diese Wegweiser alles andere als perfekt. Tatsächlich gibt es Möglichkeiten, eine Belohnung zu erhalten, die die Wahr-

scheinlichkeit für die Verbreitung der eigenen Gene sogar *reduzieren*. Dazu gehören Drogenkonsum, das Trinken großer Mengen zucker- und kohlen säurehaltiger Getränke oder täglich 18 Stunden Videospiele spielen. Außerdem: Bestünde ein direkter elektrischer Zugang zum eigenen Belohnungssystem, liefen wir Gefahr, uns ununterbrochen selbst zu stimulieren und dabei schließlich zu sterben.⁵

Nicht nur einzelne Individuen leiden bisweilen daran, dass Belohnungssignale und evolutionäre Fitness nicht immer zusammenpassen. Auf einer kleinen Insel vor der Küste Panamas lebt das Zwergfaultier. Es scheint, als seien die Tiere von einer dem Valium ähnlichen Substanz abhängig, die in den Blättern seiner Nahrungspflanze, der Roten Mangrove, vorkommt. Das Zwergfaultier ist vom Aussterben bedroht.⁶ Das zeigt, dass eine ganze Spezies untergehen kann, wenn sie eine ökologische Nische findet, in der sie ihr Belohnungssystem mit den falschen Reizen befriedigt.

Abgesehen von derart unglücklichen Fehlschlägen erhöht ein Verhalten, das Belohnungen maximiert, in natürlichen Umgebungen im Normalfall die Chance, die eigenen Gene zu verbreiten, und die Wahrscheinlichkeit, Veränderungen der Umwelt zu überleben.

Evolutionärer Beschleuniger

Lernen dient nicht allein dem Überleben und Gedeihen, es kann auch die *Evolution beschleunigen*. Wie kann das sein? Lernen verändert die DNA nicht und bei der Evolution geht es doch um die Veränderung der DNA im Laufe der Generationen. Die Verbindung zwischen Lernen und Evolution wurde 1896 von dem amerikanischen Psychologen James Baldwin⁷ sowie dem britischen Verhaltensforscher Conwy Lloyd Morgan⁸ unabhängig voneinander vorgeschlagen, aber damals nicht allgemein akzeptiert.

Sie ist heute als Baldwin-Effekt bekannt und bezeichnet die Möglichkeit, dass die Evolution entweder einen *instinktgesteuerten* Organismus hervorbringt, bei dem jede Reaktion im Voraus feststeht, oder einen *adaptiven* Organismus, der lernt, wie er richtig handelt. Nehmen wir der Einfachheit halber an, der optimale instinktgesteuerte Organismus könne als sechsstellige Zahl codiert werden, zum Beispiel 472116. Beim adaptiven Organismus wird dagegen nur die Ziffernfolge 472*** vorgegeben und der Organismus muss die drei letzten Stellen selbst durch lebenslanges Lernen bestimmen.

Wenn die Evolution nur die ersten drei Stellen festlegen muss, ist ihre Aufgabe natürlich einfacher. Der adaptive Organismus erlernt die letzten drei Stellen im Laufe seines Lebens – ein Unterfangen, für das die Evolution viele Generationen benötigt hätte. Gehen wir davon aus, dass der adaptive Organismus überlebt, stellt das Lernvermögen eine evolutionäre Abkürzung dar. Rechenmodelle scheinen den Baldwin-Effekt zu bestätigen.⁹ Kultur beschleunigt den Prozess noch mal, da eine organisierte Zivilisation den einzelnen Organismus beim Lernen schützt und ihn mit Informationen versorgt, die das Individuum nun nicht mehr selbst herausfinden muss.

