9 Großartige künstliche Intelligenz erschaffen

Danko Nikolić

"We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."
John McCarthy, Marvin Minsky, Nathaniel Rochester und Claude Shannon im Jahr 1955

Fragen, die in diesem Kapitel beantwortet werden:

- Was ist KI, und wie unterscheidet sie sich von der einfachen Erstellung von Modellen für maschinelles Lernen?
- Was braucht es, um ein großartiges KI-Produkt zu entwickeln?
- Was sind die häufigsten Fallen bei der Konzeption und Entwicklung einer KI, und wie können Sie diese Fallen vermeiden?

9.1 Wie KI mit Data Science und maschinellem Lernen zusammenhängt

Sie fragen sich vielleicht, welche Funktion ein Kapitel über künstliche Intelligenz (KI) in einem Buch über Data Science hat? Oft wird KI nur als ein schicker Name für maschinelle Lernmodelle verstanden, Modelle, die Data Scientists ohnehin im Rahmen ihrer Arbeit erstellen. Wenn das der Fall wäre, wäre KI einfach ein Teil der Data Science, und es gäbe keinen Grund, ein separates Kapitel über KI zu schreiben, da sich der Rest des Buches um diese Technologie dreht. Nun, das ist nicht ganz richtig. Es stimmt zwar, dass einer der wichtigsten – und vielleicht auch interessantesten – Teile der KI in den maschinellen Lernmodellen liegt, aber KI besteht aus viel mehr als nur maschinellem Lernen. Bei der Entwicklung eines KI-Produkts sind einige kritische Überlegungen anzustellen, die Sie normalerweise nicht in einem typischen Buch über Data Science oder sogar in anderen Kapiteln dieses Buchs finden. Wenn Sie in einem dieser Bereiche einen Fehler machen, kann Ihr Endprodukt enttäuschen. Sie können beispielsweise in eine Situation geraten, in der zu Beginn des Erstellungsprozesses alles in Ordnung zu sein scheint, das Endprodukt jedoch nicht überzeugt und die Bedürfnisse und Erwartungen der Endnutzer nicht erfüllt.

Sehen wir uns zunächst an, welche Art von Maschinen wir heute als Beispiele für KI betrachten. Was uns sofort in den Sinn kommen mag, ist vielleicht ein Roboter. Aber nicht irgendein Roboter. Die meisten Roboter sind nicht sehr intelligent. Roboter bestehen aus mechanischen Komponenten wie Armen und Aktuatoren. Und dann gibt es noch Batterien und Sensoren. Aber das allein reicht nicht aus, um einen Roboter als KI zu bezeichnen. Es gibt viele Roboter, die zwar sehr nützlich für uns sind, aber einfach nur dumm sind. Beispiele dafür sind Staubsaugerroboter in Privathaushalten und Industrieroboter in Fabrikhallen. Ausschlaggebend dafür, ob ein Roboter das Prädikat „künstlich intelligent" erhält oder nicht, ist, was er mit seiner gesamten Hardware selbstständig tun kann. Nur ein intelligenter Roboter, der über Fähigkeiten verfügt, die weit über die reine Programmierung von Bewegungen hinausgehen, wird die Ehre haben, als KI bezeichnet zu werden. Wir suchen hier nach einem Roboter, der eine Vielzahl unterschiedlicher Verhaltensweisen zeigt, sich in einer komplexen Umgebung zurechtfindet oder Aufgaben in einer Vielzahl neuartiger Situationen bewältigen kann. Stellen Sie sich zum Beispiel einen anthropomorphen Roboter vor, der in der Lage ist, einen Tisch voller schmutzigem Geschirr abzuräumen, dieses Geschirr dann manuell abzuwaschen und es schließlich abzutrocknen und in den Schrank zu stellen – und das alles, ohne etwas kaputt zu machen! Roboter mit derartigen Fähigkeiten gibt es noch nicht.

Um mit der Entwicklung eines solchen Roboters zu beginnen, könnte es bald klar sein, dass es nicht ausreicht, Deep-Learning-Modelle zu trainieren. Man kann sich dafür entscheiden, sich in hohem Maße auf Deep Learning zu verlassen, und doch wird der Roboter viel mehr brauchen als das, was Deep Learning bieten kann. Um die erforderliche Intelligenz des Roboters zu fördern, müssen wir Technologien entwickeln und einsetzen, die weit über die Möglichkeiten des maschinellen Lernens hinausgehen – und auch weit über das hinausgehen, was Data Science abdeckt. Dennoch werden Data Scientists eine entscheidende Rolle bei der Entwicklung solcher Roboter spielen. Das ist der Grund, warum Sie dieses Kapitel lesen.

Eine Art von Robotern hat die Aufmerksamkeit der Industrie auf sich gezogen, und es wurde auch viel investiert: unsere Autos. Es wurde viel Geld in die Entwicklung von Autos gesteckt, die in der Lage sind, selbstständig zu fahren und somit zu intelligenten Robotern zu werden. Das Problem des autonomen Fahrens ist nicht einfach, vor allem dann nicht, wenn das Fahrzeug in der „echten Welt" und nicht in einer kontrollierten Testumgebung fährt. Die Vielfalt der unterschiedlichen Situationen, denen das Fahrzeug begegnen kann, ist enorm. Daher stellen solche Fahrzeuge eine große Herausforderung für die Technologie dar. Vielleicht ist das Problem des autonomen Fahrens so schwierig wie das Abräumen eines Tisches mit Geschirr. Der Druck auf die Qualität der Lösung, d. h., keinen Fehler zu machen, ist ebenfalls hoch. Während unser manueller Geschirrspülroboter im schlimmsten Fall ein paar Gläser oder Teller kaputt macht, trägt ein Autoroboter eine viel größere Verantwortung: Er ist für Menschenleben verantwortlich. Dies ist ein weiterer Grund, der ein erfolgreiches selbstfahrendes Auto zu einem schwierigen Ziel macht. Dennoch hat es in diesem Bereich einige Fortschritte gegeben. Autonome Fahrzeuge sind wohl die schlausten und intelligentesten Roboter, die die Menschheit bisher gebaut hat. Und doch gibt es noch viel zu tun. Die Frage ist also: Was war nötig, um diese Maschinen intelligent zu machen? Und vor welchen Problemen und Hürden stehen diese Maschinen in Bezug auf Intelligenz noch? Ist das alles nur Data Science, um größere und intelligentere Modelle zu bauen, oder steckt da mehr dahinter?

Um diese Fragen zu beantworten, müssen wir zunächst feststellen, dass KI nicht gleichbedeutend mit einem maschinellen Lernmodell ist. Um das zu verstehen, hilft es, zwischen einem Produkt und einer kritischen Komponente zu unterscheiden, die für den Aufbau eines Produkts erforderlich ist. Ein Produkt ist viel mehr als nur seine kritischen Komponenten. Ein Messer ist mehr als nur eine Klinge, auch wenn die Klinge der entscheidende Bestandteil ist. Ein Monitor ist mehr als seine kritische Komponente, der Bildschirm. Ein Speicherstick ist mehr als ein SSD-Chip. Ein Fahrrad ist mehr als nur ein Paar Räder und Pedale. In all diesen Fällen stellen wir fest, dass ein Produkt mehr ist als seine entscheidenden Komponenten.

Wir können diesen Unterschied am Beispiel eines Autos erkennen. Ein Auto, das sich auf dem Markt verkaufen lässt und somit geeignet ist, einen Wert für den Kunden zu schaffen, ist mehr als ein Motor auf vier Rädern. Für ein Auto braucht man ein Lenkrad und Bremsen. Doch das ist noch kein vollständiges Produkt. Zu einem vollständigen Produkt gehören auch Scheinwerfer für Nachtfahrten, eine Windschutzscheibe, Türen, Fenster an den Türen und Scheibenwischer an der Windschutzscheibe. Man braucht auch eine vollständige Kabine mit Sitzen. Dann braucht man eine Heizung, eine Klimaanlage und ein Unterhaltungssystem. All dies muss in ein schönes Design verpackt werden, das dem menschlichen Auge gefällt. Erst wenn wir all dies zusammengefügt haben, haben wir ein vollständiges Produkt, das wir Auto nennen.

Eine KI ist wie ein vollständiges Produkt. Sie ist eine Maschine, die einen Dienst für einen Menschen leistet, und damit dieser Dienst zufriedenstellend erledigt werden kann, muss die Maschine vollständig sein. Man muss ein vollständiges Produkt schaffen. Ein maschinelles Lernmodell kann also eine entscheidende Komponente für eine KI sein, vielleicht das Äquivalent zu einem Motor für ein Auto. Wichtig ist jedoch, dass wir erst dann eine KI haben, wenn wir ein Produkt um diesen (maschinellen Lern-)Motor herum gebaut haben.

In der Praxis erfordert die Erstellung eines Produkts zumindest, dass das Modell in Produktion geht und eine Schnittstelle für die Erfassung der Eingaben, die in das Modell für maschinelles Lernen einfließen, eingerichtet wird und dann auch eine Form von Ausgabe erzeugt wird. Oft ist viel mehr erforderlich, um ein nützliches Produkt zu schaffen. Wie wir im Fall des autonomen Fahrzeugs gesehen haben, ist eine Menge Hardware erforderlich, um ein komplettes Auto zu bauen.

Aber es sind nicht nur „nicht intelligente“ Komponenten, die man zu maschinellen Lernmodellen hinzufügen muss, um eine KI zu schaffen. Ein tieferer Grund, warum maschinelles Lernen allein für KI nicht ausreicht, ist, dass KI-Lösungen oft sehr viel komplexer sind als das, was mit einem einzelnen maschinellen Lernmodell erreicht werden könnte. Betrachten wir zum Beispiel einen Chatbot. Nehmen wir an, dass wir außer der intelligenten Komponente nur eine minimale Schnittstelle schaffen müssen, die aus Textfeldern besteht, in die die Fragen der Benutzer eingegeben und die Antworten der Maschine ausgedruckt werden. Daraus könnte man schließen, dass es ausreichen sollte, zwischen diesen beiden Komponenten ein großes, gut trainiertes maschinelles Lernmodell zu platzieren, das das Chatten mit einem menschlichen Benutzer übernimmt. Leider funktioniert das so nicht. Jeder ausgeklügelte intelligente Chat-Assistent (man denke an Alexa, Siri, Cortana usw.) ist viel komplexer als ein einzelnes Deep-Learning-Modell.

