Szenarien
einer
kommenden
Revolution

Nick
Bostrom

Superintelligenz

Suhrkamp

Was geschieht, wenn es Wissenschaftlern eines Tages gelingt, eine Maschine zu entwickeln, die die menschliche Intelligenz auf so gut wie allen wichtigen Gebieten übertrifft? Klar ist: Eine solche Superintelligenz wäre enorm mächtig und würde uns vor Kontroll- und Steuerungsprobleme stellen, verglichen mit denen die Bewältigung des Klimawandels ein Klacks ist. Mehr noch: Vermutlich würde die Zukunft der menschlichen Spezies in den Händen dieser Superintelligenz liegen, so wie heute die Zukunft der Gorillas von uns abhängt. Zukunftsmusik? Oder doch Science-Fiction? Eindeutig Zukunftsmusik, sagt Nick Bostrom, und zwar eine, die vielleicht schon binnen eines Menschenalters erklingen wird. Damit wir verstehen, was auf uns zukommt, nimmt er uns mit auf eine faszinierende Reise in die Welt der Orakel und Genies, der Superrechner und Gehirnsimulationen, aber vor allem in die Labore dieser Welt, in denen fieberhaft an der Entwicklung einer künstlichen Intelligenz gearbeitet wird.

Bostrom skizziert mögliche Szenarien, wie die Geburt der Superintelligenz vonstattengehen könnte, und widmet sich ausführlich den Folgen dieser Revolution. Sie werden global sein und unser wirtschaftliches, soziales und politisches Leben tiefgreifend verändern. Wir müssen handeln, und zwar kollektiv, bevor der Geist aus der Flasche gelassen ist – also jetzt! Das ist die eminent politische Botschaft dieses so spannenden wie wichtigen Buches.

Nick Bostrom, geboren 1973 in Schweden, studierte Physik, Mathematik, Neurowissenschaften sowie Philosophie unter anderem am King's College in London und an der London School of Economics, wo er promoviert wurde. Derzeit ist er Professor für Philosophie am St. Cross College der Universität von Oxford und Direktor sowohl des Future of Humanity Institute als auch des Programme for the Impact of Future Technology. 2009 wurde er für seine Arbeit mit dem prestigeträchtigen Eugene R. Gannon Award for the Continued Pursuit of Human Advancement ausgezeichnet und war auf der *100 Top Global Thinkers List* von *Foreign Policy.*

Nick Bostrom

Superintelligenz

Szenarien einer kommenden Revolution

Aus dem Englischen von
Jan-Erik Strasser

Suhrkamp

eBook Suhrkamp Verlag Berlin 2014
Der vorliegende Text folgt der Erstausgabe, 2014.

Titel der Originalausgabe: *Superintelligence. Paths, Dangers, Strategies*
First Edition was originally published in English in 2014.
This translation is published by arrangement with Oxford University Press.
Erstmals erschienen 2014 bei Oxford University Press. Die Übersetzung erscheint mit freundlicher Genehmigung von Oxford University Press.

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.

Satz: Satz-Offizin Hümmer GmbH, Waldbüttelbrunn
Umschlaggestaltung: Hermann Michels und Regina Göllner

eISBN 978-3-518-73900-6
www.suhrkamp.de

Inhalt

Die unvollendete Fabel von den Spatzen

Es war die Zeit des Nestbaus, aber nach tagelanger harter Arbeit saßen die Spatzen in der Abenddämmerung beisammen, um auszuruhen und miteinander zu zwitschern.

»Wir sind alle so klein und schwach. Stellt euch vor, wie angenehm es wäre, wenn wir eine Eule hätten, die uns beim Nestbau helfen würde!«

»Genau!«, stimmte ein anderer Spatz ein. »Sie könnte bei Jung und Alt nach dem Rechten sehen.«

»Und uns Ratschläge geben und vor der Nachbarskatze warnen«, meinte ein dritter.

Darauf sprach Pastus, der Schwarmälteste: »Schicken wir unsere Späher in alle Himmelsrichtungen aus, um ein verwaistes Eulenküken oder ein Ei zu finden. Auch ein Krähenjunges oder ein kleines Wiesel könnten von Nutzen sein. Das ist vielleicht das Beste, was uns je passiert ist – zumindest seit der Eröffnung des Pavillons der nie enden wollenden Körner.«

Die ganze Schar war begeistert, und die Spatzen allüberall begannen aus vollster Kehle zu trällern.

Nur Scronkfinkle, ein einäugiger Spatz von mürrischem Gemüt, war von der Klugheit des Vorhabens nicht überzeugt. Er sprach: »Das wird unser Verderben sein. Sollten wir nicht erst bedenken, wie eine Eule sich zähmen und bändigen lässt, bevor wir sie in unsere Mitte bringen?«

Pastus erwiderte: »Eine Eule zu zähmen scheint mir ein höchst schwieriges Unterfangen zu sein. Es wird schon schwer genug werden, ein Ei zu finden. Fangen wir also damit an. Nachdem wir die Eule großgezogen haben, können wir uns der nächsten Herausforderung stellen.«

»Der Plan hat einen Haken!«, piepste Scronkfinkle, aber sein Protest ging im Tumult des auffliegenden Schwarms unter, der sich daranmachte,

Pastus' Vorhaben in die Tat umzusetzen.

Nur zwei oder drei Spatzen blieben mit Scronkfinkle zurück. Gemeinsam versuchten sie herauszufinden, wie Eulen gezähmt oder gebändigt werden könnten. Schon bald wurde ihnen klar, dass Pastus recht gehabt hatte: Es war in der Tat ein höchst schwieriges Unterfangen, zumal ihnen eine Eule zum Üben fehlte. Nichtsdestotrotz fuhren sie fort, so gut sie konnten, und in der ständigen Furcht, die anderen Spatzen könnten mit einem Ei zurückkehren, bevor eine Lösung für das Kontrollproblem gefunden wäre.

Niemand weiß, wie die Geschichte ausgeht, aber der Autor widmet dieses Buch Scronkfinkle und seinen Anhängern.

Vorwort

In Ihrem Schädel steckt etwas, das diesen Satz liest. Dieses Ding, das menschliche Gehirn, hat einige Fähigkeiten, die den Gehirnen anderer Tiere fehlen, und diesen besonderen Fähigkeiten verdanken wir unsere dominante Stellung auf der Erde. Andere Tiere haben stärkere Muskeln oder schärfere Krallen, doch wir haben die schlaueren Gehirne. Dieser bescheidene Vorteil an allgemeiner Intelligenz hat uns schließlich Sprache, Technologie und komplexe soziale Organisationen entwickeln lassen, und er wuchs im Lauf der Zeit, da jede Generation auf den Leistungen der vorigen aufbauen konnte.

Falls wir eines Tages künstliche Gehirne bauen, die das menschliche an allgemeiner Intelligenz übertreffen, dann könnte diese neue Art von Superintelligenz überaus mächtig werden. Und genau wie das Schicksal der Gorillas heute stärker von uns Menschen abhängt als von den Gorillas selbst, so hinge das Schicksal unserer Spezies von den Handlungen dieser maschinellen Superintelligenz ab.

Wir haben allerdings einen Vorteil: Wir sind diejenigen, die das Ding bauen. Im Prinzip könnten wir eine Art von Superintelligenz erschaffen, die menschliche Werte achtet, und wir hätten sicherlich gute Gründe, genau das zu tun. In der Praxis jedoch sieht das Kontrollproblem (wie können wir kontrollieren, was die Superintelligenz tun würde?) ziemlich schwierig aus. Außerdem scheint es, als hätten wir nur einen Schuss frei. Sobald eine unfreundliche Superintelligenz existiert, wird sie uns davon abhalten, sie zu ersetzen oder ihre Präferenzen zu ändern. Dann wäre unser Schicksal besiegelt.

