Catrin Misselhorn

Grundfragen der Maschinenethik

Reclam

Catrin Misselhorn

Grundfragen der Maschinenethik

Reclam

4., durchgesehene und überarbeitete Auflage

2018, 2019 Philipp Reclam jun. GmbH & Co. KG,
Siemensstraße 32, 71254 Ditzingen
Gesamtherstellung: Philipp Reclam jun. Verlag GmbH,
Siemensstraße 32, 71254 Ditzingen
Made in Germany 2019
RECLAM ist eine eingetragene Marke
der Philipp Reclam jun. GmbH & Co. KG, Stuttgart
ISBN 978-3-15-950527-5
ISBN der Buchausgabe 978-3-15-019583-3
www.reclam.de

Inhalt

Für Leonore

Einleitung: Maschinenethik – eine neue Disziplin an der Schnittstelle von Philosophie, Informatik und Robotik

Seit jeher verbindet sich mit dem Einsatz von Maschinen die Hoffnung, dass sie Menschen von Tätigkeiten entlasten, die schwer, schmutzig, gefährlich oder einfach nur unangenehm sind. Manchmal sind die Leistungen von Maschinen auch einfach schneller oder präziser als diejenigen des Menschen. Das eröffnet die Aussicht, dass die Menschen freigestellt werden für kreative und sinnstiftende Aktivitäten, aber es bringt auch die Sorge mit sich, der Mensch könne von Maschinen verdrängt werden.

So geht der Begriff *Roboter* auf ein 1920 erschienenes tschechisches Theaterstück zurück, in dem es um menschenähnliche Maschinensklaven bzw. Androiden geht, die dafür geschaffen wurden, den Menschen das Arbeiten abzunehmen.[1] Doch schlussendlich wenden sie sich gegen ihre Schöpfer und versuchen die Menschheit auszulöschen.[2] Die Digitalisierung befeuert sowohl diese Hoffnungen als auch diese Ängste noch einmal, denn sie ermöglicht die Konstruktion von Maschinen, die intelligent genug sind, um auch komplexe Aufgaben zu meistern, ohne auf die permanente Überwachung durch den Menschen angewiesen zu sein.

Je intelligenter und autonomer Maschinen werden, desto größer ist die Wahrscheinlichkeit, dass sie in Situationen geraten, die ihnen moralische Entscheidungen abverlangen. Doch können Maschinen überhaupt moralisch handeln und sollen sie es? Wie baut man eine moralische Maschine? Mit Fragen wie diesen beschäftigt sich die Maschinenethik, ein neues Forschungsgebiet an der Schnittstelle von Philosophie, Informatik und Robotik.

Die Maschinenethik unterscheidet sich von anderen Formen der Technikethik, weil ihr Gegenstand die Entwicklung einer Ethik *für* Maschinen im Gegensatz zu einer Ethik für Menschen im *Umgang mit* Maschinen ist. Es geht nicht nur um die Frage, wie die Menschen mit einer bestimmten Technologie (beispielsweise Atomkraft) aus moralischer Sicht umgehen sollten. Das Ziel ist vielmehr, darüber nachzudenken, ob und wie man Maschinen konstruieren kann, die selbst moralische Entscheidungen treffen und umsetzen können, und ob man dies tun sollte.

Dieses Thema ist in der Geschichte der Ethik neu, weil es erst mit der Entstehung des Computers überhaupt in den Bereich des Möglichen gerückt ist. Die rasante Entwicklung der künstlichen Intelligenz im letzten Jahrzehnt beflügelte die Maschinenethik und macht sie zugleich unabdingbar, damit diese technologische Entwicklung segensreich verläuft.[3] Man spricht analog zu *Artificial Intelligence* auch von *Artificial Morality*. Während Erstere das Ziel hat, intelligentes Verhalten zu modellieren oder zu simulieren, geht es Letzterer darum, künstliche Systeme mit der Fähigkeit zum moralischen Entscheiden und Handeln auszustatten.

Lange Zeit stand die Maschinenethik zu Unrecht im Verdacht, bloß Science-Fiction zu sein. Doch schon eine so simple Maschine wie ein Staubsaugerroboter steht vor moralischen Entscheidungen: Soll er einen Marienkäfer einfach einsaugen oder soll er ihn verscheuchen bzw. umfahren? Und wie sieht es mit einer Spinne aus? Soll er sie töten oder ebenfalls verschonen? Ein solcher Roboter ist in einem minimalen Sinn autonom, weil er im Unterschied zu einem konventionellen Staubsauger nicht von einem Menschen geführt wird. Die Entscheidung ist moralisch, weil sie sich darauf bezieht, ob man Tiere zu Reinigungszwecken töten darf. Gewöhnliche Staubsaugerroboter besitzen allerdings noch nicht die Fähigkeit, eine sol-

che Entscheidung zu treffen. Es gibt jedoch erste Ansätze, eine um ein Ethikmodul erweiterte Version des populären Modells *Roomba* zu schaffen, die das Leben von Insekten berücksichtigt (der Prototyp ist mit einem optionalen »Kill-Button« für Spinnen ausgestattet).[4]

Je komplexer die Einsatzbereiche autonomer Systeme sind, desto anspruchsvoller werden die moralischen Entscheidungen, die sie treffen müssen. Dies zeigt sich beispielsweise an Pflegesystemen, Kriegsrobotern und autonomen Fahrzeugen, drei paradigmatischen Anwendungsfeldern der Maschinenethik, die auch in diesem Buch behandelt werden. In allen drei Bereichen stehen autonome Systeme vor grundlegenden moralischen Entscheidungen, in denen es manchmal sogar um Leben und Tod von Menschen geht. Kann man Maschinen solche Entscheidungen überlassen, darf man es oder sollte man es gar? Das sind die grundlegenden Fragen, um die es in diesem Buch gehen wird.