Nachfolgend ist die Architektur der ursprünglichen Watson AI die Lösung – eine Maschine, die 2010 Geschichte schrieb, als sie das Spiel Jeopardy gegen die besten menschlichen

Spieler in diesem Spiel gewann. Es ist klar, dass die Organisation dieser KI viel aufwendiger war als ein einzelnes maschinelles Lernmodell. Viele ihrer Komponenten beruhen nicht einmal auf maschinellem Lernen, tragen aber dennoch zur Gesamtintelligenz von Watson bei. Es ist wichtig zu verstehen, dass nur die Maschine als Ganzes eine KI ist; keine einzelne Komponente allein ist eine. Ein großer Teil dieser Gesamtintelligenz kommt von der Architektur – wie der Rechenfluss organisiert ist und wie entschieden wird, welche Komponente wann ausgeführt wird. Es sind also nicht nur die Gewichte in den maschinellen Lernmodellen, die zur Gesamtintelligenz beitragen. Es gibt noch viel mehr, einschließlich der Regeln, nach denen die verschiedenen Modelle miteinander interagieren und sich gegenseitig helfen. Erst die vollständige Kombination aller Teile, also Watson, ist ein vollständiges Produkt und eine KI.

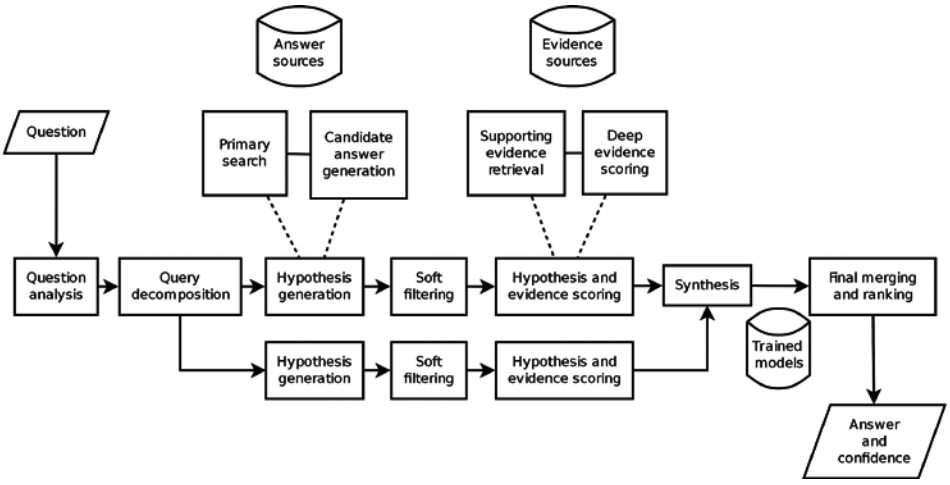

Bild 9.1 Die Architektur der ursprünglichen Watson-KI, die das Spiel Jeopardy gegen die besten menschlichen Konkurrenten gewann *(https://en.wikipedia.org/wiki/File:DeepQA.svg)*

Ähnliches gilt für die Intelligenz von autonomen Fahrzeugen. Die internen Architekturen der Algorithmen, die die Autos steuern, sind nicht einfacher als die von Watson. Außerdem werden die Autos mit der Zeit immer intelligenter und bessere Fahrer, sodass die Zahl der Komponenten und die interne Komplexität der KI-Gesamtlösungen tendenziell zunehmen.

Wichtig ist, dass viele der Komponenten solcher Lösungen auch keine Modelle des maschinellen Lernens sind, sondern andere Algorithmen verwenden. Diese anderen Komponenten können die Suche in Datenbanken, Brute-force-Ansätze zum Finden optimaler Lösungen, rein wissenschaftliche Berechnungen, regelbasierte Entscheidungsfindung und so weiter sein. Auch hier tragen alle diese Komponenten gemeinsam zur Gesamtintelligenz der KI bei.

Schließlich gibt es einen weiteren Grund, warum maschinelles Lernen und KI nicht dasselbe sind. Maschinelles Lernen wird oft für andere Zwecke als die Entwicklung intelligenter Maschinen eingesetzt. Maschinelles Lernen hat Verwendungszwecke, die über das hinausgehen, wofür KI gedacht ist. Insbesondere wird maschinelles Lernen oft als Werkzeug für die Datenanalyse eingesetzt. Der Autor dieses Kapitels hat maschinelles Lernen ausgiebig genutzt, um zu analysieren, wie das Gehirn sensorische Informationen speichert. Wir

haben maschinelle Lernmodelle trainiert, um Informationen aus Gehirnsignalen zu lesen. Entscheidend ist, dass es uns nicht darum ging, ein Produkt zu entwickeln. Vielmehr stellten wir Fragen über das Gehirn, z. B. wie lange speichert das Gehirn Informationen über ein Bild, das wir kurz auf dem Bildschirm gezeigt haben? Oder: Wie schnell kann diese Information durch einen neu dargebotenen Reiz wieder gelöscht werden? Auf diese Weise haben wir zahlreiche Erkenntnisse darüber gewonnen, wie das Gehirn sensorische Informationen aufbewahrt [1-3]. Für reine Engineers mag eine solche Anwendung des maschinellen Lernens überraschend sein. Für einen Data Scientist sollte dies jedoch nicht so unerwartet sein. Kein Wissenschaftler sollte zögern, Algorithmen des maschinellen Lernens als Analysewerkzeuge einzusetzen. Der Einsatz von maschinellem Lernen bringt große Vorteile mit sich, insbesondere in Situationen, in denen die Daten komplex und Erkenntnisse mit herkömmlichen Analysemethoden nur schwer zu gewinnen sind.

Um die Beziehung zwischen maschinellem Lernen und KI zu verstehen, ist es üblich, Venn-Diagramme zu zeichnen, wie sie in Bild 9.2 zu sehen sind. Das Venn-Diagramm auf der linken Seite ist dasjenige, das in der KI-Literatur häufig zu finden ist. Das rechte Diagramm ist jedoch korrekter, da es auch die Tatsache berücksichtigt, dass maschinelles Lernen für andere Zwecke als KI eingesetzt werden kann.

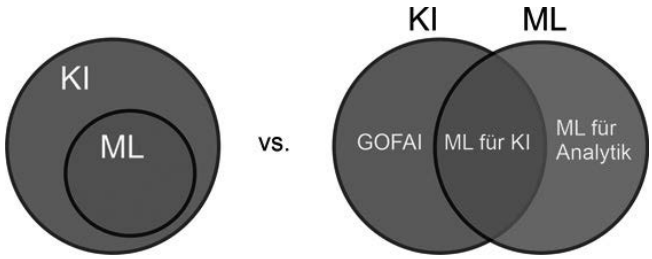

Bild 9.2 Die Beziehung zwischen KI und maschinellem Lernen (ML). Links: die Beziehung, wie sie in der Literatur häufig dargestellt wird. Rechts: eine realistischere Darstellung, die zeigt, dass maschinelles Lernen auch für andere Zwecke als KI verwendet werden kann, z. B. für die Analyse von Daten. GOFAI steht für „Good old-fashioned AI" (gute altmodische KI), die kein maschinelles Lernen einsetzt.

9.2 Eine kurze Geschichte der KI

Um zu verstehen, dass KI sich nicht ausschließlich auf maschinelles Lernen stützen muss, ist es am besten, einen Blick auf ihre Geschichte zu werfen. Die Geschichte der KI ist nämlich länger als die des maschinellen Lernens. Grob gesagt gab es bei der KI zwei große Phasen. In der ersten Phase lag der Schwerpunkt auf der Entwicklung von Algorithmen, die nichts mit maschinellem Lernen zu tun hatten. Stattdessen stützten sich diese frühen Algorithmen ausschließlich auf maschinelles Wissen, das manuell in die Maschine eingegeben wurde – von Menschen. In dieser ersten Phase galt zum Beispiel ein großer regelbasierter Entscheidungsbaum als ein hochmoderner Algorithmus für KI. Kennzeichnend für diese Form der KI ist, dass sie ihr Wissen nicht in Form von Zahlen speicherte und keine Schlussfolgerungen durch Anwendung von Gleichungen auf diese Zahlen zog. Der Grund dafür war

einfach: Es wäre für Menschen sehr schwierig gewesen, zahlenbasiertes Wissen, wie z. B. die Verbindungsgewichte von künstlichen neuronalen Netzen, in eine Maschine einzugeben. Stattdessen wurde das meiste Wissen in einer symbolischen Form gespeichert – einer Form, die für Menschen verständlich ist. Zum Beispiel konnte ein Wissensobjekt Symbole verwenden, um „wenn Fieber, dann Grippe" darzustellen. Die Schlussfolgerungen wurden durch Anwendung logischer Regeln auf diese Symbole gezogen. Die Regeln waren wiederum für Menschen verständlich.

Wir bezeichnen diese Stufe der KI häufig als *symbolische* KI (siehe auch Kapitel 10). Ein anderer allgemein bekannter Begriff ist Good-old-fashioned-AI, abgekürzt GOFAI. Die Forschung im Bereich der symbolischen KI begann bereits in den 1950er-Jahren, wobei der offizielle Geburtsort die historische Dartmouth-Konferenz im Jahr 1956 ist. Der Organisator dieser Konferenz, John McCarthy – der den Begriff „Künstliche Intelligenz" formulierte –, stellte auch die erste Programmiersprache vor, die Computern zu symbolischer Intelligenz verhelfen sollte: LISP. Die aus zwei Buchstaben bestehende Abkürzung „AI" wurde erst nach Steven Spielbergs Film „A.I. Artificial Intelligence" aus dem Jahr 2001 allgemein verwendet.

Das maschinelle Lernen trat erst später in den Bereich der KI ein und leitete die zweite – und immer noch –aktuelle – Phase ein. Tatsächlich hat sich das maschinelle Lernen erst in den 1970er-Jahren durchgesetzt, obwohl einige der Algorithmen schon viel früher existierten. Der Grund für diese Verzögerung war, dass man erst mit der Zeit erkannte, dass die symbolische KI ihre Grenzen hatte und ein anderer Ansatz erforderlich war. Ein Problem bestand darin, dass die symbolische KI nicht von sich aus lernen konnte; das Wissen musste von Menschen vermittelt werden. Dies stellte einen enormen Engpass dar, da die Menge an Wissen, die manuell aufgebaut werden musste, oft zu groß war. Daher wurde der GOFAI-Ansatz zur Steigerung der Intelligenz von Maschinen nicht mehr tragbar. Infolgedessen erreichten viele Projekte nicht die Nützlichkeitsebene und kamen über den anfänglichen Proof-of-Concept nicht hinaus; was in kleinem Maßstab gut funktionierte, konnte in größerem, nützlicherem Maßstab nicht umgesetzt werden.

Heute, in der zweiten Phase, verlassen wir uns überwiegend auf Algorithmen des maschinellen Lernens, um Wissen in Maschinen einzuspeisen und es aus großen Datensätzen in Matrizen von Modellparametern umzuwandeln. Diese Algorithmen stellen eine große Erleichterung für die manuelle Arbeit dar. Alles, was der Mensch tun muss, ist, Beispiele für intelligentes Verhalten zu liefern. Die Maschine ist dann in der Lage, die Regeln zu extrahieren, nach denen dieses Verhalten zustande kommt.

