Was ist KI?

Man könnte glauben, dass KI bereits alle Lebensbereiche durchdrungen hat. Das liegt zum Teil daran, dass der Begriff »künstliche Intelligenz« so viele verschiedene Bedeutungen hat. Je nachdem, ob Sie Science-Fiction-Literatur lesen, eine neue App verkaufen oder wissenschaftliche Forschung betreiben, wird er anders interpretiert. Wenn jemand sagt, er habe einen Chatbot mit KI-Fähigkeiten, können wir dann erwarten, dass er Meinungen und Gefühle hat wie C-3PO? Oder ist es nur ein Algorithmus, der gelernt hat, zu erraten, wie Menschen wahrscheinlich auf einen bestimmten Satz reagieren? Oder eine Tabelle, die Wörter in Ihrer Frage mit vorformulierten Antworten abgleicht? Oder ein unterbezahlter Mensch, der alle Antworten von irgendeinem Ort der Welt von Hand eintippt? Oder vielleicht sogar eine komplett vorgefertigte Unterhaltung, in der Mensch und KI vordefinierte Textzeilen vorlesen wie die Figuren in einem Theaterstück? Verwirrenderweise wurden alle diese Definitionen schon benutzt, um zu erklären, was KI bedeutet.

In diesem Buch benutzen wir den Begriff *künstliche Intelligenz* (*KI*) so, wie er heutzutage üblicherweise von Programmierern verwendet wird: als eine bestimmte Art von Computerprogramm, die als **Algorithmus für maschinelles Lernen** bezeichnet wird. Das folgende Diagramm zeigt eine Reihe von Begriffen, die wir in diesem Buch behandeln und wie sie nach dieser Definition einzuordnen sind.

Als KI bezeichnete Dinge

KI in diesem Buch:

Algorithmen für maschinelles Lernen
Deep Learning
Neuronale Netzwerke
Rekurrente neuronale Netzwerke
Markow-Ketten
Random Forests
Genetische Algorithmen
Generative Adversarial Networks
Reinforcement Learning
Smartphone-Textvorschläge
Magische Sandwich-Sortierer
Unglückliche Mörder-Roboter

In diesem Buch, aber keine KI:

Science-Fiction-KIs
Regelbasierte Programmierung
Als Roboter verkleidete Menschen
Roboter, die Skripte vorlesen
Menschen, die vorgeben, KIs zu sein
Kakerlaken mit Bewusstsein
Phantomgiraffen

Alles, was in diesem Buch als KI bezeichnet wird, ist auch ein Algorithmus für maschinelles Lernen. Daher wollen wir zunächst einmal sehen, was das eigentlich ist.

Klopf klopf, wer ist da?

Um eine KI in freier Wildbahn zu erkennen, müssen Sie zuerst den Unterschied zwischen **Algorithmen für maschinelles Lernen** (die wir in diesem Buch als KI bezeichnen) und traditionellen Programmen kennen (die Programmierer üblicherweise als **regelbasiert** bezeichnen). Wenn Sie sich jemals mit einfacher Programmierung beschäftigt oder auch nur eine HTML-Seite erstellt haben, werden Sie mit großer Sicherheit ein regelbasiertes Programm benutzt haben. Sie erstellen eine Liste mit Befehlen oder Regeln in einer Computersprache, und der Computer tut genau das, was Sie ihm gesagt haben. Um mit einem regelbasierten Programm ein Problem zu lösen, müssen Sie alle dafür nötigen Schritte kennen und beschreiben können.

Ein Algorithmus für maschinelles Lernen findet die Regeln dagegen durch selbstständiges Ausprobieren, indem er seine Fortschritte ständig mit dem vom Programmierer vorgegebenen Ziel abgleicht. Das kann eine Liste mit Beispielen sein, die imitiert werden sollen, ein Spielstand, der verbessert werden soll, oder etwas vollkommen anderes. Während die KI versucht, ihr Ziel zu erreichen, kann sie sogar Regeln und Beziehungen finden, deren Existenz der Programmierer nicht einmal geahnt hat. Die Programmierung einer KI hat mehr Ähnlichkeit damit, einem Kind etwas beizubringen, als damit, einen Computer zu programmieren.