Wie wichtig dieses Thema ist, zeigt sich auch an der großen Aufmerksamkeit, die mögliche Dilemmasituationen beim autonomen Fahren in einer breiteren Öffentlichkeit erhalten haben. Was ist, wenn ein autonomes Fahrzeug ausschließlich die Möglichkeit hat, entweder einen Menschen am Ende seines Lebens oder ein kleines Kind zu töten? Was, wenn es nur dadurch fünf Menschenleben retten kann, dass es eine auf dem Gehweg stehende Person anfährt? Ist ein besonderer Schutz für die Insassen moralisch legitim oder kommt den anderen Verkehrsteilnehmern vom moralischen Standpunkt mehr Gewicht zu?

Gelegentlich wird eingewandt, solche Szenarien seien bloß hypothetisch und würden aller Wahrscheinlichkeit nach nie eintreten. Doch die ethische Bewertung von Technologien bezieht sich häufig auf Fälle, deren Eintreten nicht sehr wahrscheinlich ist. So wurde auch eine nukleare Katastrophe lange Zeit für unwahrscheinlich gehalten, und doch gehört ein sol-

ches Szenarium zu den Kernproblemen der ethischen Diskussion der Atomkraft. Ein anderes Beispiel ist der Fall des Erlanger Babys im Jahr 1992, bei dem eine hirntote Schwangere, die in der 15. Woche einen tödlichen Unfall erlitt, über mehrere Wochen auf der Intensivstation behandelt wurde, um das ungeborene Kind auszutragen.[5] Eine solche Konstellation ist ebenfalls sehr selten, löste jedoch in der breiteren Öffentlichkeit eine heftige ethische Kontroverse aus.

Entscheidend ist also nicht die Eintrittswahrscheinlichkeit, sondern dass grundlegende ethische Fragen betroffen sind. An derartigen Fällen lässt sich die Überzeugungskraft unserer moralischen Grundsätze überprüfen, die in weniger extremen Situationen gar nicht in Zweifel gezogen würden. Genau das gilt auch für Pflegesysteme, Kriegsroboter und autonome Fahrzeuge: Diese müssen, wie wir sehen werden, über die Fähigkeit zum moralischen Entscheiden und Handeln verfügen, sollen sie ihre Funktion erfüllen. Zugleich versteht es sich nicht von selbst, dass das überhaupt möglich ist. Und selbst wenn es geht, ist noch nicht ausgemacht, dass dies von einem moralischen Standpunkt aus auch gutzuheißen ist.

Ein anderer Einwand gegen die Maschinenethik lautet, dass es nicht die Maschine ist, welche die moralische Entscheidung trifft, sondern der Mensch, der sie programmiert. Doch je größer die Fortschritte der Künstlichen Intelligenz sind, desto stärker verwischt diese Grenze. Bereits bei einem Schachprogramm ist die Vorstellung inadäquat, die Entwickler hätten es mit Anweisungen ausgestattet, die unmittelbar den bestmöglichen Zug bestimmen. Das zeigt sich schon daran, dass ein solches Programm weit besser Schach spielt als seine Programmierer, die es sicherlich nicht mit einem Schachweltmeister aufnehmen könnten.

Das Gleiche trifft in noch höherem Maß auf das Brettspiel *Go* zu. Dieses galt lange Zeit als zu komplex und kognitiv, zu

anspruchsvoll für ein künstliches System. Doch dem von *Google DeepMind* mit den Mitteln des maschinellen Lernens entwickelten Programm *AlphaGo* gelang es 2016, einige der weltbesten Spieler zu schlagen. Ende 2017 folgte *AlphaGo Zero*, das im Unterschied zu *AlphaGo* nur mit den Spielregeln ausgestattet war, und nicht mehr anhand einer großen Datenmenge menschlicher Spiele trainiert werden musste.[6] Das Programm spielte zu Beginn lediglich Partien gegen sich selbst und erreichte so eine Spielstärke, die die alte Version in kürzester Zeit weit übertraf. Die Entwickler sehen darin den Grundstein für die Entwicklung einer allgemeinen künstlichen Intelligenz, die ohne menschliches Expertenwissen auskommt und in allen möglichen Bereichen einsetzbar ist.[7]

Beide Systeme sind menschlichen Spielern überlegen, und auch die Entwickler können nicht unmittelbar bestimmen, für welche Züge sich das System entscheiden wird. Wie wir sehen werden, ist gerade dieser Mangel an Kontrolle und Vorhersehbarkeit für die Maschinenethik entscheidend.