Offensichtlich haben wir auf diese Weise einen großen Fortschritt in unserer Fähigkeit erzielt, die Intelligenz von Maschinen zu steigern. Es ist jedoch falsch anzunehmen, dass sich die Welt von der symbolischen KI entfernt hat und dass GOFAI-Algorithmen der Vergangenheit angehören. Ganz und gar nicht. Der symbolische Ansatz ist immer noch lebendig und gut. Jede komplexe KI-Lösung, die heute entwickelt wird, ist ein Mischmasch aus maschinellem Lernen und GOFAI-Komponenten. Symbolische KI ist ein nicht minder wichtiger Bestandteil. Es ist nur so, dass GOFAI-Komponenten nicht beworben werden, was mehr mit dem aktuellen Hype und den Marketingstrategien zu tun hat als mit den Fakten, wie die Maschinen unter der Haube arbeiten. Symbolische KI ist weit verbreitet. Oft entscheidet GOFAI, welcher Deep-Learning-Algorithmus als Nächstes ausgeführt wird. In anderen Fällen erhält GOFAI die Ergebnisse der maschinellen Lernmodelle, um die nächste Entschei-

dung zu treffen. In anderen Fällen unterstützt das maschinelle Lernen GOFAI bei der Suche nach einer optimalen Lösung. Oft sind die beiden Komponenten ineinander verschachtelt: Ein symbolischer Algorithmus ruft ein maschinelles Lernmodell auf, das wiederum eine andere GOFAI-Komponente um Hilfe bittet, die wiederum zum maschinellen Lernen zurückkehrt und so weiter. Die Möglichkeiten sind grenzenlos. Watson könnte ohne GOFAI-Komponenten kein Jeopardy-Spiel gewinnen. Ohne die Verwendung eines GOFAI hätte alphaGo das Go-Spiel gegen den Weltmeister Lee Sedol nicht gewinnen können (das Ergebnis war vier zu eins für die Maschine). Ein autonomes Fahrzeug beispielsweise kann nicht ohne altmodische KI-Komponenten fahren. Alexa, Siri und Co können sich ohne symbolische Teile ihrer allgemeinen Intelligenzarchitekturen nicht mit Ihnen unterhalten.

Was bedeutet das alles für einen Data Scientist, der heute mit der Entwicklung eines KI-Produkts betraut ist? Sehr wahrscheinlich wird Ihre Lösung viel mehr beinhalten müssen als nur ein Modell für maschinelles Lernen. Es wird eine Menge Technik außerhalb des maschinellen Lernens benötigt. Es wird schwierig sein, symbolische Komponenten zu vermeiden. Das bedeutet, dass Sie kluge architektonische Entscheidungen über die gesamte Lösung treffen müssen, und diese Entscheidungen werden viel mehr als nur maschinelles Lernen umfassen. Um ein effektives Produkt zu schaffen, benötigen Sie möglicherweise sogar Komponenten, die außerhalb der Technik liegen. Ein gutes Design der Schnittstelle für Ihre KI kann für den Erfolg ebenso entscheidend sein wie die Leistung des zugrunde liegenden Modells. Ähnlich wie man einen ergonomischen Griff an einer Klinge anbringen muss, um ein gutes Messer herzustellen, oder bequeme Sitze braucht, um ein gutes Auto zu bauen, muss sich Ihre KI in vielen verschiedenen Dimensionen entwickeln, um ein gutes Produkt zu präsentieren. Die Modelle des maschinellen Lernens sind nur ein Teil des Gesamtergebnisses und somit auch nur ein Teil des gesamten Kundenerlebnisses.

9.3 Fünf Empfehlungen für die Entwicklung einer KI-Lösung

Auf dem Weg zur Entwicklung einer KI-Lösung muss ein Data Scientist eine Reihe von Entscheidungen treffen. Sie als Data Scientist müssen zwangsläufig eine Architektur erstellen, die Komponenten verschiedener Typen kombiniert, die interagieren und gemeinsam die Intelligenz Ihrer Maschine hervorbringen. Vielleicht werden Sie diese Architektur mit mehreren Kästchen und Pfeilen zeichnen, wie die Zeichnung der Watson-Architektur in Bild 9.1. Die Frage ist dann: Welche Strategien können Sie anwenden, und worauf sollten Sie achten, um bestimmte häufige Fehler zu vermeiden?

9.3.1 Empfehlung Nr. 1: Seien Sie pragmatisch

In den vorangegangenen Kapiteln dieses Buches haben Sie verschiedene Rezepte zur Lösung von Data-Science-Problemen kennengelernt. All dies wird Ihnen als Einzelteile präsentiert, zum Beispiel als einzelne Algorithmen für maschinelles Lernen. Außerdem wer-

den die Teile in einer idealisierten Welt gezeigt, unabhängig vom wirklichen Leben. Wenn Sie eine echte künstliche Intelligenz – ein komplettes Produkt – entwerfen, müssen Sie darüber nachdenken, wie Sie Algorithmen für eine unvollkommene Welt auswählen. Sie müssen darüber nachdenken, wie Sie sie kombinieren können. Außerdem werden Sie Algorithmen finden und verwenden müssen, die nicht in diesem Buch beschrieben sind. Es ist wichtig, dass Sie nicht bei einem Satz von Algorithmen bleiben, nur weil sie in der Vergangenheit für Sie funktioniert haben oder weil Sie sie kennen. Erweitern Sie Ihr Wissen, wenn Sie es brauchen. Wählen Sie die Algorithmen nach ihrer Eignung für ein bestimmtes Problem aus, nicht nach ihrer Bequemlichkeit. Denken Sie daran, dass Ihr neues Problem immer etwas anders sein wird als alles, was Sie bisher gesehen haben. Seien Sie eklektisch bei der Auswahl des Werkzeugs zur Lösung der Aufgaben. Wählen Sie aus einer möglichst großen Auswahl. Schränken Sie sich nicht ein.

Bleiben Sie außerdem pragmatisch. Ihre erste Sorge sollte das Erreichen des Ziels sein. Sie müssen nicht immer die neuesten Algorithmen, das heißeste und am meisten gehypte Werkzeug verwenden. Nehmen Sie stattdessen das, was für das jeweilige Problem am besten geeignet ist. Ich habe schon erlebt, dass sich Data Scientists in bestimmte Modelle „verliebt“ haben und dann Favoriten spielen. Aber Erfolg in Data Science stellt sich nicht ein, wenn man Favoriten spielt. Ich habe Leute erlebt, die versuchen, jedes Problem mit demselben Ansatz zu lösen. Es gibt Leute, die erwarten, dass alles mit Deep Learning gelöst werden muss. Ich habe auch schon eingefleischte Fans von bayesschen Ansätzen gesehen. Sicher, sowohl die bayesschen als auch die Deep-Learning-Methoden sind charmant und haben einige attraktive Eigenschaften, die ihnen einzigartige „Superkräfte“ verleihen. Aber beide haben auch Nachteile. Tatsächlich hat jeder Ansatz, für den Sie sich entscheiden, einige Vorteile gegenüber anderen und zwangsläufig auch einige Nachteile. Ihre Aufgabe ist es, beide Seiten zu berücksichtigen und die Vor- und Nachteile abzuwägen, um eine gute Wahl zu treffen.

Es ist von größter Wichtigkeit, sich sowohl der Vorteile als auch der Nachteile einer bestimmten Methode oder eines Algorithmus bewusst zu sein. Die Nachteile sind vielleicht schwieriger zu erkennen, weil die Autoren, die über ihre neuen Methoden schreiben, dazu neigen, sich auf die positiven Aspekte zu konzentrieren. Die rosigen Bilder sind es, die sie dazu motivieren, überhaupt zu forschen und Artikel zu schreiben. Wir sollten also ein gewisses Verständnis aufbringen. Dennoch muss man sich die Fähigkeit aneignen, „zwischen den Zeilen zu lesen“ und mögliche Einschränkungen und Fallstricke zu erkennen. Ein erfahrener Data Scientist wird in der Lage sein, mögliche Nachteile einer neuen Methode zu erkennen, auch wenn sie nicht so klar formuliert sind wie die Vorteile. Entwickeln Sie diese Fähigkeit, denn sie wird Ihnen viel Kraft geben, um gute Designentscheidungen für Ihre KI-Architekturen zu treffen. Ziel ist es, sich Wissen über eine Vielzahl von Algorithmen, Modellen und Optimierungstechniken anzueignen.

Das Sortiment an Tools, aus dem man wählen kann, ist riesig. Eine einzelne Person kann wahrscheinlich nie einen vollständigen Überblick über den Bereich der Data Science haben. Um umfassende Kenntnisse über Methoden des maschinellen Lernens und KI-Algorithmen zu erlangen, ist lebenslanges Lernen erforderlich. Und man ist nie fertig. Außerdem nimmt das Tempo, mit dem neue Algorithmen vorgeschlagen werden, rapide zu, da immer mehr Menschen an diesem Thema arbeiten, Universitäten neue Abteilungen für KI und Data Science eröffnen und Regierungen mehr Geld für die KI-Forschung bereitstellen. Mit all

diesen Entwicklungen Schritt zu halten ist eine Herausforderung. Man sollte nie aufhören zu lernen, aber auch nicht erwarten, alles zu wissen.

Um sich in diesem ständig wachsenden Wald neuer Werke zurechtzufinden, ist ein gründliches Verständnis von Algorithmen Voraussetzung. Sie werden einen neuen Algorithmus besser verstehen, wenn Sie bereits ein tiefes Verständnis für einen verwandten, bereits existierenden Algorithmus haben. Ein oberflächliches Verständnis von Methoden ist nicht annähernd so leistungsfähig. Das richtige Verständnis mehrerer verschiedener Algorithmen, die jeweils zu einer anderen Kategorie gehören, ist wahrscheinlich die beste Strategie, die man verfolgen kann, um das Gebiet der Data Science zu beherrschen. Neue Algorithmen sind oft mit den bestehenden verwandt. Es kommt selten vor, dass Forscher einen völlig neuen Ansatz zur Lösung eines Problems des maschinellen Lernens entwickeln (obwohl sie gelegentlich genau das tun). Wenn Sie einen Algorithmus gut verstehen, fällt es Ihnen leicht, das Wesen seiner Vettern zu erfassen – sie werden zu einer Variation des Themas. Wenn Sie dagegen einen Algorithmus nur oberflächlich verstehen, kann eine Variation dieses Algorithmus für Sie ein Rätsel sein, und Sie haben möglicherweise Schwierigkeiten zu entscheiden, ob diese neue Variation für Ihr neues Problem hilfreich ist oder nicht.