In diesem Buch versuche ich zu verstehen, welche Herausforderung die Superintelligenz darstellt und wie wir darauf reagieren sollten. Dies ist die wahrscheinlich größte und beängstigendste Aufgabe, der die Menschheit

je gegenüberstand – und egal, ob wir sie meistern oder an ihr scheitern: Es wird wohl auch die letzte sein.

Sie werden hier keine Argumente dafür finden, dass wir kurz vor einem großen Durchbruch in der Forschung zur künstlichen Intelligenz stehen oder dass sich auch nur einigermaßen genau vorhersagen lässt, wann es so weit ist. Es sieht so aus, als ob es irgendwann in diesem Jahrhundert geschehen wird, aber sicher können wir uns dessen nicht sein.

Die ersten Kapitel zeigen mögliche Wege zur Superintelligenz auf und sagen etwas zum zeitlichen Ablauf, doch der größte Teil des Buches ist der Frage gewidmet, was danach geschieht. Wir untersuchen die Kinetik einer Intelligenzexplosion, die Formen und die Fähigkeiten einer Superintelligenz sowie die Strategien, die einem superintelligenten Akteur zur Verfügung stehen, sobald er einen entscheidenden Vorteil erlangt hat. Im Anschluss verlagern wir unseren Fokus auf das Kontrollproblem und fragen danach, was zu tun ist, damit wir das Resultat all dieser Entwicklungen überleben und es für uns günstig ausfällt. Gegen Ende des Buches treten wir schließlich einen Schritt zurück, betrachten das große Ganze, das sich aus unseren Untersuchungen ergeben hat, und machen einige Vorschläge dazu, was schon jetzt getan werden kann, um eine existentielle Katastrophe zu vermeiden.

Es war nicht leicht, dieses Buch zu schreiben. Ich hoffe, dass der nun freigeräumte Weg es anderen Forschern ermöglicht, das neue Territorium schneller und bequemer zu erreichen, sodass sie sich dort – frisch und ausgeruht – mit uns daranmachen können, die Grenzen unseres Verständnisses zu erweitern. (Und falls dieser Weg ein wenig holprig und kurvenreich ist, so hoffe ich, dass die Kritiker die ursprüngliche Unwirtlichkeit des Geländes im Nachhinein nicht unterschätzen!)

Es war nicht leicht, dieses Buch zu schreiben: Ich habe versucht, es lesbar zu machen, aber es ist mir wohl nicht ganz gelungen. Beim Schreiben hatte ich als Zielgruppe ein früheres Ich im Sinn, das am Lesen dieses Buches Gefallen gefunden hätte – und das könnte den Leserkreis ganz schön einschränken. Dennoch denke ich, dass der Inhalt vielen Menschen zugänglich ist, wenn sie ihn gründlich studieren und der

Versuchung widerstehen, jeden neuen Gedanken augenblicklich mit dem nächstliegenden Klischee zu verwechseln. Leser ohne entsprechendes Vorwissen sollten sich nicht von den paar Happen Mathematik oder Fachvokabular abschrecken lassen, denn es ist immer möglich, den Hauptgedanken aus dem Kontext zu erschließen (und umgekehrt werden diejenigen Leser, die ans Eingemachte wollen, in den Anmerkungen fündig werden)[1].

Viele der Argumente in diesem Buch sind vermutlich falsch,[2] und wahrscheinlich habe ich auch Überlegungen von entscheidender Bedeutung nicht berücksichtigt, womit einige oder alle meine Schlussfolgerungen ungültig wären. Es hat mich einige Mühe gekostet, im gesamten Text auf Nuancen und Grade der Ungewissheit hinzuweisen, indem ich ihn mit einer Unzahl von Wörtern wie »könnte«, »dürfte«, »müsste«, »vielleicht« oder »wahrscheinlich« gespickt habe. Jede Einschränkung dieser Art wurde sorgfältig und bewusst platziert, doch selbst diese Hinweise epistemischer Bescheidenheit reichen nicht aus – sie müssen um ein globales Eingeständnis der Ungewissheit und Fehlbarkeit ergänzt werden. Das ist keine falsche Bescheidenheit: Obwohl mein Buch wahrscheinlich schwere Fehler und Irreführungen enthält, stehen die in der Literatur vorgebrachten Alternativen meiner Meinung nach noch wesentlich schlechter da – einschließlich der üblichen Ansicht (der »Nullhypothese«), der zufolge wir die Aussicht auf eine Superintelligenz vorerst gefahrlos oder vernünftigerweise ignorieren können.

1. Vergangene Entwicklungen und gegenwärtige Möglichkeiten

Ganz zu Beginn werfen wir einen Blick zurück. Wenn man sich die historische Entwicklung im größtmöglichen Maßstab ansieht, so scheint sie sich als Abfolge unterschiedlicher und immer schnellerer Wachstumsmodi darzustellen. Aufgrund dieses Musters wurde angenommen, dass ein weiterer (noch schnellerer) Wachstumsmodus möglich sein könnte. Wir messen dieser Beobachtung allerdings nicht zu viel Gewicht bei: Dies ist kein Buch über »technologische Beschleunigung«, »exponentielles Wachstum« oder die diversen Vorstellungen, die manchmal unter der Rubrik der »Singularität« versammelt werden. Im Anschluss an diese Überlegungen betrachten wir die Geschichte der KI-Forschung, bevor wir uns einen Überblick über den aktuellen Stand dieser Disziplin verschaffen. Zuletzt sehen wir uns einige kürzlich durchgeführte Expertenbefragungen an und setzen uns mit unserer Unwissenheit bezüglich der Zeitpunkte zukünftiger Fortschritte auseinander.

Wachstumsmodi und Weltgeschichte

Es ist erst einige Millionen Jahre her, dass unsere Vorfahren sich durch die Baumkronen der afrikanischen Urwälder schwangen. Nach geologischen und sogar evolutionären Maßstäben ist der Homo sapiens schnell aus seinem letzten gemeinsamen Vorfahren mit den Menschenaffen hervorgegangen. Wir entwickelten die aufrechte Haltung, den opponierbaren Daumen und – am wichtigsten – einige relativ geringfügige Änderungen der Gehirngröße und neurologischen Organisation, die zu einem sprunghaften Anstieg der kognitiven

Fähigkeiten führten. Als Folge davon kann der Mensch abstrakt denken, komplexe Gedanken ausdrücken und Informationen über die Generationen hinweg kulturell weit besser anhäufen als jede andere Spezies auf diesem Planeten.

Diese Fähigkeiten ließen uns immer effizientere Produktionsmethoden entwickeln, sodass es unseren Vorfahren schließlich gelang, den Regenwald und die Savanne weit hinter sich zu lassen. Insbesondere nach der Einführung der Landwirtschaft wuchs die Bevölkerungsdichte mit der Gesamtbevölkerung an. Mehr Menschen bedeuteten auch mehr Ideen; größere Dichten bedeuteten, dass Ideen sich leichter ausbreiten und manche Menschen sich spezialisieren konnten. Diese Entwicklungen erhöhten die *Wachstumsraten* der wirtschaftlichen Produktivität und der technologischen Leistungsfähigkeit. Spätere, mit der industriellen Revolution zusammenhängende Entwicklungen führten zu einem zweiten, vergleichbar dramatischen Anstieg der Wachstumsrate.