Dieser Mangel an Kontrolle gehört zu den Gründen, aus denen eine beeindruckende Zahl von KI-Forschern und Wissenschaftlern vor einiger Zeit einen offenen Brief verfasst haben, in dem die Maschinenethik als eines der wichtigsten und drängendsten Forschungsgebiete der KI hervorgehoben wird. Zu den prominentesten Unterzeichnern gehören Nick Bostrom, Stephen Hawking, Elon Musk und Stewart Russell.[8]

Dieses Buch hat das Ziel, die Grundfragen der Maschinenethik zu thematisieren und zur kritischen Reflexion der Möglichkeit und Wünschbarkeit moralisch handelnder Maschinen anzuleiten. Es richtet sich an die Vertreter der verschiedenen Disziplinen, die an der Maschinenethik beteiligt sind, aber auch an eine breitere Öffentlichkeit. Denn die Maschinenethik geht uns alle an, weil sie die grundlegenden Fragen betrifft, wie wir uns selbst sehen und in welcher Gesellschaft wir leben

wollen. Um die Verständlichkeit zu erleichtern, findet sich am Ende des Buchs ein Glossar, in dem wichtige Fachbegriffe, die häufiger vorkommen, fasslich erläutert werden.

Der erste Teil entfaltet die allgemeinen theoretischen Grundlagen der Maschinenethik in den verschiedenen Themenbereichen. Das erste Kapitel erklärt zunächst, was künstliche Intelligenz ist, und führt in die unterschiedlichen Forschungsrichtungen ein (wer sich im Bereich der KI schon ein wenig auskennt, kann dieses Kapitel getrost überspringen).

Im zweiten Kapitel werden die philosophischen Implikationen dargelegt und diskutiert, wie sich die künstliche Intelligenz zur menschlichen Intelligenz verhält. Insbesondere ist zu prüfen, inwiefern Maschinen eigentlich über Fähigkeiten wie Denken, Bewusstsein oder Emotionen verfügen können, die für menschliche Intelligenz eine wichtige Rolle spielen. (Diejenigen, die in der Philosophie des Geistes Bescheid wissen, brauchen dieses Kapitel nur zu überfliegen. Es ganz auszulassen, ist vielleicht nicht ratsam, weil einige Weichenstellungen getroffen werden, die später Bedeutung erlangen.)

Das dritte Kapitel stellt die ethischen Grundlagen bereit. Eine allgemeine Einführung legt dar, was unter Ethik und Moral zu verstehen ist. Dann werden die drei wichtigsten ethischen Theorien erläutert, die für die Maschinenethik maßgeblich sind: Der Utilitarismus, die kantische Ethik und die Tugendethik. Außerdem wird die Maschinenethik als eine Form der angewandten Ethik eingeordnet und ihre Beziehungen zu anderen Disziplinen wie der Informationsethik oder der Computerethik dargestellt. Wer sich bereits mit diesen Disziplinen beschäftigt hat oder über Grundwissen der allgemeinen Ethik verfügt, braucht dieses Kapitel nicht so intensiv zu studieren. Auch dort werden allerdings gewisse moralphilosophische Vorentscheidungen getroffen, so dass möglicherweise Rückfragen entstehen, wenn man es überspringt.

Der zweite Teil ist der Explikation der Grundbegriffe und Methoden der Maschinenethik gewidmet. Zunächst wird diskutiert, inwieweit Maschinen moralisch handeln können. Das erste Kapitel führt eine Stufung moralischer Akteure ein. Es entfaltet einen graduellen Ansatz moralischer Handlungsfähigkeit, nach dem auch künstliche Systeme in einem funktionalen Sinn moralische Akteure sein können, wenn sie bestimmte Bedingungen erfüllen. Sie erreichen jedoch keine vollumfängliche moralische Handlungsfähigkeit wie Menschen.

Das zweite Kapitel erörtert, wie moralische Fähigkeiten in eine Maschine implementiert werden können. Dies erfordert einen interdisziplinären Ansatz, dessen Kernbereiche Philosophie und Informatik bilden. Zunächst wird die Herangehensweise an dieses Vorhaben methodologisch erklärt. Ein Blick auf die Kognitionswissenschaften erweist sich dabei als hilfreich. Dann werden drei konkrete Ansätze eingeführt, wie man Moral in einem künstlichen System implementieren kann, die jeweils eine bestimme Vorgehensweise der Softwareentwicklung mit einem bestimmten Typus ethischer Ansätze verbinden: Top-Down-Ansätze gehen von allgemeinen Regeln aus, die auf konkrete Fälle angewendet werden sollen. Bottom-Up-Ansätze versuchen hingegen, auf der Grundlage von Einzelfällen zu groben Verallgemeinerungen zu gelangen, während hybride Ansätze beide Vorgehensweisen miteinander verbinden. Welcher Ansatz vorzuziehen ist, hängt vom jeweiligen Anwendungsbereich ab.

Nachdem die Grundlagen der Moralimplementation erarbeitet wurden, erörtert das dritte Kapitel die Frage, wie sich das moralische Handeln von Mensch und Maschine zueinander verhält. Wie sich bereits im ersten Kapitel dieses Teils zeigte, sind Maschinen zwar moralische Akteure, aber sie sind nicht zu vollumfänglichem moralischen Handeln wie Men-

schen in der Lage. Dazu fehlen ihnen unter anderem Bewusstsein, Willensfreiheit und die Fähigkeit zur Selbstreflexion. Da diese Eigenschaften für die Übernahme moralischer Verantwortung wesentlich sind, können Maschinen auch nicht für ihr Handeln verantwortlich gemacht werden. Trotzdem ist zu diskutieren, inwiefern der Einsatz moralischer Maschinen die Verantwortungszuschreibung an Menschen unterminiert, so dass am Ende möglicherweise niemand für ihr Handeln die Verantwortung trägt. Diese Frage wird im dritten Teil bezogen auf die Anwendungspraxis erneut aufgegriffen.