Man kann den Algorithmus immer an den Daten ausprobieren und sehen, was passiert. Es gibt auch Tools, mit denen man automatisch mehrere Algorithmen ausprobieren und den besten auswählen kann (als autoML bezeichnet). Aber damit kommt man nicht weit. Man kann ein autonomes Fahrzeug nicht entwickeln, indem man wahllos verschiedene Architekturen ausprobiert. Wenn man KI entwickelt, muss man das gute alte menschliche Denken einsetzen – und zwar eine ganze Menge. In diesem Fall wollen Sie die Entscheidungsfindung minimieren, indem Sie die Algorithmen an Ihren Daten ausprobieren. Natürlich werden Sie das irgendwann tun müssen, daran besteht kein Zweifel. Der Unterschied zwischen einem erfahrenen KI-Entwickler und einem unerfahrenen Entwickler besteht jedoch darin, dass ersterer die Aufgabe mit mehr Nachdenken und weniger Ausprobieren bewältigen kann. Erfahrene Menschen können die Möglichkeiten im Kopf durchgehen, ohne den Algorithmus auf den Daten trainieren zu müssen. Dank des erweiterten Wissens können sie erkennen, dass etwas nicht gut funktionieren wird, noch bevor sie es ausprobieren. Das spart eine Menge Zeit.

Was kann Ihnen noch helfen, gute Entscheidungen zu treffen? Eine gute Idee ist es, Ihre zukünftige Architektur zu zeichnen, bevor Sie mit der Programmierung beginnen. Legen Sie die Details fest und versuchen Sie, den Datenfluss im System mental zu simulieren. Stellen Sie sich bei jedem Schritt eine Frage: Sehe ich einen Grund, warum dieser Schritt scheitern oder Schwierigkeiten bereiten könnte? Wenn Sie mögliche Probleme sehen, gehen Sie diese Probleme sofort an. Pragmatisch ist es, die schwächsten Punkte zuerst anzugehen. Hoffen Sie nicht, dass ein Wunder geschieht, nachdem Sie sich mit dem einfachen Teil beschäftigt haben.

Es ist eine weit verbreitete Ansicht, dass mit genügend Rechenleistung und einer ausreichenden Menge an Daten alles möglich ist: dass alles von einer Maschine gelernt werden kann. Obwohl an dieser Aussage etwas Wahres dran ist, gibt es auch eine Menge Unwahrheiten. Auf einige dieser Punkte werde ich später in diesem Kapitel eingehen. Die Quintessenz ist, dass die blinde Verfolgung der Strategie „mehr Daten mit mehr Rechenleistung“ fast garantiert zu Problemen führen wird. Es ist viel besser, Ihre Algorithmen gründlich zu bereinigen, indem Sie Ihr Verständnis von Statistik, maschinellem Lernen und KI im Allgemeinen nutzen. Vertrauen Sie auf Big Data und Rechenleistung als Ihre letzte Ressource.

Sicherlich müssen Sie verschiedene Konzepte ausprobieren. Und Sie werden die Ergebnisse dieser Versuche als Feedback nutzen müssen. Sie werden Ihnen zeigen, wie Sie sich verbessern können. Es ist wichtig zu erkennen, dass Ihre Iterationen viel schneller und effektiver sein werden, wenn Sie besser verstehen, was Sie tun.

Das Denken ist vergleichsweise schwer. Das Codieren und Ausführen von Modellen ist vergleichsweise einfach. Wenn Sie sich jedoch nicht davor scheuen, den schwierigen Teil zu tun, werden Sie wahrscheinlich den Wettbewerbsvorteil erlangen, den Sie brauchen, um ein Produkt zu schaffen, das der Markt braucht und das ihm gefällt.

Vergessen Sie nicht, dass eine Person nicht alles wissen kann. Stellen Sie ein Team aus Personen mit unterschiedlichen Fachkenntnissen zusammen. Lassen Sie alle mitarbeiten; jeder sollte ein Mitspracherecht haben. Stellen Sie sicher, dass Sie das Talent jedes Einzelnen für Ihr Endprodukt nutzen können.

9.3.2 Empfehlung Nr. 2: Erleichtern Sie Maschinen das Lernen – schaffen Sie induktive Verzerrungen

Es gibt eine einfache Wahrheit über Algorithmen für maschinelles Lernen: Einige lernen schneller und besser als andere. In manchen Fällen genügen schon wenige Beispiele, um eine hohe Leistung zu erreichen. In anderen Fällen sind Millionen von Beispielen erforderlich. Es gibt zwar viele Gründe für diese Unterschiede, aber es gibt einen Grund, den Sie selbst beeinflussen können: Ein Faktor, der die Lerneffizienz eines Algorithmus bestimmt, sind seine induktiven Verzerrungen. Induktive Verzerrungen sind wie ein Stück Wissen, das einem Algorithmus hinzugefügt wird und es ihm ermöglicht, einige Lernschritte zu überspringen und schneller und sicherer zum Ziel zu kommen. Im wahrsten Sinne des Wortes ermöglichen induktive Verzerrungen Algorithmen, voreilige Schlüsse zu ziehen. Und wenn Sie die richtigen induktiven Verzerrungen eingefügt haben, wird Ihr Algorithmus auch die richtigen Schlussfolgerungen ziehen.

Was ist also eine induktive Verzerrung? Es handelt sich um eine Veranlagung, eine bestimmte Beziehung in den Daten zu finden (d. h. zu folgern, zu induzieren). Induktive Verzerrungen helfen dem Algorithmus, eine bestimmte Beziehung zu finden, selbst wenn die Beweise sehr schwach sind und andernfalls Millionen von Datenpunkten durchlaufen werden müssten. Induktive Verzerrungen sind eine Art Vorurteil, um eine bestimmte Art von Muster in den Daten zu erkennen.[1] Wenn Ihr mathematisches Modell beispielsweise aus Sinus- und Kosinusfunktionen besteht und Sie hauptsächlich die Parameter solcher Funktionen (z. B. Amplitude und Phase eines Sinus) anpassen, dann wird Ihr Modell wahrscheinlich in der Lage sein, solche Funktionen in den Daten zu finden, selbst bei kleinen Datenmengen. Mit anderen Worten: Das Modell neigt dazu, eine Sinuswelle zu finden.

Was die Leute dazu verleitet, die Bedeutung der induktiven Verzerrungen zu ignorieren, ist die Tatsache, dass man theoretisch dieselbe Art von Sinusmodell verwenden kann, um andere Funktionen als Sinuswellen anzunähern. Man könnte Millionen von Sinuswellen kombinieren, um eine Potenzgesetzfunktion genau zu approximieren. Aber das ist viel schwieriger. Man benötigt ein größeres Modell – d. h. eines mit einer größeren Anzahl elementarer

[1] Induktive Verzerrungen haben nichts mit Verzerrungen in den Daten zu tun, das ist ein ganz anderes Problem.

Sinuswellen und damit einer größeren Anzahl von Parametern –, und man benötigt mehr Daten zum Trainieren.[2] Diese Beziehung gilt für jedes Modell und für alle Daten. Mit ausreichend großen Deep-Learning-Algorithmen kann man fast alles approximieren. Ähnliches gilt für Entscheidungsbäume, die groß genug sind (siehe Abschnitt 6.2.3 über Entscheidungsbäume). Es gibt sogar ein mathematisches Theorem, das universelle Approximationstheorem[3], das beweist, dass ein künstliches neuronales Netz mit nur einer verborgenen Schicht jede beliebige mathematische Funktion approximieren kann, sofern genügend Neuronen in der verborgenen Schicht vorhanden sind [4]. Wo liegt also das Problem, wenn wir alles approximieren können? Warum sollten wir uns Gedanken über die Hinzufügung induktiver Verzerrungen machen, wenn Modelle jede Funktion auch ohne sie approximieren können? Das offensichtlichste Problem habe ich bereits angedeutet: Wenn die induktiven Verzerrungen des Modells nicht gut mit den Daten übereinstimmen, braucht man viele Daten, ein großes Modell und viele Berechnungen. Das bedeutet auch, dass während des Trainings und der Erstellung des Modells mehr CO_2 in die Atmosphäre abgegeben wird. Das alles sind keine guten Nachrichten.

Andererseits können Sie die Modellgröße reduzieren, wenn Sie die richtigen induktiven Verzerrungen hinzufügen. Sie können es dann mit weniger Datenpunkten trainieren, da dieses schlankere Modell nicht so leicht in die lokalen Minima der Überanpassung fällt.[4] Die Vorteile der induktiven Verzerrungen sind der Grund dafür, dass wir so viele verschiedene Modelle haben. Jedes Problem unterscheidet sich ein wenig von jedem anderen Problem und kann daher mit einem spezielleren Satz von Gleichungen optimal angegangen werden. Für jedes Problem gibt es theoretisch ein optimales Modell, das genau auf dieses Problem spezialisiert ist. Daher wird uns der Platz für die Erfindung neuer Modelle nie ausgehen. Die Liste aller möglichen Modelle ist unendlich; wir werden nie das Ende dieser Liste erreichen.

Ich lernte die Macht induktiver Verzerrungen in der Praxis bei einer Gelegenheit kennen, bei der mein Team und ich eine Überanpassung in neuronalen Netzen für tiefes Lernen herbeiführen wollten. Unser Ziel war es, einen Algorithmus zu testen, der die Überanpassung in einer Situation des One-shot Learnings reduziert, und unser Ansatz war folgender: Eine unbegrenzte Menge an Daten für das Training des One-shot-Learning-Algorithmus (siehe Kapitel 10)[5] zu erzeugen, eine Überanpassung auf diesem Datensatz zu induzieren und dann das Netz mit unserem neuen Algorithmus vor der Überanpassung zu „retten". Meine Idee war es, unsere „unbegrenzten Daten" mit einem Deep-Learning-Netz mit einem zufälligen Satz von Gewichten zu erstellen und dann ein anderes naives Deep-Learning-Netz zu trainieren, um dieselben zufälligen Mappings zu lernen. Wir waren zuversichtlich, dass wir auf diese Weise eine Überanpassung erzeugen könnten, aber wir wurden eines Besseren belehrt: Wir reduzierten die Größe des Trainingsdatensatzes immer weiter, aber das neue Netz wollte nicht überanpassen. Die Leistung bei den Testdaten blieb gut, manchmal sogar mit nur zehn oder 20 Datenpunkten. Zunächst waren meine Kollegen und ich

---

[2] Die Fourier-Transformation ist ein Instrument, mit dem sich beurteilen lässt, wie komplex ein auf Sinuswellen basierendes Modell für eine Zeitreihe sein muss. Zeitreihen, die periodisch sind und den Formen von Sinuswellen ähneln, können durch einfache Modelle angenähert werden. Andere benötigen komplexe Modelle und viele Parameter.

[3] *https://en.wikipedia.org/wiki/Universal_approximation_theorem*

[4] *https://en.wikipedia.org/wiki/Overfitting*

[5] Hier kann man etwas über One-shot Learning erfahren: *https://en.wikipedia.org/wiki/One-shot_learning*

verwirrt. Wie war das möglich? Es sollte sich um sehr schwer zu erlernende Daten handeln, mit komplexen zufälligen Beziehungen in einem mehrdimensionalen Raum. Wie konnte das Netz diese Beziehungen mit nur einer kleinen Anzahl von Beispielen lernen? Dieses Lernen war auch dann noch effizient, wenn wir die Architektur des Netzes, die Anzahl der Schichten und die Größe der einzelnen Schichten änderten. Die Fähigkeit zum effizienten Lernen der Daten war robust.