Solche Anstiege haben gewichtige Konsequenzen. Vor ein paar hunderttausend Jahren, in der frühen menschlichen (oder hominiden) Vorgeschichte, war das Wachstum so langsam, dass es etwa eine Million Jahre gedauert hätte, um die Produktionskapazitäten so weit zu erhöhen, dass das Existenzminimum einer weiteren Million Menschen gesichert gewesen wäre. Bis zum Jahr 5000 v. Chr. stieg die Wachstumsrate in der Folge der Erfindung der Landwirtschaft dann so weit an, dass das gleiche Wachstum in nur zwei Jahrhunderten erreicht wurde. Heute, nach der industriellen Revolution, dauert es noch 90 Minuten.[1]

Auch die gegenwärtige Wachstumsrate wird also schon eindrucksvolle Ergebnisse liefern, wenn sie noch eine gewisse Zeit stabil bleibt. Wächst die Wirtschaft weiterhin so wie in den letzten 50 Jahren, dann wird die Welt im Jahr 2050 rund 4,8-mal reicher und im Jahr 2100 etwa 34-mal reicher sein als heute.[2]

Doch die Aussicht auf ein stabiles exponentielles Wachstum verblasst im Vergleich zu einem erneuten sprunghaften *Anstieg der Wachstumsrate*, der mit denjenigen der landwirtschaftlichen oder der industriellen Revolution vergleichbar wäre. Auf der Grundlage historischer Wirtschafts-

und Bevölkerungsdaten hat der Ökonom Robin Hanson die Verdopplungszeit des Weltwirtschaftswachstums geschätzt. Für Gemeinschaften von Jägern und Sammlern im Pleistozän kommt er auf 224 000 Jahre, für Agrargesellschaften auf 909 Jahre und für Industriegesellschaften auf 6,3 Jahre.[4] (Hansons Modell zufolge herrscht in der gegenwärtigen Epoche eine Mischung aus landwirtschaftlichen und industriellen Wachstumsmodi vor – die Weltwirtschaft als ganze verdoppelt sich noch nicht alle 6,3 Jahre.) Bei einem weiteren Übergang zu einem anderen Wachstumsmodus ähnlichen Ausmaßes wäre eine sich etwa alle zwei Wochen verdoppelnde Weltwirtschaft die Folge.

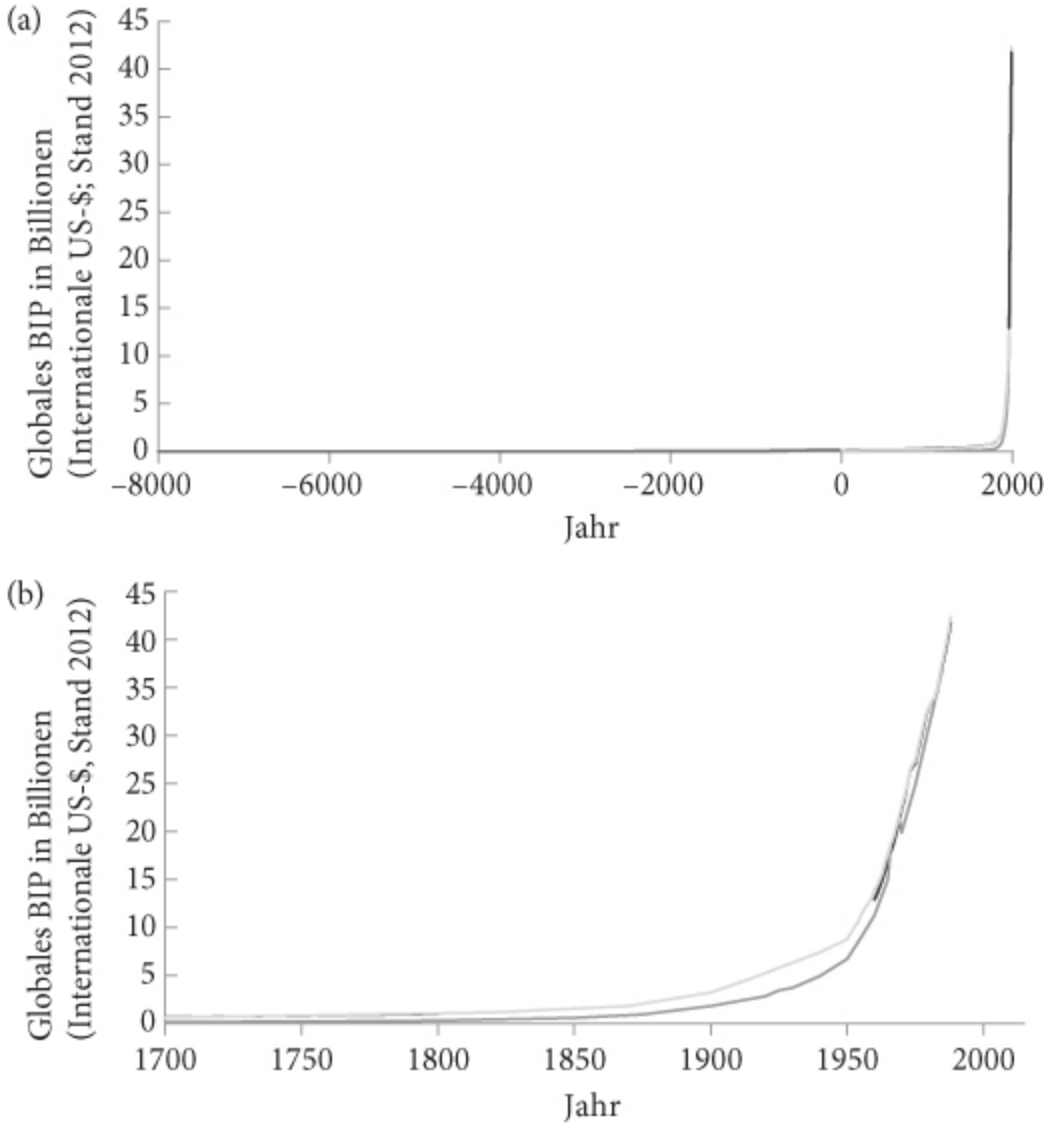

Abb. 1: Langzeitbetrachtung des globalen BIP. Auf einer linearen Skala aufgetragen, ähnelt die Geschichte der Weltwirtschaft einer flachen Linie, die sich an die x-Achse schmiegt, bis sie plötzlich senkrecht nach oben schießt. (a) Selbst wenn wir uns nur auf die letzten 10 000 Jahre beschränken, bleibt das Muster im Wesentlichen ein 90-Grad-Winkel. (b) Erst in den letzten rund 100 Jahren steigt die Kurve sichtbar an. (Die verschiedenen Linien des Graphen entsprechen verschiedenen Datensätzen, die leicht unterschiedliche Schätzwerte liefern.[3])

So etwas erscheint derzeit unvorstellbar, aber in früheren Zeiten mag die Annahme ebenso absurd gewesen sein, die Weltwirtschaft könne sich irgendwann binnen eines einzigen Menschenlebens mehrmals verdoppeln – und doch halten wir genau diese außergewöhnliche Situation heute für völlig normal.

Der Gedanke einer kommenden technologischen Singularität ist mittlerweile weit verbreitet, wofür sowohl Vernor Vinges bahnbrechender Aufsatz als auch die Schriften von Ray Kurzweil und anderen gesorgt haben.[5] Der Begriff »Singularität« wurde jedoch in vielen unterschiedlichen Bedeutungen ge- und missbraucht und ist mittlerweile von einer unheiligen (wenngleich fast millenaristischen) Aura techno-utopischer Konnotationen umgeben.[6] Da das meiste davon für unsere Argumentation ohne Belang ist, können wir Klarheit schaffen, indem wir auf das Wort »Singularität« verzichten und stattdessen präzisere Begriffe nutzen.

Die mit der Singularität zusammenhängende Idee, die uns hier interessiert, ist die Möglichkeit einer *Intelligenzexplosion*, insbesondere die Aussicht auf eine maschinelle Superintelligenz. Es mag Leute geben, die durch Wachstumsdiagramme (wie die in *Abbildung 1* präsentierten) davon überzeugt wurden, dass ein weiterer Wachstumsmodus vor der Tür steht, der mit den landwirtschaftlichen oder industriellen Revolutionen vergleichbar ist. Diese Leute könnten daraus schließen, dass ein Szenario, in dem die Verdopplungszeit der Weltwirtschaft nur wenige Wochen beträgt, die Erschaffung von Intelligenzen voraussetzt, die viel schneller und effizienter sind als die uns vertrauten biologischen. Um die Möglichkeit einer solchen Revolution ernst zu nehmen, müssen wir jedoch

keine Punkte auf Kurven auftragen oder Wirtschaftswachstumsdaten extrapolieren. Wie wir sehen werden, gibt es dafür stärkere Gründe.