Im dritten Teil werden drei zentrale Anwendungsbereiche der Maschinenethik diskutiert: Das erste Kapitel setzt sich mit Pflegesystemen auseinander, das zweite mit autonomen Waffensystemen und das dritte mit dem autonomen Fahren. In allen drei Bereichen wird der Entwicklungsstand der Maschinenethik anhand konkreter Beispiele erläutert. Daran anknüpfend werden anwendungsspezifische Herausforderungen und Probleme untersucht. Schließlich wird vor dem Hintergrund eines breiteren gesellschaftlichen Kontexts erwogen, ob der Einsatz von Maschinen mit Moral aus einer ethischen Perspektive jeweils gutzuheißen ist oder nicht. Dabei wird ein Verständnis von angewandter Ethik vorausgesetzt, das nicht in einer deduktiven Anwendung von Moralprinzipien auf Anwendungsbereiche besteht, sondern in der Mobilisierung moralischer Intuitionen und kontextspezifischer Argumente. Dies hat den Vorteil, den fundamentalen Dissens zu umgehen, der im Hinblick auf die richtige allgemeine Theorie der Moral besteht.

Ein wichtiges Anliegen dieses Buchs ist es, zu zeigen, dass diese Fragen nicht auf einer allgemeinen Ebene entschieden werden können, sondern nur bezogen auf spezifische Anwendungskontexte. Am positivsten stellt sich der Einsatz von Maschinen mit Moral im Bereich Pflege dar. Autonome Waffen-

systeme und Fahrzeuge hingegen werfen grundlegende moralische Einwände auf.

Als Ausblick wird am Ende das Verhältnis von Maschinenethik und Singularität thematisiert. Dieses Kapitel habe ich hinzugefügt, weil ich in Diskussionen und Interviews immer wieder auf dieses Thema angesprochen wurde. Die Singularitätsthese behauptet, dass es in absehbarer Zeit aufgrund der Fortschritte der KI Maschinen geben wird, die die Menschen an Intelligenz weit übertreffen. Damit sind im Kontext der Maschinenethik drei Befürchtungen verbunden: Die eine besagt, dass die Konstruktion von Maschinen mit der Fähigkeit zum moralischen Handeln einen ersten Schritt hin zu einer solchen Superintelligenz darstellen könnte. Die zweite Sorge besteht darin, solche superintelligenten Maschinen könnten auch moralische Ansprüche gegenüber den Menschen erheben. Drittens wird befürchtet, die Superintelligenz könnte die Herrschaft über die Menschheit übernehmen.

Wir werden die Argumente für die Singularitätsthese untersuchen und feststellen, dass es in absehbarer Zeit weder empirisch noch philosophisch Grund zur Annahme einer Superintelligenz gibt. Der in diesem Buch verfolgte Ansatz stellt deshalb einen Mittelweg dar: Auf der einen Seite wird für die Möglichkeit argumentiert, Maschinen mit der Fähigkeit zum moralischen Entscheiden und Handeln in einem funktionalen Sinn auszustatten. Auf der anderen Seite bedeutet das nicht, dass auch Fähigkeiten wie Bewusstsein, Willensfreiheit oder Selbstreflexion, die für menschliche Intelligenz und Moralität zentral sind, auf diesem Weg reproduziert werden können. Die Arbeit an einer Ethik für Maschinen geht deshalb nicht zwangsläufig mit der Annahme einher, auch der Mensch sei letztlich nur eine moralische Maschine. Umgekehrt werden wir aber im Zusammenhang mit einigen Überlegungen zu der schwedischen Science-Fiction-Serie *Real Humans* feststellen,

dass sich echte Menschlichkeit auch und gerade im Umgang mit Maschinen niederschlägt.

Die Debatte um die Singularität ist irreführend, weil nicht darin die größte Gefahr für die Menschheit liegt. Dieses Buch will vielmehr deutlich machen, dass Maschinen, die moralisch handeln können, ebenso faszinierend und verlockend wie moralisch problematisch sind. Denn sie bringen grundlegende Veränderungen unseres Selbstverständnisses und unseres gesellschaftlichen Zusammenlebens mit sich. Wir sollten darüber nachdenken, ob und in welchen Bereichen wir diese Veränderungen wollen. Die Zeit drängt angesichts der Tatsache, dass längst technisch und wirtschaftlich motivierte Fakten geschaffen werden.

1. Theoretische Grundlagen

1. Künstliche Intelligenz

Künstliche Intelligenz (oder kurz: KI) spielt in vielen Bereichen eine wichtige Rolle.[1] Sie steckt in Apps, Suchalgorithmen, Robotern, Fahrassistenten oder *Smart Watches*. Diese digitalen Technologien werden immer intelligenter. Dies betrachten einige mit Euphorie, andere mit Sorge. Doch was ist eigentlich Intelligenz? An dieser Frage scheiden sich die Geister. Es ist noch nicht einmal klar, ob Intelligenz überhaupt ein einheitliches Phänomen darstellt.

Was ist Intelligenz?