Die Geschichte über den Baldwin-Effekt ist faszinierend, aber unvollständig, denn sie geht davon aus, dass Lernen und Evolution immer dasselbe Ziel haben. Anders ausgedrückt: Sie besagt, dass jedes interne Rückkopplungssignal, über das die Stoßrichtung des Lernfortschritts im Organismus definiert wird, perfekt dem Ziel der evolutionären Fitness dient. Wir haben jedoch am Beispiel des Zwergfaultiers gesehen, dass diese Annahme nicht korrekt ist. Die integrierten Lernmechanismen liefern nur einen groben Anhaltspunkt für die langfristigen Auswirkungen einer beliebigen Handlung auf die evolutionäre Fitness. Außerdem muss man sich diese Frage stellen: »Wie ist das Belohnungssystem eigentlich entstanden?« Die Antwort: Im Rahmen eines evolutionären Prozesses, bei dem ein zumindest einigermaßen auf die evolutionäre Fitness abgestimmter Rückkopplungsmechanismus internalisiert wurde.¹⁰ Ein Lernmechanismus, der den Organismus vor möglichen Partnern weglaufen ließe und stattdessen hin zu Raubtieren, würde sich nicht lange halten.

Wir haben es also dem Baldwin-Effekt zu verdanken, dass Neuronen mit ihrer Fähigkeit, zu lernen und Probleme zu lösen, im Tierreich so weit verbreitet sind. Dabei ist es wichtig, zu begreifen, dass es der Evolution egal ist, ob Sie ein Gehirn haben oder interessanten Gedanken nachhängen. Für die Evolution sind Sie nur ein *Agent*, also jemand, der handelt. Würdige geistige Merkmale wie logisches Schlussfolgern, zielgerichtetes Planen, Weisheit, Witz, Vorstellungskraft und Kreativität können für die Intelligenz eines Agenten ausschlaggebend sein – oder auch nicht. Ein Grund für die Faszination der künstlichen Intelligenz besteht darin, dass sie uns einen möglichen Weg zeigt, diese Phänomene besser zu verstehen: Wir könnten begreifen, wie diese geistigen Merkmale intelligentes Verhalten ermöglichen und warum

es nicht möglich ist, ohne sie ein wirklich intelligentes Verhalten zu schaffen.

Individuelle Rationalität

Seit Anbeginn der antiken griechischen Philosophie ist das Konzept der Intelligenz mit den Fähigkeiten des *erfolgreichen* Wahrnehmens, Nachdenkens und Handelns verknüpft.¹¹ Im Laufe der Jahrhunderte wurde dieses Konzept sowohl in einem erweiterten Sinn angewandt als auch präziser definiert.

Neben anderen hat Aristoteles untersucht, was erfolgreiches Nachdenken ausmacht: Deduktionsverfahren, die auf Basis wahrer Prämissen zu wahren Konklusionen führen. Er hat ebenfalls untersucht, wie sich Menschen entscheiden, was sie tun sollen, manchmal auch als *praktische Vernunft* bezeichnet. Und er konstatierte, dass dabei der Schluss, dass bestimmte Handlungen zu einem gewünschten Ziel führen, eine Rolle spielt:

»Unsere Überlegung betrifft nicht das Ziel, sondern die Mittel, es zu erreichen. Der Arzt überlegt nicht, ob er heilen, der Redner nicht, ob er überzeugen [...] soll, sondern nachdem man sich ein Ziel gestellt hat, sieht man sich um, wie und durch welche Mittel es zu erreichen ist; wenn es durch verschiedene Mittel möglich scheint, sieht man zu, durch welches es am leichtesten und besten erreicht wird; und wenn es durch eines regelrecht verwirklicht wird, fragt man wieder, *wie* es durch *dasselbe* verwirklicht wird, und wodurch wiederum jenes, bis man zu der ersten Ursache gelangt, [...] und das, was bei der Zergliederung als Letztes herauskommt, ist bei der Verwirklichung durch die Handlung das erste. Stößt man auf eine Unmöglichkeit, so steht man von der Sache ab, z.B. wenn sich Geldmittel erforderlich zeigen, die man nicht aufbringen kann. Erscheint die Sache aber möglich, so nimmt man sie in die Hand.«¹²

Diese Ausführungen scheinen in der westlichen Welt den Grundtenor des Verständnisses von Rationalität für die nächsten gut 2.000 Jahre bestimmt zu haben. Ihnen zufolge ist das »Ziel«, das der Mensch erreichen will, unveränderlich und vorbestimmt und die vernünftige