Große Erwartungen

Mit Maschinen, die dem Menschen an allgemeiner Intelligenz gleichkommen – die also über gesunden Menschenverstand ebenso verfügen wie über die Fähigkeit zu lernen, zu schlussfolgern und zu planen, um komplexe Herausforderungen in einer Vielzahl von natürlichen und abstrakten Bereichen zu meistern –, wurde seit der Erfindung des Computers in den 1940er Jahren gerechnet. Zu jener Zeit wurde das Erscheinen solcher Maschinen oft etwa 20 Jahre in die Zukunft verlegt.[7] Seitdem hat sich das erwartete Ankunftsdatum mit einer Geschwindigkeit von einem Jahr pro Jahr nach hinten verschoben, sodass manche Futuristen, die sich mit der Möglichkeit der künstlichen allgemeinen Intelligenz befassen, auch heute noch glauben, dass es binnen zwei Jahrzehnten intelligente Maschinen geben wird.[8]

Zwei Dekaden sind eine ideale Zeitspanne für Propheten eines radikalen Wandels: kurz genug, um Aufmerksamkeit zu erregen und relevant zu sein, aber lang genug, um zu plausibilisieren, dass bis dahin eine Reihe von Durchbrüchen stattgefunden haben könnte, die derzeit nur zu erahnen sind. Zum Vergleich: Die meisten Technologien, die in fünf oder zehn Jahren einen großen Einfluss auf die Welt haben werden, sind bereits in begrenztem Einsatz, und diejenigen, die die Welt binnen 15 Jahren revolutionieren, existieren wahrscheinlich schon als Prototypen. Zwanzig Jahre dürften auch in der Nähe des Karriereendes eines typischen Prognostikers liegen, was das Risiko mindert, mit einer kühnen Vorhersage den eigenen Ruf zu ruinieren.

Aus der Tatsache, dass viele Prognosen in der Vergangenheit zu vorschnell waren, folgt allerdings nicht, dass eine künstliche Intelligenz (KI) unmöglich ist oder niemals entwickelt werden wird.[9] Der Hauptgrund für die Verzögerung ist, dass die technischen Probleme bei der Konstruktion intelligenter Maschinen größer sind, als die frühen Pioniere

dachten. Dies aber lässt das konkrete Ausmaß der Probleme ebenso offen wie die Frage, wie weit wir jetzt von deren Überwindung entfernt sind. Manchmal hat ein Problem, das zunächst hoffnungslos kompliziert erscheint, eine überraschend einfache Lösung (zugegebenermaßen kommt das Umgekehrte vermutlich häufiger vor).

Im nächsten Kapitel werden wir verschiedene Wege betrachten, die zu einer maschinellen Intelligenz auf menschlichem Niveau führen könnten. Aber schon an dieser Stelle sei daran erinnert, dass diese nicht das Ziel der Reise darstellt, egal, wie viele Zwischenstopps wir auf dem Weg dorthin einlegen müssen. Bereits der nächste Halt danach, nur eine kurze Strecke entfernt, heißt *übermenschliche* maschinelle Intelligenz. Der Zug wird in Menschendorf nicht anhalten oder auch nur abbremsen, sondern wahrscheinlich einfach durchrasen.

Der Mathematiker I. J. Good, der im Zweiten Weltkrieg Chefstatistiker der Gruppe um Alan Turing war, die die deutschen Geheimcodes entschlüsselte, dürfte die zentralen Aspekte dieses Szenarios als Erster formuliert haben. In einer häufig zitierten Passage aus dem Jahr 1965 schrieb er:

> Eine ultraintelligente Maschine sei definiert als eine Maschine, die alle geistigen Anstrengungen jedes noch so schlauen Menschen bei weitem übertreffen kann. Da die Konstruktion von Maschinen solch eine geistige Anstrengung ist, könnte eine ultraintelligente Maschine noch bessere Maschinen konstruieren; zweifellos würde es dann zu einer »Intelligenzexplosion« kommen, und die menschliche Intelligenz würde weit dahinter zurückbleiben. Die erste ultraintelligente Maschine ist also die letzte Erfindung, die der Mensch je machen muss, vorausgesetzt, die Maschine ist fügsam genug, um uns zu sagen, wie man sie unter Kontrolle hält.[10]

Heute scheint es offensichtlich zu sein, dass große existentielle Risiken mit einer solchen Intelligenzexplosion verbunden wären und das Thema daher mit größtem Ernst zu behandeln ist, selbst wenn klar wäre (was nicht der Fall ist), dass nur eine relativ kleine Wahrscheinlichkeit dafür besteht. Die meisten Pioniere der künstlichen Intelligenz aber zogen die Möglichkeit einer übermenschlichen KI nicht in Betracht, obwohl sie von einer unmittelbar bevorstehenden KI menschlichen Niveaus überzeugt waren. Es ist, als ob ihr Mut zum Mutmaßen vom Ersinnen der radikalen

Möglichkeit intelligenter Maschinen so erschöpft worden wäre, dass sie sich das Korollar – superintelligente Maschinen – nicht mehr ausmalen konnten.

Auch die Möglichkeit von Risiken ließen sie nicht zu,[11] ja, sie taten nicht einmal so, als ob irgendwelche Sicherheitsbedenken oder moralische Skrupel bei der Schaffung von künstlichen Intelligenzen und potentiellen Computerdiktatoren eine Rolle spielen könnten: ein Versäumnis, das sogar vor dem Hintergrund der wenig beeindruckenden Standards der Technikfolgenabschätzung jener Zeit erstaunt.[12] Wenn es so weit ist, müssen wir darauf hoffen, dass wir nicht nur die technologische Kompetenz haben werden, eine Intelligenzexplosion einzuleiten, sondern auch die darüber hinaus erforderlichen Fähigkeiten besitzen, um die Detonation zu überleben.

Ehe wir uns jedoch dem zuwenden, was vor uns liegt, wird es sinnvoll sein, einen kurzen Blick auf die Geschichte der maschinellen Intelligenz von ihren Anfängen bis heute zu werfen.

Zeiten der Hoffnung und der Hoffnungslosigkeit

Im Sommer 1956 kamen am Dartmouth College zehn Wissenschaftler zu einem sechswöchigen Workshop zusammen, die sich alle für neuronale Netze, Automatentheorie und das Studium der Intelligenz interessierten. Dieses Treffen wird häufig als die Geburtsstunde des Forschungsfelds der künstlichen Intelligenz betrachtet, und viele der Teilnehmer gelten heute als dessen Gründerväter. Ihr Optimismus kommt in einem Antrag an die Rockefeller-Stiftung zum Ausdruck, die Mittel für den Workshop bereitstellte:

> Wir schlagen vor, dass eine zweimonatige, zehnköpfige Untersuchung der künstlichen Intelligenz durchgeführt wird. […] Die Studie soll auf der Grundlage der Vermutung durchgeführt werden, dass jedes Merkmal des Lernens oder der Intelligenz überhaupt im Prinzip so genau beschreibbar ist, dass eine Maschine es simulieren kann. Es wird der Versuch gemacht werden, herauszufinden, wie man Maschinen baut, die Sprache verwenden, abstrahieren und Begriffe bilden, die Arten von Problemen lösen, welche derzeit den Menschen vorbehalten sind, und die sich selbst vervollkommnen. 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 daran arbeitet.

Seit diesen kühnen Worten sind sechs Jahrzehnte vergangen, in denen sich auf dem Feld der künstlichen Intelligenz Begeisterungsstürme und hohe Erwartungen mit Rückschlägen und Enttäuschungen abwechselten.