Der Harvard-Psychologe Howard Gardner unterscheidet sieben Formen der Intelligenz: visuell-räumlich, körperlich-kinästhetisch, musikalisch, interpersonal und intrapersonal, sprachlich und logisch-mathematisch.[2] Um sich nicht auf bestimmte Intelligenzkriterien festlegen zu müssen, wird in der KI häufig der Mensch als Maßstab genommen, um intelligentes Verhalten zu definieren. So schreibt John McCarthy (1927–2011)[3] 1955 in einem der Gründungsdokumente der KI, künstliche Intelligenz habe die Aufgabe, Maschinen zu konstruieren, die sich auf eine Art und Weise verhalten, die man bei Menschen als intelligent bezeichnen würde.[4] Nun wäre die Beherrschung der Grundrechenarten bei einem Menschen sicherlich ein Zeichen von Intelligenz. Doch die Schlussfolgerung, dass ein simpler Taschenrechner intelligent ist, ja intelligenter als die meisten Menschen, weil er schneller rechnet und weniger Fehler macht, erscheint absurd. Intelligenz erschöpft sich nicht darin, ein bestimmtes kognitives Problem zu lösen, sondern es kommt darauf an, *wie* das geschieht.

Das lässt sich mit Hilfe eines anderen Beispiels verdeutlichen. Bei dem Spiel *Drei Gewinnt* wetteifern zwei Spieler darum, auf einem quadratischen Spielfeld, das aus 3 × 3 Feldern besteht, als Erster drei Zeichen in eine Zeile, Spalte oder Diagonale zu setzen. Die Zahl der möglichen Spiele ist begrenzt und beläuft sich auf 255 168. Es ist ziemlich einfach, ein Programm zu schreiben, das alle möglichen Spielfolgen erzeugt und diejenigen auszeichnet, die gewinnen. Den jeweils optimalen Spielzug muss das Programm dann nur noch von einer Liste ablesen. Allerdings lassen sich nicht viele kognitive Probleme auf diese Art und Weise lösen. Außerdem würden die meisten diese Methode, das Spiel zu spielen, wohl nicht als intelligent bezeichnen.

Auch wenn es schwierig ist, den Begriff der Intelligenz hieb- und stichfest zu definieren, lässt dieses Beispiel einige Rückschlüsse darauf zu, welche Aspekte wichtig sind. So wird Intelligenz häufig mit der Fähigkeit verbunden, zu *lernen*. Ein Programm, das die Regeln von *Drei Gewinnt* nicht kennt und aus der Beobachtung einer Reihe von Spielen darauf schließt, welche Regeln das Spiel ausmachen und welches die besten Strategien sind, um zu gewinnen, würden wir durchaus als intelligent gelten lassen. Auch die Fähigkeit, sich auf *neue Situationen* einzustellen, gehört zum intelligenten Verhalten. Tritt eine bislang nicht bekannte Spielsituation auf und gelingt es, darauf die passende Reaktion zu finden, so ist auch das eine intelligente Leistung. Schließlich ist auch die Fähigkeit zur *Verallgemeinerung* von Bedeutung für Intelligenz. Ein Programm, das in der Lage ist, alle möglichen Brettspiele zu lernen, scheint intelligenter zu sein als eines, das nur *Drei Gewinnt* spielen kann.

Turing Maschine

Künstliche Intelligenz stützt sich auf Computertechnologie. Das abstrakte Modell, das einem herkömmlichen Digitalcomputer zugrunde liegt, ist die Turingmaschine. Diese wurde um 1936 von dem Mathematiker Alan Turing (1912–1954) entwickelt.[5] Es handelt sich dabei nicht um einen konkreten Gegenstand, sondern um ein mathematisches Objekt, das es erlaubt, die Begriffe des Algorithmus und der Berechenbarkeit zu formalisieren, die für die Funktionsweise eines Computers ausschlaggebend sind. Ein Algorithmus ist umgangssprachlich ausgedrückt eine formale Anleitung zur schrittweisen Lösung eines Problems: Das kann Euklids (3. Jh. v. Chr.) Algorithmus zur Berechnung des größten gemeinsamen Teilers zweier natürlicher Zahlen sein, aber auch ein Kochrezept.

Eine Turingmaschine erlaubt die mathematische Darstellung eines Algorithmus. Sie besteht aus einem Lese- und Schreibkopf, einem Speicherband sowie Verarbeitungsregeln. Eine Berechnung erfolgt durch die schrittweise Manipulation von Symbolen, die mit Hilfe des Lese- und Schreibkopfs vom Speicherband abgelesen, nach den Regeln verarbeitet und dann wieder auf das Speicherband geschrieben werden. Diese Symbole lassen sich als Zahlen interpretieren. Die Turingmaschine beschreibt eine Funktion, die von einem bestimmten Ausgangswert zu einem bestimmten Ergebnis führt. Gemäß der sog. *Church-Turing-These* ist jede intuitiv berechenbare Funktion auch von einer Turingmaschine berechenbar.[6] Da der Begriff der *intuitiv berechenbaren Funktion* nicht formalisierbar ist, lässt sich diese These nicht beweisen. Sie wird jedoch in der Informatik üblicherweise akzeptiert. Allerdings ist nicht jede Funktion berechenbar, wie bereits Turing selbst gezeigt hat.[7] So gibt es keinen Algorithmus, der für alle Algorithmen bestimmt, ob sie zu einem Ende gelangen. Man spricht vom

Halteproblem, weil die Turingmaschine bei dem Versuch, einen solchen Algorithmus auszuführen, nicht zu einem Ende kommt, sondern unendlich weiterläuft.[8]

Die ersten physikalischen Umsetzungen digitaler Computer, die dem Prinzip der Turingmaschine entsprechen, stammen freilich nicht von Turing, sondern von dem Bauingenieur Konrad Zuse (1910–1995) und dem Mathematiker John von Neumann (1903–1957). Die Von-Neumann-Architektur ist bis heute maßgeblich für die Arbeitsweise der meisten Computer. Die Ausgangshypothese der KI lässt sich nun so formulieren, dass intelligentes Verhalten algorithmisierbar und somit grundsätzlich durch eine Turingmaschine berechenbar ist.