Regelbasierte Programmierung

Angenommen, wir wollten regelbasierte Programmierung benutzen, um einem Computer beizubringen, wie man Klopf-klopf-Witze* erzählt. Zuerst müssen wir herausfinden, nach welchen Regeln die Witze funktionieren. Ich würde ihre Struktur analysieren und feststellen, dass sie nach einer Art Formel ablaufen:

```
Klopf klopf.
Wer ist da?
[Name/Begriff]
[Name/Begriff] Wer?
[Name/Begriff] [Pointe]
```

Sobald wird diese Formel kennen, gibt es nur noch zwei Stellen, die das Programm erzeugen muss: [Name/Begriff] und [Pointe]. Aber auch hierfür werden Regeln gebraucht.

Ich könnte eine Liste mit Namen/Begriffen und dazu passenden Pointen erstellen, wie etwa diese:

Namen/Begriffe	***Pointen***
Mani	*Manitu. Für Mani tu ich alles.*
Muh	*Muh-Tation.*
Salat	*Klopf-Salat.*
Werner	*Wer nervt mich andauernd mit diesen Klopf-klopf-Witzen?*
Karla	*Karla Gerfeld.*

* Eine Form von Witzen, die besonders in den USA sehr populär ist und so ähnlich funktioniert wie die Häschen-Witze (»Hattu Möhrchen?«) im deutschsprachigen Raum.

Namen/Begriffe	*Pointen*
Tupper	*Tupperware.*
Australien	*Kängurus kommen Aus Tralien.*
Corona	*Corona-Impfstoff. Bitte warten! Bitte warten!*

Jetzt kann der Computer einen Klopf-klopf-Witz erzeugen, indem er ein Paar aus Name und Pointe auswählt und in die Vorlage einfügt. Hierbei werden allerdings *keine neuen Witze* erzählt, sondern nur solche, *die wir bereits kennen*. Ich könnte versuchen, das etwas interessanter zu machen, indem wir [Manitu] zum Beispiel durch [Manipulation] ersetzen. Dann kann das Programm einen neuen Witz erzeugen:

```
Klopf klopf.
Wer ist da?
Mani
Mani wer?
Manipulation. Du wirst jetzt ganz müde.
```

Ich könnte [Manitu] durch [Manifest], [Manila] oder irgendetwas anderes ersetzen, und der Computer könnte noch mehr neue Witze produzieren. Mit genug Regeln könnte ich vermutlich Hunderte neuer Witze ausgeben lassen.

Je nachdem, wie ausgeklügelt das Programm sein soll, könnte ich sehr viel Zeit mit dem Festlegen neuer Regeln verbringen. Ich könnte eine Liste mit vorhandenen Pointen finden und nach einer Möglichkeit suchen, sie in das nötige Format für die Klopf-klopf-Witze umzuwandeln. Ich könnte sogar versuchen, Ausspracheregeln zu programmieren oder Reime, Homophone oder kulturelle Anspielungen und so weiter. Diese könnte der Computer dann zu neuen interessanten Pointen zusammensetzen. Wenn ich das schlau genug anstelle, kann das Programm Pointen erstellen, die es nie zuvor gehört hat. (Allerdings hat eine Person, die das tatsächlich ausprobiert hat, festgestellt, dass die vom Algorithmus verwendeten Wörter und Phrasen so altmodisch und verworren waren, dass sie heutzutage niemand mehr verstehen würde.) Dennoch, egal wie ausgeklügelt meine Witzeregeln auch sein mögen: Ich gebe dem Computer immer genau Anweisungen dazu, wie er das Problem lösen soll.

KI trainieren

Wenn ich dagegen eine KI darauf trainiere, Klopf-klopf-Witze zu erzählen, lege ich keine Regeln fest. Die KI muss die Regeln selbst herausfinden.

Ich übergebe ihr nur eine Reihe schon bekannter Klopf-klopf-Witze und ein paar absolut notwendige Spielregeln: »Hier hast du ein paar Klopf-klopf-Witze. Versuche, mehr davon herzustellen.« Und das Rohmaterial, mit dem die KI arbeiten soll? Ein Sack voll zufälliger Buchstaben und Satzzeichen.