Die Zeit der anfänglichen Begeisterung, die mit dem Treffen in Dartmouth begann, wurde später von John McCarthy (dem Hauptorganisator der Veranstaltung) als die »Kein Trick!«-Ära bezeichnet. In jenen Tagen bauten die Forscher einfache Systeme, um die damals oft zu hörenden Behauptungen der Form »Keine Maschine wird jemals X können!« zu widerlegen. Diese Systeme konnten eine abgespeckte Version von X in einer »Mikrowelt« (einem wohldefinierten, begrenzten Bereich) ausführen, sodass die prinzipielle Machbarkeit einer Maschine, die X in der echten Welt tun könnte, nachgewiesen war. Ein solches frühes System, *Logic Theorist*, war in der Lage, die meisten Theoreme des zweiten Kapitels von Whiteheads und Russells *Principia Mathematica* zu beweisen, und kam sogar auf einen Beweis, der viel eleganter war als der ursprüngliche. Eine Weiterentwicklung des Programms, der *General Problem Solver*, war im Prinzip fähig, eine Vielzahl formal spezifizierter Probleme zu lösen.[13] Dadurch war die Vorstellung, Maschinen könnten »nur numerisch denken«, widerlegt und gezeigt, dass Maschinen zur Deduktion und zum Erfinden logischer Beweise fähig sind.[14] Außerdem schrieb man Programme zum Lösen von Analysis-Aufgaben für Erstsemester, von visuellen Analogieproblemen von der Art, wie sie in einigen IQ-Tests auftauchen, sowie von einfachen verbalen Algebra-Aufgaben.[15] Der Roboter *Shakey* (so getauft wegen seiner Neigung, während des Betriebs zu zittern) demonstrierte, wie logisches Denken und Wahrnehmung integriert und zur Planung und Steuerung körperlicher Aktivität genutzt werden können,[16] und das Programm *ELIZA* imitierte einen Psychotherapeuten der Rogers-Schule.[17] Mitte der 1970er Jahre zeigte das Programm *SHRDLU*, wie ein simulierter Roboterarm in einer simulierten Welt, bestehend aus geometrischen Klötzchen, Anweisungen befolgen konnte; außerdem beantwortete es auf Englisch eingegebene Fragen

dazu.[18] In den folgenden Jahrzehnten wurden Systeme entwickelt, die zeigten, dass Maschinen Werke im Stil verschiedener klassischer Komponisten schaffen, angehende Ärzte in bestimmten klinischen Diagnoseaufgaben übertreffen, autonom Autos steuern und patentierbare Erfindungen machen können.[19] Es gab sogar eine KI, die sich neue Witze ausdachte.[20] (Ihr Humor war zwar nicht besonders hochklassig, aber Kinder fanden die Pointen offenbar durchaus unterhaltsam.)

Die Methoden, die man bei den ersten Programmen erfolgreich angewandt hatte, versagten jedoch bei allgemeineren oder schwierigeren Problemen. Ein Grund dafür ist, dass diese Methoden auf der Suche nach der richtigen Antwort der Reihe nach alle Möglichkeiten prüfen, auch wenn deren Anzahl explosionsartig anwächst. Solche Methoden eignen sich gut für einfache Problemfälle, sie versagen jedoch, wenn die Dinge komplizierter werden. Möchte man ein Theorem beweisen, dessen Beweis – in einem deduktiven System mit einer Schlussregel und fünf Axiomen – fünf Zeilen lang ist, dann kann man einfach jede der 3125 möglichen Kombinationen daraufhin prüfen, ob sie zum gewünschten Ergebnis führt. Eine solche Exhaustionsmethode funktioniert auch noch für sechs- und siebenzeilige Beweise; wird die Aufgabe jedoch schwieriger, stößt sie bald an ihre Grenzen. Die Suche nach einem 50-zeiligen Beweis dauert nicht zehn Mal so lang wie diejenige nach dem fünfzeiligen: Will man alle Möglichkeiten prüfen, dann sind das 5^{50} ($\sim 8{,}9 \times 10^{34}$) mögliche Sequenzen – und das ist auch mit den schnellsten Supercomputern nicht zu schaffen.

Um das Problem der kombinatorischen Explosion zu überwinden, braucht man Algorithmen, die Regularitäten im Zielbereich ebenso ausnutzen wie Vorwissen, indem sie heuristische Suchen, Planung und flexible, abstrakte Repräsentationen verwenden – alles Fähigkeiten, die die frühen KI-Systeme so gut wie nicht besaßen. Ihre Leistung litt auch darunter, dass sie mit Unsicherheit, der Abhängigkeit von fehleranfälligen und schlecht fundierten symbolischen Repräsentationen, mit Datenknappheit und starken Einschränkungen bezüglich der Speicherkapazität und Prozessorgeschwindigkeit nur schlecht zurande

kamen. Mitte der 1970er Jahre wurden diese Probleme immer stärker als solche wahrgenommen. Die Erkenntnis, dass viele KI-Projekte letztendlich zum Scheitern verurteilt waren, führte zur ersten »KI-Eiszeit«: eine Zeit der Sparmaßnahmen, in der die Förderung ab- und die Skepsis zunahm und in der die KI aus der Mode kam.

Ein neuer Frühling begann in den frühen 1980er Jahren, als Japan eine finanzkräftige Public-Private-Partnership, das sogenannte Computerprojekt der fünften Generation, startete. Es ging um die Entwicklung einer massiv-parallelen Computerarchitektur, die den Stand der Technik revolutionieren und als Plattform für eine künstliche Intelligenz dienen sollte. Genau zu dieser Zeit erreichte auch die Faszination für das japanische »Nachkriegswirtschaftswunder« ihren Höhepunkt. Westliche Regierungen und Wirtschaftsführer versuchten verzweifelt, hinter das Geheimnis des japanischen Erfolgs zu kommen, um es zu kopieren. Als Japan beschloss, groß in die KI-Forschung zu investieren, zogen einige andere Länder prompt nach.

In den folgenden Jahren verbreiteten sich *Expertensysteme*, das heißt regelbasierte Programme, die zur Unterstützung von Entscheidungsträgern gedacht waren, enorm. Sie konnten einfache Schlussfolgerungen aus einer Datenbasis ziehen, die mühsam codiertes und von Hand eingegebenes menschliches Expertenwissen enthielt. Hunderte dieser Expertensysteme wurden gebaut. Die kleineren Systeme hatten allerdings kaum einen Nutzen, und die größeren stellten sich als teuer in Entwicklung, Bewertung, Aktualisierung und in der Regel auch zu umständlich in der Bedienung heraus; ein Computer mit einem einzigen Programm war zu unpraktisch. In den späten 1980er Jahren ging auch diese Wachstumszeit vorüber.