Die Idee, rationales Denken zu mechanisieren, ist nicht neu. Schon Aristoteles (384–322 v. Chr.) schuf mit der Syllogistik einen Mechanismus, der notwendig von bestimmten Prämissen zu einer Konklusion führt. Das berühmteste Beispiel ist der Schluss:

Prämisse 1: Alle Menschen sind sterblich.
Prämisse 2: Sokrates ist ein Mensch.
Konklusion: Sokrates ist sterblich.

Aristoteles' Syllogistik blieb jedoch informell. Erst im 19. Jahrhundert gelang es, die Logik zu formalisieren und für die Mathematik fruchtbar zu machen. Einen Meilenstein bilden die Arbeiten Gottlob Freges (1848–1925).[9] Er entwickelte eine formale Sprache und damit zusammenhängend formale Beweise, die dann von Bertrand Russell (1872–1970) und Alfred North Whitehead (1861–1947) in ihrem Buch *Principia Mathematica* kritisiert und weitergeführt wurden.[10]

Die Grundidee der Aussagenlogik besteht darin, Symbole für ganze Sätze und ihre Bindeglieder (wie z. B. »und«, »oder« und »wenn … dann …«) einzuführen. Der Satz »Wenn es reg-

net, wird die Straße nass« lässt sich beispielsweise als »wenn p, dann q« (symbolisch: $p \rightarrow q$) darstellen. Die Prädikatenlogik differenziert die logische Formulierung weiter aus, indem unterschiedliche Symbole für das logische Subjekt und das logische Prädikat eines Satzes verwendet werden. Der Satz »Sokrates ist ein Mensch« kann beispielsweise durch »Ma« ausgedrückt werden, wobei »M« für das Prädikat »ist ein Mensch« und »a« für Sokrates steht. Hinzukommen Quantoren, die festlegen, für welche Objekte einer Grundmenge ein Prädikat gilt. So lässt sich die erste Prämisse im Schluss »Alle Menschen sind sterblich« durch »$(\forall x)\,(Mx \rightarrow Sx)$« (»für alle x gilt: wenn x M ist, dann ist x auch S«) symbolisieren.[11] Weitere Symbole gibt es für Klassen, Glieder der Klassen und für die Beziehungen zwischen der Zugehörigkeit und dem Einschluss der Glieder. Die logische Formalisierung macht auch Aussagen in einer natürlichen Sprache der Verarbeitung durch eine Turingmaschine zugänglich.

Symbolverarbeitungsansatz versus künstliche neuronale Netze

Von hier aus ist es nicht weit zu der Annahme, dass Symbolverarbeitung eine notwendige und hinreichende Bedingung für intelligentes Verhalten ist. Diese Behauptung wurde 1975 von dem Kognitionspsychologen Allen Newell (1927–1992) und dem Sozialwissenschaftler Herbert A. Simon (1916–2001) aufgestellt.[12] Damit ist einerseits gemeint, dass physikalische Symbolverarbeitungssysteme zu intelligentem Verhalten fähig sind. Andererseits wird impliziert, dass auch menschliche Intelligenz sich auf physikalische Symbolverarbeitungsprozesse zurückführen lässt. Das menschliche Gehirn wäre demnach letztlich nichts anderes als eine »Symbolverarbeitungsmaschine aus Fleisch«[13]. Man spricht in diesem Zusammenhang des-

halb auch vom computationalen Modell des Geistes. Newells und Simons These hatte großen Einfluss auf die Entwicklung der KI. Sie bringt ein Forschungsprogramm auf den Punkt, das auch als *Good Old-Fashioned AI* oder *GOFAI* bezeichnet wird.[14]

Das legt nahe, dass es weitere Formen der KI gibt, die von anderen Annahmen ausgehen. Entsprechend der einen Auffassung ist es ein Fehler, gänzlich von der Beschaffenheit des menschlichen Gehirns zu abstrahieren, wenn wir Intelligenz verstehen und nachbilden möchten. Dieser Ansatz versucht, ein Analogon der Neuronenverbände, die das menschliche Gehirn ausmachen, mit den Mitteln der Informatik zu konstruieren. Da es besonders auf die Verbindung zwischen den einzelnen Neuronen ankommt, spricht man auch von *Konnektionismus*.[15]

Der Aspekt der Intelligenz, den dieser Ansatz in den Vordergrund stellt, ist die Fähigkeit zu lernen. Der Konnektionismus wird sogar häufig mit dem maschinellen Lernen gleichgesetzt. Künstliche neuronale Netze haben nämlich die Fähigkeit, in großen Datenmengen Muster zu erkennen. Auch wenn die Bezeichnung altmodisch für den Symbolverarbeitungsansatz nahelegt, dass der Konnektionismus eine neuere Entwicklung ist, reichen die ersten Versuche, künstliche neuronale Netze zu entwickeln, schon bis in die Pionierphase der KI Mitte des 20. Jahrhunderts zurück.[16] Allerdings schienen die Erfolge zunächst gering zu sein, so dass der Ansatz erst in den 1980er Jahren an Popularität gewann.