Es dauerte ein paar Tage, bis wir erkannten, dass das Modell, von dem wir hofften, es würde sich übermäßig anpassen, dazu „verdammt" war, da es perfekte induktive Verzerrungen für die Daten aufwies. Wir verwendeten die gleichen ReLu- und sigmoide Transferfunktionen für die Generierung der Daten und für das Modell, das die Daten lernte, was die Arbeit des Lernmodells im Grunde sehr einfach machte. Dies veranschaulichte mir, wie mächtig induktive Verzerrungen sein können: Dasselbe Netz kann eine Million Beispiele benötigen, um etwas zu lernen, das für seine induktiven Verzerrungen kontraintuitiv ist, wie z. B. das Erkennen einer Blume auf einem Foto, und nur zehn Beispiele, um etwas zu lernen, das für jedes andere Modell hochkomplex ist, aber für dieses spezielle Netz vollkommen intuitiv ist. Das liegt daran, dass das Netz genau die richtigen induktiven Verzerrungen hat.[6]

Induktive Verzerrungen geben uns eine Menge Möglichkeiten, mit denen wir bei der Entwicklung von Modellen spielen können. Das Spiel ist zweidimensional. Die eine Dimension bezieht sich auf die Art der induktiven Verzerrungen: Sollen wir ReLu- oder Sigmoid-Übertragungsfunktionen verwenden, oder sollen wir Tangens oder sogar Sinuswellen verwenden? Auf diese Weise ändern wir die Annahmen, die das Modell über die Welt macht. Wir können eine Annahme durch eine andere ersetzen und dadurch die induktiven Verzerrungen ändern. Ein lineares Modell geht von einer bestimmten linearen Beziehung zwischen Daten aus. Ein Entscheidungsbaum geht von einer anderen Annahme aus. Und so weiter.

Die andere Dimension, in der wir mit induktiven Verzerrungen spielen können, ist die Frage, wie eng die Annahmen sind, die wir treffen wollen. Wir können die Annahmen lockerer fassen, was im Grunde bedeutet, dass wir ein Modell mit mehr Parametern haben. Wir können auch ein strengeres Modell mit weniger Parametern erstellen. Indem wir einem neuronalen Netz mehr Einheiten (Neuronen) hinzufügen, lockern wir seine Annahmen. Modelle, die für ein bestimmtes Problem gut geeignet sind, d. h. genau die richtige Menge an induktiven Verzerrungen haben, können oft mit nur einer Handvoll Parameter großartige Arbeit leisten. Die größten Modelle haben heute Milliarden von Parametern. Diese Modelle sind ziemlich entspannt: Es gibt eine ganze Menge verschiedener Dinge, die sie möglicherweise lernen können.

Wie wir bereits erwähnt haben, hat dies direkte Auswirkungen auf die Datenmenge, die zum Lernen benötigt wird. Ein strenges Modell ist natürlich in der Lage, aus nur wenigen Datenpunkten zu lernen, vorausgesetzt, die induktiven Verzerrungen sind korrekt. Wenn die induktiven Verzerrungen nicht korrekt sind, wird ein kleines Modell niemals gut passen, egal wie viele Datenpunkte Sie ihm zum Training geben. Die einzigen beiden Möglichkeiten zur Verbesserung sind entweder die Vergrößerung des Modells (mit einer entsprechenden Vergrößerung des Datensatzes) oder die Korrektur der induktiven Verzerrungen. Daher können Sie auch mit schlechten induktiven Verzerrungen Daten gut anpassen; alles,

[6] Später erfuhr ich, dass jemand denselben Fehler wie wir gemacht und eine ganze Abhandlung veröffentlicht hatte, ohne sich des von uns entdeckten Problems der induktiven Verzerrung bewusst zu sein, was zu der falschen Schlussfolgerung führte, dass neuronale Netze nicht anfällig für Überanpassung sind [5].

was Sie brauchen, sind genügend Parameter und genügend Daten. Deep Learning fällt in die letztgenannte Klasse von Modellen, die nicht spezialisiert sind, lockere Annahmen haben und eine große Datenmenge erfordern. In Bild 9.3 finden Sie die Beziehung zwischen der Anzahl der erforderlichen Daten (ausgedrückt als „Trainingsaufwand“) und der Strenge des Modells (ausgedrückt als „Spezialisierung“) für verschiedene Arten von Modellen. Die strengsten Modelle sind die Gesetze der Physik. Für $E = mc^2$ gibt es z. B. nur einen Parameter, der angepasst werden muss, nämlich c. Dann kann man das „Modell“ verwenden, um E aus m vorherzusagen.

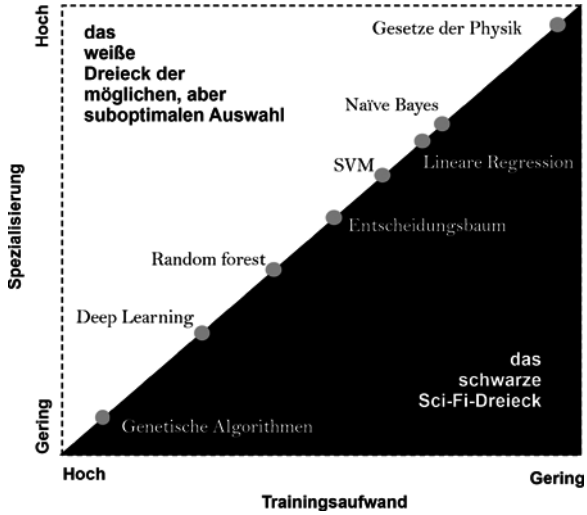

Bild 9.3 Verschiedene Modelle haben unterschiedliche Fähigkeiten zu lernen. Einige benötigen eine große Datenmenge und einen hohen Trainingsaufwand. Andere können mit nur wenigen Beispielen schnell lernen. Ein Modell ist optimal, wenn es irgendwo auf der Diagonalen liegt: In diesem Fall wurde das richtige Modell für die Aufgabe gewählt. Wenn die benötigte Datenmenge und der Trainingsaufwand für den gegebenen Spezialisierungsgrad zu groß sind, dann machen Sie etwas falsch, auch wenn Ihr Modell gut funktioniert (das weiße Dreieck). Es ist unmöglich, ein gut funktionierendes Modell zu haben, das gleichzeitig generisch ist und eine geringe Datenmenge zum Lernen benötigt. Das kann nur in der Fantasie passieren, und manchmal hoffen Data Scientists ganz naiv, ein solches Modell zu finden.

Wie können Sie sich also dieses Wissen über induktive Verzerrungen zunutze machen? Sie können solche Verzerrungen in Ihre Modelle einbauen, damit diese besser und schneller lernen. Auf diese Weise können Sie Modelle kleiner, schneller und zuverlässiger machen. Sie müssen nur die richtigen induktiven Verzerrungen finden. Manchmal müssen Sie auch das Gegenteil tun, nämlich das Modell vergrößern und damit seine Annahmen lockern. Sie müssen herausfinden, was der richtige Ansatz für Ihr Problem ist. Wenn Sie schon einmal eine Hyperparameter-Abstimmung durchgeführt haben[7], dann haben Sie bereits erste Erfahrungen mit der Anpassung der induktiven Verzerrungen von Modellen gemacht. Wenn Sie über gut strukturierte Validierungs- und Trainingsdatensätze verfügen, haben Sie die

[7] *https://en.wikipedia.org/wiki/Hyperparameter_optimization*

18 Vertrauenswürdige KI

Rania Wazir

"All algorithms should be seen as untrustworthy until proven otherwise."
Cathy O'Neil

Fragen, die in diesem Kapitel beantwortet werden:

- Wie sieht der derzeitige Rechtsrahmen für vertrauenswürdige KI aus, insbesondere in der EU?
- Wer sind die möglichen KI-Stakeholder?
- Was ist Fairness in der KI, und wie wird Bias definiert?
- Was sind die verschiedenen Metriken zur Messung der Auswirkungen von Algorithmen auf die Fairness?
- Was sind mögliche Techniken, um unerwünschte Verzerrungen abzuschwächen?
- Wie können Daten und Modelle dokumentiert werden, um ihre Transparenz, Nutzbarkeit und Sicherheit zu verbessern?
- Welche Methoden gibt es, um Modellentscheidungen zu erklären?

Die breite Klasse an Technologien, die unter den Begriff KI fallen – von Expertensystemen bis hin zu datenwissenschaftlichen Anwendungen und Lösungen, die auf maschinellem Lernen basieren –, revolutionieren die Industrie, durchdringen die meisten Wirtschaftssektoren und haben das Potenzial, der Wirtschaft, der Gesellschaft und der Umwelt zu nutzen. Wie sich in den letzten Jahren gezeigt hat, sind diese Technologien jedoch auch mit Risiken verbunden[1] [2] [3]. Mit der Aufdeckung von Beispielen für Stereotypisierung und Diskriminierung, Bedenken hinsichtlich Arbeitnehmerrechte und nachteilige Auswirkungen auf demokratische Grundsätze und die Umwelt hat die Skepsis in der Öffentlichkeit zugenommen. Damit die KI-Technologien weiterhin eine rasch wachsende Akzeptanz finden und ihr nützliches Potenzial entfalten können, wird die Nachfrage nach KI-basierten Systemen, denen man vertrauen kann, steigen. Für die Anbieter von KI-Systemen führt dieses Vertrauen zu einer erhöhten Akzeptanz von Produkten, bei denen es vorhanden ist, und zu rechtlichen

[1] O'Neil, C., Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Broadway Books, 2017.
[2] Agentur der EU für Grundrechte (FRA), Getting the Future Right
[3] Kate Crawford, AI Now Report 2019

und rufschädigenden Konsequenzen, wenn dieses Vertrauen verletzt wird. Im folgenden Kapitel werden wir in der Praxis untersuchen, was Vertrauen in KI-Systeme bedeutet, insbesondere im Zusammenhang mit maschinellem Lernen und datenwissenschaftlichen Lösungen, sowie wer die Interessengruppen sind, die berücksichtigt werden müssen. Hinzu kommen einige praktische Implementierungsschritte, die den Entwicklungsprozess begleiten können.