Das japanische Großprojekt verfehlte seine Ziele, und ebenso erging es den entsprechenden Vorhaben in den Vereinigten Staaten und Europa. Eine zweite Eiszeit brach an. Kritiker konnten zu Recht beklagen, dass »die Geschichte der KI-Forschung bis heute stets aus sehr begrenzten Erfolgen in einzelnen Bereichen besteht, die sofort von dem Misserfolg abgelöst werden, die gerade erst geweckten größeren Erwartungen zu

erfüllen«.[21] Vorhaben, die irgendetwas mit »künstlicher Intelligenz« zu tun hatten, wurden von privaten Geldgebern mehr und mehr gemieden, und selbst unter Akademikern und deren Geldgebern hatte das Label »KI« fortan einen schlechten Beigeschmack.[22]

Die Forschungsarbeit ging jedoch unvermindert weiter, und so setzte in den 1990er Jahren allmählich wieder Tauwetter ein. Frischen Wind brachten neue Alternativen zum traditionellen logizistischen Paradigma (oft »*Good Old-Fashioned Artificial Intelligence*«, GOFAI, genannt), das sich auf hochstufige Symbolmanipulation konzentriert und seinen Höhepunkt mit den Expertensystemen der 1980er Jahre erreicht hatte. Die jetzt in Mode kommenden Techniken, darunter neuronale Netze und genetische Algorithmen, versprachen, einige Mängel des GOFAI-Ansatzes zu überwinden, insbesondere die Empfindlichkeit der klassischen KI-Programme (die in der Regel kompletten Unsinn produzierten, sobald ihre Programmierer das kleinste Detail übersahen). Die neuen Techniken erwiesen sich als robuster. Neuronale Netze etwa wiesen die Eigenschaft der Fehlertoleranz auf: Kleinere Schäden führten in der Regel zu einer geringfügig schlechteren Leistung und nicht zum Totalabsturz. Noch wichtiger war, dass diese Systeme aus der Erfahrung lernen konnten, indem sie Beispielfälle auf natürliche Art und Weise verallgemeinerten und verborgene statistische Muster in ihrem Input fanden.[23] Die Netze eigneten sich daher gut zur Mustererkennung und für Klassifikationsprobleme. Ein Beispiel: Wenn ein neuronales Netz mit einer Menge von Sonarsignalen gefüttert wurde, dann lernte es, die akustischen Profile von U-Booten, Minen und Meerestieren besser zu unterscheiden als menschliche Experten – und zwar ohne dass jemand vorher herausfinden musste, wie diese Profile genau zu kategorisieren oder ihre Eigenschaften zu gewichten wären.

Obwohl es schon seit den späten 1950er Jahren einfache neuronale Netze gab, erlebte das Feld mit der Erfindung des »Backpropagation-Algorithmus« eine Renaissance, denn dieser ermöglichte es, mehrschichtige neuronale Netze zu trainieren.[24] Solche Netze, bei denen eine oder mehrere zwischengeschaltete (»verborgene«) Schichten von

Neuronen zwischen der Eingangs- und der Ausgangsschicht liegen, können eine viel breitere Palette von Aufgaben erledigen als ihre einfacheren Vorgänger.[25] Zusammen mit immer leistungsfähigeren Computern ermöglichten diese algorithmischen Verbesserungen den Bau von Systemen, die auf vielen Gebieten praktisch einsetzbar waren.

Im Vergleich zu den herkömmlichen übergenauen, aber störanfälligen regelbasierten Systemen schnitten die echten Gehirnen ähnelnden neuronalen Netze deutlich besser ab. In der Folge wurde ein neuer »Ismus« geprägt, der *Konnektionismus*, der die Bedeutung massiv-paralleler subsymbolischer Verarbeitung betonte. Mehr als 150 000 wissenschaftliche Arbeiten sind seitdem über künstliche neuronale Netze veröffentlicht worden, und sie stellen nach wie vor einen wichtigen Ansatz innerhalb des maschinellen Lernens dar.

Evolutionsbasierte Methoden – darunter genetische Algorithmen und genetische Programmierung – sind ein weiterer Ansatz, dessen Entstehung zum Ende der zweiten KI-Eiszeit beitrug. Ihr Einfluss auf das Fach war vermutlich geringer als jener der neuronalen Netze, sie erregten jedoch große öffentliche Aufmerksamkeit. In evolutionären Modellen betrachtet man eine Population von möglichen Lösungen (das können Datenstrukturen oder Programme sein); durch deren Mutation oder Rekombination werden dann zufällig neue Lösungsmöglichkeiten erzeugt. In regelmäßigen Abständen wird die Population durch Anwendung eines Auswahlkriteriums (einer Fitnessfunktion) beschnitten, und nur die besser angepassten Lösungen kommen eine Runde weiter. Nach Tausenden von Generationen nimmt die durchschnittliche Qualität der Lösungen in der Population allmählich zu. Wenn er funktioniert, kann so ein Algorithmus ein sehr breites Spektrum von Problemen effizient lösen. Die Lösungen können dabei völlig neuartig und kontraintuitiv sein; sie ähneln natürlichen Strukturen oft stärker als alles, was ein Ingenieur entwerfen würde. Im Prinzip muss der Mensch dabei kaum mehr tun, als die – oft sehr simple – Fitnessfunktion zu spezifizieren; in der Praxis erfordern evolutionäre Methoden insbesondere bei der Entwicklung eines guten Darstellungsformats jedoch Geschick und Einfallsreichtum. Ohne eine

effiziente Möglichkeit, in Frage kommende Lösungen zu codieren (ohne eine genetische Sprache, die latenten Strukturen im Zielbereich entspricht), neigt eine evolutionäre Suche nämlich dazu, entweder nie ans Ende zu kommen oder in einem lokalen Optimum steckenzubleiben. Und selbst wenn ein gutes Darstellungsformat gefunden wird, ist ein evolutionärer Prozess aufwändig zu simulieren und fällt in der Praxis oft der kombinatorischen Explosion zum Opfer.

Neuronale Netze und genetische Algorithmen sind Beispiele für Methoden, die in den 1990er Jahren als aufregende Alternativen zum stagnierenden GOFAI-Paradigma galten. Hier soll es aber nicht darum gehen, für diese beiden Methoden zu werben oder sie den vielen anderen Techniken des maschinellen Lernens vorzuziehen. Eine der wichtigsten theoretischen Entwicklungen dieser Zeit war gerade die allmähliche Erkenntnis, dass sich vordergründig disparate Techniken als Spezialfälle eines einzigen mathematischen Paradigmas verstehen lassen. Viele Arten von künstlichen neuronalen Netzen beispielsweise können als Klassifikatoren betrachtet werden, die eine besondere Form der statistischen Berechnung (die Maximum-Likelihood-Methode) durchführen.[26] Diese Perspektive erlaubt einen Vergleich neuronaler Netze mit einer größeren Klasse von Algorithmen, die anhand von Beispielen zu klassifizieren lernen: »Entscheidungsbäume«, »logistische Regressionsmodelle«, »Support Vector Machines«, »Naive Bayes«, »k-nächste-Nachbarn-Regression« und andere.[27] In ähnlicher Weise kann man genetische Algorithmen so betrachten, als vollzögen sie ein »stochastisches Bergsteigen«, das wiederum eine Teilmenge einer größeren Klasse von Optimierungsalgorithmen ist. Jeder dieser Algorithmen zum Klassifizieren oder zum Durchsuchen eines Lösungsraums hat seine eigenen Stärken und Schwächen, die mathematisch untersucht werden können. Die Algorithmen unterscheiden sich in puncto Rechenzeit, Speicherplatzbedarf und induktive Vorentscheidungen sowie in der Einfachheit, mit der externe Inhalte eingearbeitet werden können oder menschliche Experten imstande sind, diese ganzen Abläufe zu durchschauen.

Hinter dem Durcheinander des maschinellen Lernens und kreativen Problemlösens steht also eine Reihe mathematisch genau spezifizierter Kompromisse. Das Ideal ist der perfekte bayesianische Akteur, der alle verfügbaren Informationen wahrscheinlichkeitstheoretisch optimal nutzt. Da dieses Ideal unerreichbar ist, weil es die Kapazitäten jedes physisch möglichen Computers weit übersteigt (siehe *Kasten 1*), gleicht die KI-Forschung folglich der Suche nach einem Mittelweg: Es geht darum, sich dem bayesianischen Ideal anzunähern und gleichzeitig die besten oder allgemeinsten Lösungen zu opfern, um in der realen Welt Probleme effizient lösen zu können.