Danach hole ich mir erst mal einen Kaffee.

Die KI macht sich an die Arbeit.

Zuerst versucht sie, ein paar Buchstaben zu finden, die in Klopf-klopf-Witzen vorkommen. Im Moment sind die Rateversuche dabei komplett zufällig. Der erste Versuch könnte also alles Mögliche sein, zum Beispiel: »qasdnw,m sne?mso d.« Soweit die KI im Moment weiß, wird so ein Klopf-klopf-Witz erzählt.

Dann sieht sich die KI an, was Klopf-klopf-Witze *tatsächlich* sein sollen. Sehr wahrscheinlich liegt sie dabei ziemlich falsch. »Also gut«, sagt sich die KI und passt ihre eigene Struktur ein wenig an, damit sie beim nächsten Mal etwas besser rät. Dabei ist die Stärke der Änderungen begrenzt, um zu verhindern, dass sich die KI jedes neu gefundene Textstück merkt. Aber schon mit minimalen Anpassungen findet die KI heraus, dass sie zumindest nicht vollkommen falsch liegt, wenn sie nur *k*s und Leerzeichen errät. Nachdem sie sich den ersten Stapel Klopf-klopf-Witze angesehen und ein paar Korrekturen vorgenommen hat, ist das nun Folgende ihre Vorstellung davon, wie ein Klopf-klopf-Witz auszusehen hat:*

```
   k k k  k  k
 kk    k kkkok
k kkkk

kk
 kk  k  kk

klokp  k

     k
 k
```

Das ist jetzt nicht gerade der beste Klopf-klopf-Witz der Welt. Aber mit diesem Startpunkt kann die KI mit einem zweiten Stapel Witze weitermachen und dann mit noch einem. Bei jedem Durchgang verfeinert sie ihre Witzeformel weiter, um zu immer besseren Ergebnissen zu gelangen.

* Wir haben die englischen Ausgaben der KI für dieses Buch ins Deutsche übersetzt, damit die Beispiele für deutsche Leser:innen besser nachvollziehbar sind.

Nach ein paar Runden mit Rateversuchen und Anpassungen mehr hat die KI weitere Regeln gelernt. Sie weiß nun, dass am Ende einer Zeile gelegentlich ein Fragezeichen auftauchen muss. Sie lernt, Vokale zu benutzen (insbesondere *o*). Sie versucht sich sogar an der Verwendung von Kommata.

```
noo,
Lnop noo
Kof?
hnos h st
wrst oa , a

asutWen
klo ada
pf kla
w is
e
```

Wie gut stimmen ihre Regeln für die Erzeugung von Klopf-klopf-Witzen mit der Realität überein? Offenbar fehlt noch etwas.

Um einen guten Klopf-klopf-Witz zu erzählen, muss sie herausfinden, in welcher *Reihenfolge* die Buchstaben angeordnet werden müssen. Auch hier beginnt sie mit ein paar Rateversuchen. Folgt auf ein *o* immer ein *w*? Offenbar keine so gute Idee. Aber dann rät sie, dass auf ein *o* oft *pf* folgt. Großartig. Wieder hat die KI einen Fortschritt erzielt. So sieht für sie aktuell der perfekte Witz aus:

```
Weopf
Weopf
Weopf
Weopf
Weopf Weopf Weopf
Weopf Weopf
Weopf
Weopf
```

Das ist immer noch kein Klopf-klopf-Witz – es klingt mehr nach einem Huhn mit einem Sprachfehler. Die KI muss also weitere Regeln finden.

Sie sieht sich die Trainingsdaten erneut an und sucht nach neuen Möglichkeiten, »opf« zu benutzen. Dabei versucht sie, Kombinationen zu finden, die noch besser auf die bestehenden Beispiele für Klopf-klopf-Witze passen.

```
klpf wopf worpf
WKl wWlopf
Kmopf
er is Westa Wler
looo ooop
Keee?
eerr
lop lo,p Wler s rsit
ea lo oo pf KropfWnopf Woors
Dapf
lop K opf
Kop
wee
KKopf Klopf Daopf Weompf
```

Diese Entwicklung dauert nur wenige Minuten. Als ich mit meinem Kaffee zurückkomme, hat die KI schon herausbekommen, dass es sehr, sehr gut zu den Beispielen passt, wenn die Witze mit »Klopf klopf, wer ist da?« beginnen. Sie entscheidet sich, diese Regel *immer* anzuwenden.