Unsere Aufgabe wird es sein, die vielen unterschiedlichen Anforderungen miteinander zu verweben, um ein kohärentes Bild zu schaffen, das den Entwicklungsprozess von KI-Systemen von Anfang bis Ende begleiten kann. Wir beginnen mit dem rechtlichen und Soft-Law-Rahmen, indem wir uns prominente Ethikrichtlinien sowie bestehende und künftige Vorschriften und Standards ansehen. Vertrauen hat für verschiedene KI-Stakeholder unterschiedliche Bedeutungen – daher machen wir einen kurzen Abstecher zur Identifizierung von KI-Stakeholdern, bevor wir uns auf die Fragen der Fairness in der KI und der Erklärbarkeit konzentrieren. Dieses Kapitel erhebt keinen Anspruch auf Vollständigkeit, sondern soll vielmehr Anbietern und/oder Nutzern von KI-Systemen, die vertrauenswürdige Produkte entwickeln bzw. einsetzen wollen, eine Orientierungshilfe bieten.

18.1 Rechtlicher und Soft-Law-Rahmen

Seit 2016 gibt es eine explosionsartige Zunahme von sogenannten „Ethikrichtlinien“ für KI. Tatsächlich gab es 2019 bereits über 80 veröffentlichte Richtlinien.[4] Von akademischen Forschungsinstituten bis hin zu großen Technologieunternehmen, von internationalen Nichtregierungsorganisationen bis hin zu staatlichen Regierungen – alle haben ihren Beitrag dazu geleistet, was „ethische“ KI ausmacht. Leider sind die meisten Richtlinien sehr allgemein gehalten und gehen bei den Grundsätzen, die sie für eine „ethische“ KI für notwendig erachten, weit auseinander. Der Untersuchung von Jobin et al.[5] zufolge gibt es fünf allgemeine Grundsätze, auf die sich mindestens die Hälfte der Leitlinien bezieht: Transparenz, Gerechtigkeit und Fairness, Nicht-Malefizierung, Verantwortung und Schutz der Privatsphäre; ihre genaue Bedeutung und die entsprechenden Umsetzungsstrategien sind jedoch wiederum unterschiedlich.

Einige der wichtigsten internationalen Ethik-Leitlinien zur KI sind:

- OECD-Grundsätze zur künstlichen Intelligenz[6]
- UNESCO-Empfehlung zur Ethik der KI[7]
- UNICEF-Leitlinien zu künstlicher Intelligenz für Kinder[8]

[4] Jobin, Anna, Marcello Ienca und Effy Vayena. „The global landscape of AI ethics guidelines“. Nature Machine Intelligence 1.9 (2019): 389 – 399.

[5] Jobin, Anna, Marcello Ienca und Effy Vayena. „The global landscape of AI ethics guidelines“. Nature Machine Intelligence 1.9 (2019): 389 – 399.

[6] *https://www.oecd.ai/ai-principles*

[7] *https://unesdoc.unesco.org/ark:/48223/pf0000373434*

[8] *https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children*

- EU-HLEG-Leitlinien für vertrauenswürdige KI[9]
- EU-Whitepaper zu KI[10]

Eine vertrauenswürdige KIgeht jedoch über ethische Aspekte hinaus. Eine offensichtliche zusätzliche Anforderung ist ein Qualitätsgebot: Das System sollte robust, zuverlässig und sicher sein. Die OECD-Prinzipien beispielsweise richten sich an Regierungen und andere staatliche Akteure und sollen als Leitfaden für die Förderung der Entwicklung vertrauenswürdiger KI dienen. Sie schlagen die folgenden fünf Hauptprinzipien vor:[11]

1. **Integratives Wachstum, nachhaltige Entwicklung und Wohlstand.** Eine allgemeine Voraussetzung für den Nutzen vertrauenswürdiger KI ist die Verbesserung der menschlichen Fähigkeiten, der Abbau von Ungleichheiten und der Schutz der Umwelt.
2. **Auf den Menschen ausgerichtete Werte und Fairness.** Eine vertrauenswürdige KI muss die Rechtsstaatlichkeit und die Menschenrechte achten, einschließlich des Rechts auf Freiheit, des Rechts auf Würde und Autonomie, des Rechts auf Privatsphäre und Datenschutz sowie des Rechts auf Nichtdiskriminierung.
3. **Transparenz und Erklärbarkeit.** Verlangt eine verantwortungsvolle Offenlegung von Informationen über das KI-System, um das allgemeine Verständnis für solche Systeme zu fördern, die Betroffenen auf ihre Interaktionen mit einem KI-System aufmerksam zu machen und es den von einem KI-System Betroffenen zu ermöglichen, dessen Ergebnisse zu verstehen und anzufechten.
4. **Robustheit, Sicherheit und Schutz.** Rückverfolgbarkeit von Datensätzen, Prozessen und Entscheidungen sowie geeignete Maßnahmen zum Risikomanagement, um Risiken wie Sicherheit, IT-Sicherheit, Datenschutz und Verzerrungen zu vermeidenin jeder Phase des Lebenszyklus eines KI-Systems.
5. **Rechenschaftspflicht.** Alle Akteure, die an der Entwicklung, dem Einsatz oder dem Betrieb von KI-Systemen beteiligt sind, sollten entsprechend ihrer Rolle für das ordnungsgemäße Funktionieren der KI-Systeme verantwortlich gemacht werden, einschließlich der Gewährleistung, dass die oben genannten Anforderungen erfüllt werden.

Die hochrangige EU-Sachverständigengruppe für KI hat eine noch umfangreichere Liste von Anforderungen an eine vertrauenswürdige KI aufgestellt, die sich an Entwickler, Anbieter und Nutzer von KI-Systemen richtet. Eine vertrauenswürdige KI muss legal, ethisch und robust sein und sollte die folgenden Anforderungen erfüllen:[12]

1. **Menschliches Handeln und Aufsicht.** Einschließlich Grundrechte, menschliches Handeln und menschliche Aufsicht
2. **Technische Robustheit und Sicherheit.** Einschließlich Widerstandsfähigkeit gegen Angriffe, Sicherheit, Ausweichplan, Genauigkeit, Zuverlässigkeit und Reproduzierbarkeit

---

[9] *https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai*

[10] *https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en*

[11] *https://www.oecd.ai/ai-principles*

[12] High Level Expert Group on Artificial Intelligence set up by the European Commission, „Ethics Guidelines for Trustworthy AI“, April 2019, S. 14. Abgerufen von *https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai*

3. **Datenschutz und Data Governance.** Einschließlich Achtung der Privatsphäre, der Qualität und Integrität der Daten und des Datenzugangs
4. **Transparenz.** Einschließlich Rückverfolgbarkeit, Erklärbarkeit und Kommunikation
5. **Vielfalt, Nichtdiskriminierung und Fairness.** Einschließlich der Vermeidung unfairer Voreingenommenheit (Bias), Zugänglichkeit und universellem Design sowie der Beteiligung von Interessengruppen
6. **Gesellschaftliches und ökologisches Wohlergehen.** Einschließlich Nachhaltigkeit und Umweltfreundlichkeit, sozialer Auswirkungen, Gesellschaft und Demokratie
7. **Rechenschaftspflicht.** Dazu gehören Überprüfbarkeit, Minimierung und Meldung negativer Auswirkungen, Abgleichungen und eventuelle Rechtsmittel.

Die HLEG-Leitlinien sind vielleicht eine der praktischsten Leitlinien, die es bisher gibt. Sie vermitteln ein klares Verständnis der den Anforderungen zugrundeliegenden Überlegungen und geben Informationen darüber, wie sie in der Praxis umgesetzt werden können. Auf der Grundlage des Leitfadens hat die Gruppe auch die Bewertungsliste für vertrauenswürdige KI (ALTAI) entwickelt,[13] ein Instrument, das Anbietern, Entwicklern und Nutzern von KI-Systemen helfen soll zu bewerten, inwieweit ihr KI-System die sieben Anforderungen an eine vertrauenswürdige KI erfüllt.

18.1.1 Normen

Der Weg von den Leitlinien zur praktischen Umsetzung ist lang, und die Regulierung und internationale Normen sind notwendige Zwischenschritte. Mehrere internationale Normungsorganisationen arbeiten aktiv an der Erstellung der für die Gewährleistung vertrauenswürdiger KI erforderlichen Normen:

- **IEEE Ethically Aligned Design:** *https://ethicsinaction.ieee.org/#series*. Das IEEE hat eine eigene Reihe von ethischen Leitlinien, die fast 300 Seiten umfassen.[14] Diese werden durch die Normenreihe 7000 ergänzt, in der besondere Aspekte der ethischen KI spezifiziert werden. Die ersten beiden, die veröffentlicht wurden, behandeln allgemeine Grundsätze des ethischen Designs und Spezifikationen für die Messung der Auswirkungen von autonomen und intelligenten Systemen auf das menschliche Wohlbefinden.
- **ISO/IEC-Normen zu KI und vertrauenswürdiger KI:** *https://www.iso.org/committee/6794475.html*. ISO und IEC haben einen gemeinsamen Ausschuss eingerichtet, der sich mit künstlicher Intelligenz befasst. Mehrere Normen und technische Berichte sind bereits veröffentlicht worden, und viele weitere sind in Vorbereitung. Insbesondere der kürzlich veröffentlichte ISO/IEC TR 24028: Overview of trustworthiness in artificial intelligence (Überblick über die Vertrauenswürdigkeit künstlicher Intelligenz)[15] gibt einen Überblick über die Anforderungen und Fallstricke bei der Entwicklung und dem Einsatz eines vertrauenswürdigen KI-Systems und kann als Fahrplan für künftige Normenspezifikationen angesehen werden.

[13] *https://altai.insight-centre.org/*
[14] *https://ethicsinaction.ieee.org/#ead1e*
[15] *https://www.iso.org/standard/77608.html*

- **NIST-Standards für vertrauenswürdige und verantwortungsvolle KI:** *https://www.nist.gov/programs-projects/trustworthy-and-responsible-ai*. Das Projekt des NIST umfasst Standards für verschiedene Schlüsselaspekte der vertrauenswürdigen KI, einschließlich eines kürzlich veröffentlichten Entwurfs zur Abschwächung schädlicher Verzerrungen[16] sowie bereits veröffentlichte Standards zu Erklärbarkeit und Sicherheit.
- **CEN-CENELEC-Ausschuss für künstliche Intelligenz:** *https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/*. CEN und CENELEC haben das neue gemeinsame Komitee als Reaktion auf das Whitepaper der Europäischen Kommission zur künstlichen Intelligenz und die deutsche Normungs-Roadmap für künstliche Intelligenz gegründet.[17]

18.1.2 Verordnungen

Insbesondere in der EU wurde die Entwicklung einer digitalen Strategie vorangetrieben, die über Leitlinien hinausgeht und der KI-Branche eine gewisse Regulierung auferlegt. Der erste Rechtsakt in dieser Richtung war die Allgemeine Datenschutzverordnung (DSGVO), die 2018 in Kraft getreten ist. Weitere Verordnungen sind in Vorbereitung – zum Beispiel das Gesetz über digitale Dienste (DSA) und das Gesetz über digitale Märkte (DMA), deren Ziel es ist, den „Gatekeeper"-Effekt sehr großer Online-Plattformen zu verringern und den Nutzern und Verbrauchern mehr Transparenz und Wahlmöglichkeiten gegenüber diesen Plattformen zu schaffen (DSA) und kleineren Akteuren den Eintritt in die Plattformökonomie und den Wettbewerb zu ermöglichen (DMA). Während diese Verordnungen jedoch Elemente mit direkten Auswirkungen auf die Datenerhebung und die Transparenz von KI-Systemen enthalten, ist die zentrale Verordnung, die sich mit KI befasst, die EU-KI-Verordnung, die im April 2021 als Entwurf veröffentlicht wurde.