Kasten 1: Ein optimaler bayesianischer Akteur

Ein idealer bayesianischer Akteur startet mit einer »apriorischen Wahrscheinlichkeitsverteilung«, einer Funktion, die jeder »möglichen Welt« (jeder maximal spezifischen Weise, wie die Welt sein könnte) eine Wahrscheinlichkeit zuweist.[28] Diese Funktion beinhaltet einen induktiven Bias, sodass einfacheren möglichen Welten höhere Wahrscheinlichkeiten zugeordnet sind. (Die Einfachheit einer möglichen Welt lässt sich zum Beispiel durch die Angabe ihrer »Kolmogorov-Komplexität« formal definieren. Dieser Maßeinheit liegt die Länge des kürzesten Computerprogramms zugrunde, das eine vollständige Beschreibung der Welt liefert.)[29] Die Funktion beinhaltet zudem sämtliches Hintergrundwissen, das die Programmierer dem Akteur zugestehen wollen.

Sobald der Akteur neue Informationen erhält, aktualisiert er seine Wahrscheinlichkeitsverteilung, indem er diese gemäß Bayes' Theorem konditionalisiert.[30] Konditionalisierung ist diejenige mathematische Operation, die die Wahrscheinlichkeit jener Welten, die unvereinbar mit den neu erhaltenen Informationen sind, auf null setzt und die Wahrscheinlichkeitsverteilung über die restlichen möglichen Welten renormalisiert. Das Ergebnis ist eine »aposteriorische Wahrscheinlichkeitsverteilung« (die der Akteur im nächsten Schritt wiederum als neue apriorische Wahrscheinlichkeitsverteilung verwenden kann). Mit jeder neuen Beobachtung konzentriert sich die Wahrscheinlichkeitsmasse des Akteurs auf die schrumpfende Menge der möglichen Welten, die im Einklang mit dieser Beobachtung stehen, wobei einfachere Welten automatisch wahrscheinlicher sind.

Stellen wir uns die Wahrscheinlichkeit als einen Haufen Sand auf einem großen Blatt Papier vor. Das Blatt ist in Bereiche verschiedener Größe unterteilt, wobei jeder Bereich einer möglichen Welt entspricht: je größer die Fläche, desto einfacher die Welt. Auf dem ganzen Blatt liegt eine gleichmäßige Sandschicht: Das ist unsere apriorische Wahrscheinlichkeitsverteilung. Immer, wenn eine Beobachtung einige mögliche Welten ausschließt, entfernen wir den Sand von den entsprechenden Bereichen und verteilen

ihn gleichmäßig auf die übrigen. Die Gesamtmenge der Sandkörner auf dem ganzen Blatt ändert sich also nie, sie konzentriert sich nur auf immer weniger Gebiete, nämlich jene, für die die Beobachtungen sprechen. Dies ist ein Bild des Lernens in seiner reinsten Form. (Um die Wahrscheinlichkeit einer *Hypothese* zu berechnen, bestimmen wir einfach die Menge an Sand in allen Bereichen, die denjenigen möglichen Welten entsprechen, in denen die Hypothese wahr ist.)

Damit haben wir eine Lernregel definiert. Um einen Akteur zu bekommen, brauchen wir auch noch eine Entscheidungsregel. Dazu statten wir den Akteur mit einer »Nutzenfunktion« aus, die jeder möglichen Welt eine Zahl zuweist, die angibt, wie wünschenswert die entsprechende Welt gemäß den basalen Präferenzen des Akteurs ist. Bei jedem Schritt wählt der Akteur nun die Handlung mit dem höchsten erwarteten Nutzen aus.[31] (Um diese zu finden, könnte er alle möglichen Handlungen auflisten und anschließend eine bedingte Wahrscheinlichkeitsverteilung berechnen – die Wahrscheinlichkeitsverteilung, die sich ergibt, wenn die aktuelle Wahrscheinlichkeitsverteilung die Beobachtung berücksichtigt, dass die Handlung gerade ausgeführt wurde. Schließlich könnte der Akteur auch den Erwartungswert der Handlung berechnen: die Summe des Wertes jeder möglichen Welt, multipliziert mit der bedingten Wahrscheinlichkeit dieser Welt unter der Annahme, dass die Handlung stattgefunden hat.)[32]

Die Lernregel und die Entscheidungsregel definieren zusammen einen »Optimalitätsbegriff« für einen Akteur (im Wesentlichen benutzen KI-Forschung, Erkenntnistheorie, Wissenschaftstheorie, Ökonomie und Statistik den gleichen Optimalitätsbegriff).[33] In der Realität ist solch ein Akteur unmöglich, weil die erforderlichen Berechnungen undurchführbar sind und genau wie in der klassischen KI-Forschung zu einer kombinatorischen Explosion führen. Um das zu verstehen, betrachte man eine winzige Teilmenge aller möglichen Welten: diejenigen Welten, in denen nur ein einziger Bildschirm existiert. Der Bildschirm hat eine Auflösung von 1000 × 1000 Pixeln, von denen jedes entweder an oder aus ist. Selbst diese Teilmenge aller möglichen Welten ist gewaltig: Die $2^{1000 \times 1000}$ möglichen Bildschirmzustände übertreffen die Anzahl aller Berechnungen, die im beobachtbaren Universum jemals stattfinden dürften. Folglich können wir noch nicht einmal die Welten dieser winzigen Teilmenge aller möglichen Welten aufzählen, geschweige denn aufwändigere Berechnungen für jede einzelne von ihnen durchführen.

Optimalitätsbegriffe können aber auch dann von theoretischem Interesse sein, wenn sie physikalisch nicht realisierbar sind. Sie geben ein Ideal vor, anhand dessen wir heuristische Annäherungen beurteilen können, und manchmal lässt sich daraus schließen, was ein optimaler Akteur in einem speziellen Fall tun würde. Wir werden einige alternative Optimalitätsvorstellungen für künstliche Akteure in Kapitel 12 kennenlernen.

Das gleiche Bild zeigt sich in den Arbeiten der letzten Jahrzehnte zu probabilistischen graphischen Modellen, beispielsweise Bayes'schen

Netzen. Diese liefern wichtige Einsichten zum Begriff der Kausalität[34] und ermöglichen eine prägnante Art der Darstellung probabilistischer und bedingter Unabhängigkeitsbeziehungen, die in einem bestimmten Bereich gelten. (Solche Unabhängigkeitsbeziehungen zu nutzen ist unabdingbar, wenn man eine kombinatorische Explosion vermeiden will, die ein großes Problem sowohl für probabilistische Schlüsse als auch für logische Ableitungen darstellt.)

Lernprobleme aus einzelnen Bereichen mit dem allgemeinen Problem des bayesianischen Schlussfolgerns in Beziehung zu setzen hat den Vorteil, dass neue, effizientere Algorithmen in diesem Bereich dann zu unmittelbaren Verbesserungen auf vielen anderen Gebieten führen. Zudem können Forscher verschiedener Disziplinen ihre Ergebnisse so leichter vergleichen. Fortschritte bei Monte-Carlo-Näherungsverfahren beispielsweise sind von direktem Nutzen für das maschinelle Sehen, die Robotik oder die komputationale Genetik, und graphische Modelle und die bayesianische Statistik sind zu einem gemeinsamen Schwerpunkt der Forschung in vielen Bereichen geworden, zu nennen wären etwa maschinelles Lernen, statistische Physik, Bioinformatik, kombinatorische Optimierung und die Kommunikationstheorie.[35] Formale Resultate aus anderen Gebieten machten zahlreiche Fortschritte auf dem Feld des maschinellen Lernens möglich (und zudem profitierten Anwendungen auf diesem Gebiet enorm von schnelleren Computern und der besseren Verfügbarkeit großer Datensätze).