Es wäre allerdings ein Missverständnis, davon auszugehen, dass künstliche neuronale Netze Gehirnstrukturen eins zu eins abbilden. Vielmehr handelt es sich um mathematische Modelle für Computerprogramme, die bestimmten Organisationsprinzipien biologischer neuronaler Netze nacheifern. Die computationalen Neurowissenschaften streben zwar an, die Funktionsweise des Gehirns mit Hilfe künstlicher neuronaler

Netze zu verstehen und auf dem Computer zu simulieren. Doch für viele KI-Anwendungen kommt es nicht auf die biologischen Details an, sondern darauf, praktikable Lösungen für bestimmte Probleme zu finden.

Biologisch inspirierte Grundeinheit des Konnektionismus ist das Neuron. Neuronen sind über Synapsen miteinander verbunden, über die sie Signale senden oder empfangen. In biologischen Netzen handelt es sich um chemische Reaktionen, bei künstlichen neuronalen Netzen werden diese durch elektrische Schaltkreise simuliert. Wenn diese Signale eine gewisse Schwelle überschreiten, feuert das Neuron und überträgt seine Aktivität auf diese Art und Weise auf ein anderes Neuron. Während die Struktur des biologischen Gehirns noch recht unübersichtlich ist, liegt künstlichen neuronalen Netzen ein klarer hierarchischer Aufbau zugrunde. Die Neuronen sind schichtweise angeordnet und verfügen über Verbindungen zu den Neuronen der nächsthöheren und der nächstniedrigeren Schicht.

Die Neuronen auf der untersten Ebene bekommen Input von außerhalb des Netzes, beispielsweise Bildpixel von einer Kamera. Die darauf folgenden Schichten werden »versteckte Schichten« (*Hidden Layers*) genannt: Die Neuronen dieser Schichten erhalten jeweils Input von den Neuronen der nächstniedrigeren Schicht und geben Signale an die nächsthöhere Schicht weiter. Die höchste Schicht stellt den Output des Netzes bereit, beispielsweise die Ausgabe, dass es sich bei einem bestimmten Bild um eine Katze handelt. Künstliche neuronale Netze, die nur aus Ein- und Ausgabeschicht bestehen, sind zwar prinzipiell möglich, aber die versteckten Schichten erhöhen die Datenverarbeitungsmöglichkeiten eines Netzes erheblich.

Ein entscheidender Punkt ist, dass die Verbindungen zwischen den Neuronen unterschiedlich gewichtet werden kön-

nen. Wiederum stand eine biologische Vorstellung Pate. Der Psychologe Donald O. Hebb (1904–1985) formulierte 1949 die nach ihm benannte Hebb'sche Regel.[17] Sie besagt, dass die Verbindungen zwischen Neuronen durch gleichzeitiges Feuern verstärkt und ohne gleichzeitige Aktivität abgeschwächt werden. Je stärker die Verbindung zwischen zwei Neuronen ist, desto weniger Erregung bedarf es, damit ein Signal des einen Neurons das andere zum Feuern bringt. Die Verstärkung der neuronalen Verbindung bildet für Hebb die neuronale Grundlage von Lernprozessen und wird in abgewandelter Form häufig in künstlichen neuronalen Netzen verwendet. Durch Verstärkung und Abschwächung werden einige Neuronen aktiviert, andere nicht. Es entsteht ein spezifisches Aktivitätsmuster. Auf diese Art und Weise kann ein neuronales Netz beispielsweise darauf trainiert werden, Bilder von Katzen anhand spezifischer Muster zu erkennen, die diesen Bildern zugrunde liegen.

Dieser Lernprozess kann überwacht sein, d. h. dem Netz werden Bilder mit Katzen und solche ohne Katzen präsentiert. Dabei wird jeweils angegeben, auf welchem Bild eine Katze zu sehen ist. Es ist aber auch möglich, dem Netz nur Katzenbilder vorzulegen und keine weiteren Einschränkungen zu machen. In diesem Fall lernt das Netz nicht überwacht, und es bleibt ihm selbst überlassen, die relevanten Ähnlichkeiten herauszufinden. Die Gesichter von Katzen, ihre Körperform, Schwanz, Pfoten oder Schnurrhaare stellen sich zwar in ganz verschiedenen Posen, Farben und Blickwinkeln dar, dennoch bilden sie Muster, die ein neuronales Netz erkennen kann, wenn es mit sehr vielen Katzenbildern konfrontiert wird.

Ob menschliches Lernen ähnlich funktioniert, ist fraglich, schon weil wir nicht so großer Datenmengen bedürfen, um etwas zu lernen. Uns scheint es beim Lernen auf grundlegendere Zusammenhänge von Eigenschaften in der Welt anzukommen und nicht nur auf Ähnlichkeitsmuster. Gleichwohl ist das

Verfahren sehr erfolgreich und kann anhand ganz verschiedener Daten durchgeführt werden, z. B. anhand von Dokumenten, Bildern, Videos, Röntgenaufnahmen, Facebook Likes, Sprachaufnahmen oder Börsentransaktionen. Ist von *Big Data* die Rede, geht es zumeist um diese Form der KI.

Obwohl künstliche neuronale Netze möglicherweise anders lernen als Menschen, ist es dennoch aufgrund ihrer hohen kognitiven Leistungsfähigkeit (sie übertreffen die menschlichen Mustererkennungsfähigkeiten um ein Vielfaches) gerechtfertigt, von künstlicher Intelligenz zu sprechen. Das ist ein weiterer Grund für die Annahme, dass künstliche Intelligenz nicht immer menschenähnlich funktionieren muss.