- **Digitale Strategie der EU:** *https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en*
- **DSGVO (GDPR):** *https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en*
- **DSA:** *https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en*
- **DMA:** *https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en*
- **EU-Entwurf eines KI-Gesetzes:** *https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence*

Der Entwurf des KI-Gesetzes richtet sich an alle KI-Systeme, die in der EU in Verkehr gebracht oder in Betrieb genommen werden. Er verfolgt bei der Regulierung von KI einen risikobasierten Ansatz, bei dem das Risiko nicht nur physische oder psychische Schäden, sondern auch Risiken für die Grundrechte umfasst. Der Entwurf des KI-Gesetzes definiert den Begriff „KI" absichtlich weit und schließt viele Algorithmen ein, deren Einordnung als

[16] *https://doi.org/10.6028/NIST.SP.1270-draft*
[17] *https://www.din.de/en/innovation-and-research/artificial-intelligence*

„KI“ heftige Diskussionen ausgelöst hat: nicht nur Algorithmen des maschinellen Lernens, sondern auch logikbasierte Methoden und Expertensysteme, statistische und Bayesische Verfahren, Optimierung und Suche. Die vollständige Liste ist in Anhang I des Entwurfs des KI-Gesetzes zu finden.

Der Entwurf zum KI-Gesetz nennt derzeit vier Arten von Anwendungen, die verboten sind. Dazu gehören unterschwellige Manipulationen, soziales Scoring und Gesichtserkennung:

- KI-Systeme, die Menschen manipulieren und sie zu Verhaltensweisen verleiten können, die ihnen selbst oder anderen physisch oder psychisch schaden.
- KI-Systeme, die die Schwächen bestimmter Gruppen aufgrund ihres Alters oder einer geistigen oder körperlichen Behinderung ausnutzen und zu einem Verhalten führen können, das für sie selbst oder andere physisch oder psychisch schädlich ist.
- Soziales Scoring durch öffentliche Stellen
- Die Verwendung von biometrischen Echtzeit-Fernerkennungssystemen in öffentlich zugänglichen Räumen zu Strafverfolgungszwecken (dieses Verbot ist jedoch mit mehreren Ausnahmen verbunden).

Der Hauptinhalt der vorgeschlagenen Verordnung ist jedoch für Anwendungen mit hohem Risiko bestimmt. Diese sind in Anhang II – der eine Liste von Anwendungen enthält, die bereits einer sektoralen Regulierung unterliegen und für die der Rechtsakt zusätzliche Verpflichtungen vorsieht – und in Anhang III aufgeführt, in dem acht neue Anwendungsbereiche genannt werden, wobei in jedem Bereich spezifische Anwendungsfälle mit hohem Risiko genannt werden. Anhang II umfasst u.a. KI-Systeme, die in Spielzeug, Maschinen, medizinischen Geräten, in der Luftfahrt, in Kraftfahrzeugen und anderen Verkehrsmitteln eingesetzt werden. Die in Anhang III aufgeführten Anwendungsbereiche sind:

1. Biometrische Identifizierung und Kategorisierung von natürlichen Personen
2. Verwaltung und Betrieb von kritischen Infrastrukturen
3. Allgemeine und berufliche Bildung
4. Beschäftigung, Arbeitnehmermanagement und Zugang zur selbstständigen Arbeit
5. Zugang zu und Inanspruchnahme von wesentlichen privaten und öffentlichen Diensten und Leistungen
6. Strafverfolgung
7. Verwaltung von Migration, Asyl und Grenzkontrollen
8. Rechtspflege und demokratische Prozesse

Das Neue an Anhang III ist, dass der Entwurf des KI-Gesetzes der Kommission das Recht vorbehält, neue Anwendungsfälle in den Anhang aufzunehmen, wenn diese zu einem der acht Anwendungsbereiche gehören und sich herausstellt, dass sie ein hohes Risiko für Sicherheit, Gesundheit oder Grundrechte darstellen. Dies ermöglicht es der Kommission, erneute parlamentarische Verhandlungen über eventuelle Änderungen zu umgehen, und verschafft ihr ein gewisses Maß an Flexibilität, um auf neue Erkenntnisse über Schäden zu reagieren.

Die vorgeschlagene Verordnung stellt einige Anforderungen an Anbieter von KI-Systemen mit hohem Risiko, wenngleich in den meisten Fällen keine externe Prüfung erforderlich ist und eine Selbstbewertung ausreicht. Die wichtigsten Anforderungen beziehen sich auf

Datenqualität und Governance (Artikel 10), Risikobewertung und Risikomanagementsysteme (Artikel 9), Modellleistungstests (Artikel 15) und Modelldokumentation (Artikel 11, Anhang IV).

18.2 KI-Stakeholder

KI-Systeme sind in komplexe Ökosysteme eingebettet, an denen ein breites Spektrum von Akteuren beteiligt ist. Um die Risiken für Bias in der KI-Anwendung zu verstehen und sie zu mindern, muss man die verschiedenen Akteure, ihre Rollen und ihre Bedürfnisse erfassen. Die folgende Liste kann als Leitfaden dienen, ist aber keineswegs erschöpfend.

- **Datenprovider:** Organisation/Person, die die vom KI-Anbieter verwendeten Daten sammelt, verarbeitet und bereitstellt.
- **KI-Provider:** Organisation/Person, die KI-Systeme entwickelt. Innerhalb der Organisation können spezifische zusätzliche Rollen identifiziert werden.
  - Management und Vorstand
  - Rechtsabteilung/Abteilung für Unternehmensverantwortung
  - Datenschutzbeauftragte
  - SystemarchitektInnen, DateningenieurInnen
  - EntwicklerInnen, IngenieurInnen für maschinelles Lernen, DatenwissenschaftlerInnen
  - Qualitätssicherung
- **KI-Nutzer:** Organisation/Person, die ein KI-System einsetzt. Innerhalb der Organisation können spezifische zusätzliche Rollen identifiziert werden.
  - Management und Vorstand
  - Rechtsabteilung/Abteilung für Unternehmensverantwortung
  - Qualitätssicherung
  - Datenschutzbeauftragte
  - SystemarchitektInnen, DateningenieurInnen
  - Personalwesen
  - Beschaffung
  - Personen, die direkt mit dem neuen KI-System arbeiten müssen oder deren Arbeitsplätze durch das neue KI-System ersetzt werden
- **KI-Subjekt:** Organisation/Person, auf die sich die Ergebnisse/Vorhersagen des KI-Systems beziehen.
- **Zertifizierungsstelle:** Organisation, die die Einhaltung festgelegter Standards bescheinigt.
- **Regulierungsbehörde:** Behörde, die Leistungskriterien für KI festlegt, die in ihrem Zuständigkeitsbereich eingesetzt wird.

- **Die breitere Gesellschaft, z. B. Menschenrechtsorganisationen, Verbraucherschutzorganisationen, Umweltschutzorganisationen und Medien:** Sie müssen über die Anforderungen an vertrauenswürdige KI informiert werden und sollten in der Lage sein, deren Einhaltung zu einzufordern.

18.3 Fairness in der KI

Was ist ein fairer Algorithmus? Die Diskussionen finden hauptsächlich auf Englisch statt, weshalb wir uns des Weiteren auf die englischen Begriffe und ihre Definitionen beziehen. Hierzu ein Zitat aus dem Oxford English Dictionary:

> **Fairness**[18]**:** *Impartial and just treatment or behaviour without favouritism or discrimination.*

Diese Definition ist noch nicht umsetzbar – um zu bestimmen, ob ein KI-System fair ist, muss das Konzept irgendwie quantifiziert werden. Fairness ist jedoch ein soziales Konstrukt, das vom Kontext und von kulturellen/gesellschaftlichen Normen abhängt. Dies hat dazu geführt, dass es viele verschiedene Definitionen von Fairness gibt (21 und mehr),[19] jede mit ihrer eigenen mathematischen Formulierung (Fairness-Metrik) – wie im Folgenden beschrieben wird. Um die Verwirrung noch zu vergrößern, werden in der einschlägigen (meist englischen) Literatur die Begriffe „unfair Algorithm" und „biased Algorithm" häufig synonym verwendet.

> **Bias** *(laut Oxford English Dictionary)*[20]*: Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.*
>
> *1.1 A concentration on or interest in one particular area or subject.*
>
> *1.2 A systematic distortion of a statistical result due to a factor not allowed for in its derivation.*

Diese Vermengung von unfair und bias mag natürlich erscheinen, wenn man die Hauptdefinition von Bias betrachtet. Dennoch ist es wichtig zu bedenken, dass jedes Klassifizierungsmodell einen Bias haben muss, um zu funktionieren. Nehmen wir zum Beispiel einen Klassifizierer, der zwischen Bildern von Säugetieren und Vögeln unterscheiden muss. Es muss eine Tendenz haben, Bilder von Tieren mit Flügeln als Vögel zu klassifizieren. Wäre es hingegen völlig frei von Bias (Voreingenommenheit), könnte es überhaupt keine Unterscheidung treffen und würde alle Objekte in dieselbe Kategorie einordnen. Daher ist eine erste Klarstellung erforderlich: Algorithmen müssen *unerwünschten* Bias vermeiden – also Vorurteile, die auf einem geschützten Merkmal oder einer falschen Korrelation beruhen und für die jeweilige Aufgabe nicht relevant sind.

---

[18] Unparteiische und gerechte Behandlung oder Verhalten ohne Bevorzugung oder Diskriminierung.

[19] Verma, S. und Rubin, J. (2018), „Fairness Definitions Explained", Proceedings of the International Workshop on Software Fairness (FairWare), S. 1 – 7.

[20] Neigung oder Vorurteil für oder gegen eine Person oder Gruppe, insbesondere in einer Weise, die als ungerecht empfunden wird.
1.1 Konzentration auf oder Interesse an einem bestimmten Bereich oder Thema.
1.2 Systematische Verzerrung eines statistischen Ergebnisses durch einen bei der Herleitung nicht berücksichtigten Faktor.

Darüber hinaus gibt es in den Ingenieurs- und Statistikkreisen bereits eine bestimmte Art von unerwünschtem Bias: Bias im Sinne der Definition 1.2 (statistische Verzerrung). Dies führt häufig zu Verwirrung und Missverständnissen bei der Diskussion über Bias im maschinellen Lernen: Einfach ausgedrückt kann ein „fairer" Algorithmus immer noch statistische Verzerrungen aufweisen, während ein System, das frei von statistischem Bias (Verzerrungen) ist, dennoch unfair sein kann.