Der Stand der Technik

Die künstliche Intelligenz übertrifft die menschliche bereits auf zahlreichen Gebieten. Wie der Überblick in *Tabelle 1* zeigt, schlagen heutige KI-Programme ihre besten menschlichen Gegner etwa in einer Vielzahl von Spielen.[36]

Tabelle 1: *KI-Programme im Vergleich mit menschlichen Spielern*

Tabelle 1: *KI-Programme im Vergleich mit menschlichen Spielern*

Dame	Übermenschlich	Arthur Samuels Dame-Programm, im Jahr 1952 geschrieben und sukzessive verbessert (1955 wird maschinelles Lernen eingebaut), wird das erste Programm, das ein Spiel besser beherrscht als sein Schöpfer.[37] 1994 schlägt *CHINOOK* den amtierenden Weltmeister und gewinnt damit als erstes Programm eine offizielle Weltmeisterschaft in einem Geschicklichkeitsspiel. Im Jahr 2002 »lösen« Jonathan Schaeffer und sein Team das Damespiel: Sie schreiben ein Programm, das immer den bestmöglichen Zug macht (und dafür eine Alpha-Beta-Suche mit einer Datenbank von 39 Billionen Endspielstellungen kombiniert). Eine perfekte Spielweise auf beiden Seiten führt zum Unentschieden.[38]
Backgammon	Übermenschlich	1979: Das Backgammon-Programm *BKG* von Hans Berliner besiegt den Weltmeister – das erste Computerprogramm, dem dies (in einem Schaukampf) in irgendeinem Spiel gelingt, auch wenn Berliner dies später dem Würfelglück zuschreibt.[39] 1992: Das Backgammon-Programm *TD-Gammon* von Gerry Tesauro verbessert sich mithilfe von Zeitliche-Differenz-Methoden (eine Form des Verstärkungslernens) und wiederholtem Spielen gegen sich selbst und erreicht so Weltklasseniveau.[40] Inzwischen haben Backgammon-Programme die besten menschlichen Spieler weit hinter sich gelassen.[41]

Tabelle 1: *KI-Programme im Vergleich mit menschlichen Spielern*

Traveller TCS	Übermenschlich in Zusammenarbeit mit menschlichem Spieler[42]	Sowohl 1981 als auch 1982 gewinnt Douglas Lenats Programm *Eurisko* die US-Meisterschaft in Traveller TCS (einem futuristischen Marinekriegsspiel), woraufhin die Regeln geändert werden, um *Euriskos* unorthodoxe Spielweise zu verhindern.[43] *Eurisko* besaß Heuristiken sowohl zur Zusammenstellung seiner Flotte wie zur Änderung seiner Heuristiken.
Othello	Übermenschlich	1997: Das Programm *Logistello* gewinnt jedes Spiel eines Sechs-Spiele-Matches gegen den Weltmeister Takeshi Murakami.[44]
Schach	Übermenschlich	1997: *Deep Blue* schlägt den Schachweltmeister Garri Kasparow. Kasparow behauptet, in einigen Zügen des Computers Anzeichen für echte Intelligenz und Kreativität entdeckt zu haben.[45] Seitdem sind Schachprogramme weiterhin stetig besser geworden.[46]
Kreuzworträtsel	Expertenniveau	1999: Das Programm *Proverb* übertrifft den durchschnittlichen Kreuzworträtsellöser.[47] 2012: Das Programm *Dr. Fill* von Matt Ginsberg lässt im American Crossword Puzzle Tournament drei Viertel seiner menschlichen Kontrahenten hinter sich. (*Dr. Fills* Leistungen sind uneinheitlich. Es löst das für Menschen schwierigste Rätsel vollständig, kommt aber bei unüblicheren Varianten, bei denen die Antworten rückwärts oder diagonal gegeben werden müssen, ins Schleudern.)[48]
Scrabble	Übermenschlich	Ab dem Jahr 2002 übertrifft Scrabble-Software die besten menschlichen Spieler.[49]

Tabelle 1: *KI-Programme im Vergleich mit menschlichen Spielern*

Bridge	Weltklasseniveau	Seit 2005 ist Bridge-Software den besten menschlichen Bridge-Spielern ebenbürtig.[50]
Jeopardy!	Übermenschlich	2010: Der von IBM gebaute Computer *Watson* besiegt die zwei besten menschlichen Spieler, Ken Jennings und Brad Rutter.[51] Jeopardy! ist eine Fernsehshow, in der Quizfragen zu Geschichte, Literatur, Sport, Geographie, Popkultur, Wissenschaft und anderen Themen beantwortet werden müssen. Die Fragen werden in Form von Hinweisen und Wortspielen gestellt.
Poker	Unterschiedlich	Poker-Programme erreichen nicht ganz das Niveau der besten Spieler in großen Texas-Hold'em-Runden, erzielen aber übermenschliche Resultate bei anderen Pokervarianten.[52]
Freecell	Übermenschlich	Durch genetische Algorithmen erzeugte Heuristiken für das Solitär-Spiel Freecell (das in seiner allgemeinen Form NP-vollständig ist) ergeben ein Programm, das in der Lage ist, hochkarätige menschliche Spieler zu besiegen.[53]
Go	Sehr starkes Amateurniveau	2012 hat die Zen-Serie von Go-Programmen bei schnell ausgetragenen Spielen den 6. Dan (das Niveau eines sehr starken Amateurspielers) erreicht; sie nutzt Monte-Carlo-Baumsuchen und Methoden des maschinellen Lernens.[54] Go-Programme haben sich in den letzten Jahren um etwa einen Dan pro Jahr verbessert. Wenn es dabei bleibt, dürften sie in etwa 10 Jahren die Weltmeisterschaft erringen.

Diese ganzen Erfolge mögen heute vielleicht nicht mehr besonders eindrucksvoll erscheinen, aber das liegt daran, dass wir unsere Standards

beständig nach oben schrauben. Die Schachkunst zum Beispiel hielt man einst für den Inbegriff des menschlichen Intellekts – in den Worten mehrerer Experten der späten 1950er Jahre: »Wenn es gelänge, eine schachspielende Maschine zu entwickeln, dürfte man zum Kern der menschlichen geistigen Anstrengungen vorgedrungen sein.«[55] John McCarthy klagte einmal: »Sobald es funktioniert, nennt es keiner mehr KI«,[56] und da scheint etwas Wahres dran zu sein.

In einer wichtigen Hinsicht jedoch waren Schachcomputer keine so großartige Erfindung, wie viele sich das vorgestellt hatten. Früher einmal war vielleicht nicht ohne Grund angenommen worden, ein auf Großmeisterniveau spielendes Programm müsse über ein hohes Maß an *allgemeiner* Intelligenz verfügen[57] und das Schachspiel erfordere es, abstrakte Begriffe zu erlernen, Strategien und flexible Pläne zu entwickeln, eine breite Palette von ausgeklügelten logischen Schlussfolgerungen zu ziehen und vielleicht sogar das Denken des Gegners zu modellieren. Aber nein, es stellte sich heraus, dass ein speziell für das Schachspielen entwickelter Algorithmus völlig ausreicht.[58] Mithilfe der schnellen Prozessoren, die gegen Ende des 20. Jahrhunderts verfügbar wurden, ist eine solche KI kaum zu schlagen, dennoch ist sie beschränkt: Sie spielt Schach – und kann nicht anders.[59]

Andere Bereiche erwiesen sich als *komplizierter* als zunächst gedacht, und es gab geringere Fortschritte. Der Informatiker Donald Knuth formulierte es so: »Die KI kann mittlerweile so ziemlich alles, was ›Denken‹ erfordert, aber kaum etwas von dem, was Menschen und Tiere ›gedankenlos‹ tun – das ist irgendwie viel schwieriger!«[60] Die Bildanalyse, das Erkennen von Objekten oder das Kontrollieren eines Roboters in einer natürlichen Umgebung haben sich als äußerst anspruchsvoll erwiesen. Dennoch wurden und werden auch hier inzwischen viele Fortschritte gemacht, wobei kontinuierliche Verbesserungen der Hardware von Nutzen sind.

Der gesunde Menschenverstand und das Verstehen einer natürlichen Sprache haben sich ebenfalls als kompliziert herausgestellt. Heute ist die Ansicht verbreitet, dass es sich dabei um »KI-vollständige« Probleme