Noch eine wichtige Beobachtung ist, dass die Kluft zwischen dem Symbolverarbeitungsansatz und dem Konnektionismus in praktischer Hinsicht gar nicht so tief ist, wie sie zunächst zu sein scheint. Für einige Anwendungen ist GOFAI die bessere Wahl, für andere sind es künstliche neuronale Netze. Der Symbolverarbeitungsansatz hat seine Stärke bei Problemen, die abstraktes Schließen erfordern, während künstliche neuronale Netze besser sind, wenn es um Musterkennung auf der Grundlage verrauschter Daten geht, die verzerrt, instabil oder nicht eindeutig sind. Neuronale Netze kommen hingegen an ihre Grenzen, wenn es keine oder nur wenige Daten gibt, um ein Problem zu lösen, oder wenn es auf Fehlerfreiheit ankommt.

Ebenso wie am Symbolverarbeitungsansatz wurde jedoch auch am Konnektionismus grundsätzliche Kritik geäußert. Ein einflussreicher Einwand lautet, dass künstliche neuronale Netze einige zentrale Charakteristika von Intelligenz nicht modellieren können, insbesondere die Produktivität und Systematizität, die kognitive Leistungen ausmachen.[18]

Nehmen wir als Beispiel die Sprache: Wir können potentiell unendlich viele sprachliche Äußerungen verstehen. Als endli-

che Wesen ist es für uns unmöglich, dass wir jeden einzelnen Satz, den wir äußern oder verstehen können, irgendwann einmal auswendig gelernt haben. Der Schlüssel zu unserem Sprachverständnis liegt darin, dass wir auf der Grundlage eines endlichen Vokabulars und einer endlichen Menge grammatischer Regeln unendlich viele neue sprachliche Äußerungen erzeugen und verstehen können. Denn komplexe sprachliche Äußerungen lassen sich auf diese Weise systematisch auf einfachere Einheiten zurückführen. Diese Eigenschaften sind ein Markenzeichen des Symbolverarbeitungsansatzes, für den Konnektionismus sind sie hingegen schwieriger einzufangen – zumindest sofern er sich rigoros vom Symbolverarbeitungsparadigma absetzt.

Ob eine solche Entgegensetzung zutrifft, ist jedoch fraglich, denn künstliche neuronale Netze sind zwar anders aufgebaut als eine Turingmaschine, im Hinblick auf die Berechenbarkeit aber äquivalent. Es könnte sich auch einfach um unterschiedliche Beschreibungsebenen handeln. Möglicherweise realisieren konnektionistische Netze auf subsymbolischer Ebene kognitive Prozesse, die sich auch symbolisch beschreiben lassen. So könnte die Fähigkeit, die ein künstliches neuronales Netz erwirbt, das auf Spracherkennung trainiert wird, dadurch beschrieben werden, dass es eben die linguistischen Grundeinheiten und Regeln ihrer Verknüpfung erlernt. Beide Ansätze werden in der Praxis auch miteinander kombiniert. *Google DeepMind* arbeitet mit der *Neural Turing Machine (NMT)* an einem solchen System. Die Vision besteht darin, auf diese Art und Weise einen Computer zu schaffen, der selbst Programme für ihm unbekannte Situationen entwickeln kann.

Embodiment und verhaltensbasierte KI

Künstliche neuronale Netze bilden aber nicht die einzige Alternative zum klassischen Symbolverarbeitungsparadigma. In den 1980er Jahren kam ein neuer Ansatz ins Spiel, der unter den Begriffen der *Nouvelle AI*, *Embodied AI* oder *verhaltensbasierte KI* bekannt wurde. Während der Konnektionismus den Fehler des klassischen Paradigmas darin sah, Intelligenz gänzlich unabhängig von der Beschaffenheit des menschlichen Gehirns zu betrachten, fordert die verhaltensbasierte KI, auch den Körper und seine Umwelt mit einzubeziehen.[19] Die beiden zentralen Schlagworte, die diesen Ansatz charakterisieren, sind daher *Verkörperung* und *Situiertheit* der Kognition.[20]

Im Vordergrund stehen nun nicht mehr anspruchsvolle geistige Leistungen, sondern sensomotorische Fähigkeiten, über die auch verhältnismäßig einfache Organismen wie Insekten verfügen. Dieser Ansatz bleibt folgerichtig nicht bei Computermodellen stehen, sondern führt zum Bau von Robotern, die über einen Körper, Sensoren und Aktoren verfügen. Ziel ist es, Systeme zu schaffen, die die Fähigkeit haben, ihre Umwelt wahrzunehmen, sich in ihr zu bewegen und Objekte zu manipulieren, so dass sie sich erhalten und unter Umständen sogar reproduzieren können. Wie Rodney Brooks, einer der Pioniere dieser Richtung, betont, muss man von Anfang an mit Robotern arbeiten, die sich in der wirklichen Welt und nicht nur in einer künstlich vereinfachten Modellwelt behaupten können.[21]

Dieser Ansatz kommt einer wirkungsmächtigen Grundsatzkritik entgegen, die von Hubert Dreyfus (1929–2017) schon früh gegen das Projekt der künstlichen Intelligenz geltend gemacht wurde.[22] Ausgehend von Martin Heideggers (1889–1976) Begriff des *In-der-Welt-seins* kritisiert er die KI (insbesondere den Symbolverarbeitungsansatz), weil er der Einbettung der menschlichen Intelligenz in eine Welt nicht gerecht