Der Kern des Problems liegt in der Definition: Eine „systematische Verzerrung eines statistischen Ergebnisses" setzt voraus, dass eine „Grundwahrheit" (oder ein „wahrer Wert" bekannt ist, sodass eine systematische Verzerrung durch Vergleich festgestellt werden kann. Aber was ist diese „Grundwahrheit"? Wenn es sich dabei, wie bisher üblich, um den aktuellen Parameterwert der Bevölkerung handelt, dann sollte es nicht überraschen, dass z. B. ein Einstellungsalgorithmus für eine Ingenieurstelle, der auf der Grundlage historischer Beschäftigungsdaten trainiert wurde, Frauen benachteiligen würde, eben weil er den Status quo genau widerspiegelt (und daher keine statistische Verzerrung aufweist). Dies ist nicht nur eine bloße Hypothese – man denke nur an das von Amazon verworfene, auf maschinelles Lernen gestützte Rekrutierungstool.[21] Umgekehrt könnte es als notwendig erachtet werden, statistische Verzerrungen in den Algorithmus einzubauen, um mehr Geschlechtergerechtigkeit zu erreichen und „fairness" in den Algorithmus einzubauen. Natürlich könnte dieser Widerspruch zwischen statistischer Verzerrung und Fairness nicht aufkommen, wenn die „Grundwahrheit" als ein idealisiertes Ziel angesehen würde (d. h. die ideale Geschlechterverteilung bei den Ingenieuren). Dies ist jedoch ein umstrittenes Thema, und eine Änderung der Terminologie würde immer noch das grundlegende Problem, wie denn die ideale Verteilung aussehen sollte, nicht lösen. Aus diesem Grund vermeiden viele aktuelle Fairness-Kennzahlen die Verwendung einer „Grundwahrheit" als Referenzparameter.

Um Verwirrung zu vermeiden, werden wir in diesem Kapitel den Begriff Bias verwenden, um die Eingaben in ein maschinelles Lernmodell (oder allgemeiner ein KI-System) oder dessen Eigenschaften zu beschreiben. Fairness hingegen wird verwendet, um die Auswirkungen von modellbasierten Ergebnissen oder Vorhersagen auf verschiedene geschützte Bevölkerungsgruppen zu beschreiben. Dies steht auch im Einklang mit einer wachsenden Anzahl an Veröffentlichungen, in denen versucht wird, Quellen von Bias in KI-Systemen zu identifizieren und zu reduzieren, und in denen Fairness-Metriken zur Bewertung der Auswirkungen von Modellen verwendet werden.

18.3.1 Bias

Bias kann in vielen Formen auftreten und in den Lebenszyklen von maschinellem Lernen und Data Science in verschiedenen Phasen auftreten. Es können vier Hauptphasen identifiziert werden:

1. Der Bias kann in den Trainings- oder Testdaten liegen. Eine große Datenmenge befreit Datensammler nicht automatisch von den traditionellen statistischen Datenfehlern. Stichprobenverzerrung, Auswahlverzerrung und Antwortausfall sind nur einige der häu-

[21] Dastin, J. (2018), „Amazon scraps secret AI recruiting tool that showed bias against women", Reuters, 11. Oktober 2018.

figsten Fallen, die Daten für Unwissende bereithalten. Wie das obige Beispiel für das Training eines Einstellungsalgorithmus anhand historischer Daten zeigt, kann aufgrund menschlicher Vorurteile Bias selbst dann in den Daten bestehen, wenn das Verfahren zur Beschaffung der Daten statistisch korrekt war. Die Einstellungsdaten, die zum Trainieren des Algorithmus verwendet werden, könnten den Status quo genau widerspiegeln – und somit die derzeitige gesellschaftliche Voreingenommenheit gegenüber Frauen im Ingenieurwesen kodieren und aufrechterhalten. Worteinbettungen und Sprachmodelle sind ein weiteres Beispiel für diese Art von Bias – der Text, der zum Trainieren dieser Modelle verwendet wird, ist voller gesellschaftlicher Vorurteile, sodass die Worteinbettungen nicht nur allgemeine semantische Muster, sondern auch geschlechtsspezifische[22] und ethnische[23] Stereotypen und Vorurteile widerspiegeln.

2. Bias kann auch beim Entwurf des Algorithmus in das System einfließen – so könnte ein Klassifizierungssystem aufgrund der Kategorien, für die es entwickelt wurde (schwarz/weiß, männlich/weiblich),[24] verzerrt sein; Bias kann bei der Entwicklung von Merkmalen auftreten (einige Merkmale könnten für einige Gruppen aussagekräftiger sein als für andere, und die Auswahl von Merkmalen auf der Grundlage der Gesamtgenauigkeit könnte dazu führen, dass das Modell für einige Gruppen schlechter abschneidet) oder bei der Wahl des zu verwendenden Algorithmus (zu einfache Algorithmen können beispielsweise die Daten nicht richtig erfassen und zu Verzerrungen in den Modellen führen). Eine besonders heimtückische Form von Bias kann in den Algorithmusentwurf einfließen, wenn versucht wird, ein Konzept zu modellieren, das nicht vollständig quantifizierbar ist – z. B. bei der Zulassung an einer Universität, wenn Aufzeichnungen über zuvor zugelassene Studenten verwendet werden, um ein Modell zur Erkennung erfolgreicher Kandidaten für ein Doktorandenprogramm zu trainieren[25] (in Wirklichkeit werden dadurch einfach die Präferenzen und Vorurteile früherer Zulassungsausschüsse modelliert); oder bei der Verwaltung der Krankenhausversorgung, wenn die Kosten für die Gesundheitsversorgung als Indikator für die Schwere der zu behandelnden Krankheit verwendet werden.[26]
3. Bias kann auch post hoc in das System einfließen, zum Beispiel bei der Interpretation der Modellergebnisse. Oder Entscheidungen, die auf Modellvorhersagen beruhen, könnten sich auf Daten auswirken, die dann wieder in einen Online-Lernalgorithmus eingespeist werden, was zur Bildung von unkontrollierbaren Rückkopplungsschleifen führt[27] und bestehenden Bias in den Daten oder im Modell noch verstärkt.

---

[22] Bolukbasi, T., Chang, K.-W., Zou, J., Saligramma, V., Kalai, A. (2016), „Man is to computer programmer as woman is to homemaker? debiasing word embeddings“, Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, S. 4356 – 4364.

[23] Manzini, T., Yao Chong, L., Black, A. W., Tsvetkov, Y. (2019), „Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings“, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, S. 615 – 621.

[24] Leufer, D. (2021), „Computers are binary, people are not: how AI systems undermine LGBTQ identity“, Access Now, April 2021.

[25] Burke, L. (2020), U of Texas will stop using controversial algorithm to evaluate Ph. D. applicants, Inside Higher Ed, 14 December 2020.

[26] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019), „Dissecting racial bias in an algorithm used to manage the health of populations“, Science, Vol. 366, S. 447 – 453.

[27] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. und Venkatasubramanian, S. (2018), „Runaway feedback loops in predictive policing“, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, Vol. 81, S. 160 – 171.

4. Und schließlich ist der Einsatz von KI auch anfällig für Bias: von der zeitlichen Abweichung über die unangemessene Nutzung (in einem anderen als dem beabsichtigten Kontext) und von feindlichen Angriffen (man denke nur an den berüchtigten Chatbot Tay von Microsoft[28]) bis hin zum selektiven Einsatz (z. B. Verwendung von Vorhersagemodellen zur Ermittlung von Noten für Kinder in größeren Klassen, aber Verwendung menschlicher Bewertungen zur Ermittlung von Noten für Kinder in kleineren Klassen)[29].

Es ist nicht möglich, alle erdenklichen Arten von Bias aufzulisten, die in ein Modell für maschinelles Lernen einfließen können; wir beschreiben aber im Folgenden kurz einige der häufigsten Formen von Bias.[30]

- **Menschlicher kognitiver Bias:** Jede Art von Voreingenommenheit, die bei der Verarbeitung und Interpretation von Informationen durch Menschen auftreten kann
- **Gesellschaftlicher Bias:** Voreingenommenheit und Vorurteile, die sich aus einem sozialen, kulturellen oder historischen Kontext ergeben
- **Confirmation Bias:** Eine Tendenz, Modellvorhersagen zu akzeptieren, die mit den eigenen bereits bestehenden Überzeugungen übereinstimmen
- **Group Attribution Bias:** Tritt auf, wenn angenommen wird, dass das, was für ein Individuum in einer Gruppe gilt, auch für alle Mitglieder dieser Gruppe gilt
- **Automation Bias:** Eine Tendenz, sich zu sehr auf die Ergebnisse eines Vorhersagemodells zu verlassen
- **Zeitliche Verzerrung:** Bias, der dadurch entsteht, dass Unterschiede in den beobachteten/gemessenen Größen im Zeitverlauf nicht berücksichtigt werden
- **Stichprobenverzerrung (Sampling Bias):** Tritt auf, wenn die Daten nicht nach dem Zufallsprinzip aus der vorgesehenen Grundgesamtheit gezogen werden, sodass einige Personen mit größerer Wahrscheinlichkeit in die Stichprobe aufgenommen werden als andere
- **Repräsentationsverzerrung:** Tritt auf, wenn Einzelpersonen oder Gruppen in einer Studie systematisch von der Population von Interesse abweichen. Dies kann zwar den Fall der Stichprobenverzerrung einschließen, ist aber ein breiteres Konzept. Selbst wenn die Daten nach dem Zufallsprinzip aus der Gesamtbevölkerung entnommen werden, kann der Stichprobenumfang bzw. die Datenqualität für bestimmte Untergruppen gering sein, was zu Ergebnissen führt, die sich nicht gut auf diese Untergruppen verallgemeinern lassen.
- **Measurement Bias:** Diese Art von Messfehlern kann auftreten, wenn die im Modell verwendeten Merkmale und/oder Bezeichnungen Stellvertreter (Proxys) für die tatsächlich interessierende Größe sind, wodurch möglicherweise systematische Fehler zwischen dem, was beabsichtigt ist, und dem, was tatsächlich gemessen wird, entstehen (wie im oben genannten Beispiel der Verwendung von Gesundheitskosten zur Messung der Schwere einer Krankheit)[31].

---

[28] The Guardian (2016), „Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot", 26. März 2016.

[29] Elbanna, A., Engesmo, J. (2020), „A-level results: why algorithms get things so wrong – and what we can do to fix them", The Conversation, August 19, 2020.

[30] Suresh, H. und Guttag, J. (2021), „A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle", arXiv preprint, *https://arxiv.org/pdf/1901.10002.pdf*

[31] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019), „Dissecting racial bias in an algorithm used to manage the health of populations", Science, Vol. 366, S. 447–453.