Allerdings dauert es etwas länger, den Rest der Formel herauszubekommen. Dabei verfängt sich die KI immer wieder in einer Art Schleife, in der sie mit sich selbst »Stille Post« spielt.

```
Klopf klopf.
Wer ist da?
lien
Austra wer?
Aaani wer?
Alien wer?
Aulali wer?
trali wer?
stilien wer?
```

Schon bald bekommt sie auch den Rest der Formel heraus, allerdings noch ohne richtige Wörter (die KI hat keine Ahnung, was »Wörter« eigentlich sind):

```
Klopf klopf.
Wer ist da?
Wern
Hane wer?
Glarm sind dann denen da
```

Manchmal ist die Pointe auch etwas zu lang:

```
Klopf klopf.
Wer ist da?
Dieelieb
Manipizz wer?
Hingeh den letzt emotiwan aut ehn weonges eich anet ehn tee
   soor dein onithit das wo dein thasg glib tine wenn dasle ter
   den dieung!
```

Irgendwann im Verlauf des Trainings lernt die KI den Witz mit der lippenlosen Kuh, der ab sofort ihr »Absolut Großartigster Lieblingswitz Aller Zeiten« ist. Falls Sie den Witz nicht kennen – hier ist er:

Klopf klopf.
Wer ist da?
Eine lippenlose Kuh.
Welche lippenlose Kuh?
Eine lippenlose Kuh macht uuh, uuuuuh, uuuuh.

Für die KI ist dies aktuell die höchste Form des Humors. Also glaubt die KI zunächst für einige Zeit, dass die Pointe jedes Klopf-klopf-Witzes mit *uuh, uuuuuh, uuuuh* endet.

```
Klopf klopf.
Wer ist da?
Alex
Alim wer?
Wuh ku zuh den do do duhhr zu dem suuuh uuu uuuu uuuh uu uu uu
   uuuuuu uuuh uuuuuh uuu uuu uuh uuuu uuu uu...
```

Okay, die KI hat eine neue Regel gefunden. Nach und nach (und vermutlich etwas widerstrebend) nimmt sie wieder etwas Abstand von der »Kuh ohne Lippen«-Pointe. Sie hat aber weiterhin Probleme, die eine erstaunliche Ähnlichkeit mit zwei Personen haben, die sich Klopf-klopf-Witze über eine schlechte Telefonverbindung erzählen und dabei offenbar vergessen, wer denn jetzt den Witz erzählt:

```
Klopf klopf.
Wer ist da?
```

```
Alter
Wasser wer?
Wasser wer wer?
Wer wer?
Wasser damit macht gerne blau?
```

Schließlich findet die KI die korrekte Struktur eines Klopf-klopf-Witzes. Jetzt erzeugt sie Witze, die technisch zwar funktionieren, aber immer noch wenig Sinn ergeben. Viele der Witze scheinen Plagiate aus den Trainingsdaten zu enthalten:

```
Klopf klopf.
Wer ist da?
Mani
Mani wer?
Raubüberfall! Her mit der Knete!

Klopf klopf.
Wer ist da?
Salat
Salat wer?
Eine lippenlose Kuh kann gar nicht sprechen.

Klopf klopf.
Wer ist da?
Australien
Australien wer?
Aus Tralien Gerfeld.
```

Und dann ... produzierte die KI ... tatsächlich einen Witz ... selbstständig, ohne hierfür auf die Trainingsdaten zurückzugreifen. Der Witz war nicht nur verständlich, sondern sogar ... lustig?

```
Klopf klopf.
Wer ist da?
Mark
Mark wer?
Ich Mark Klopf-klopf-Witze.
```

Hat die KI also tatsächlich begonnen, Klopf-klopf-Witze und Wortspiele zu verstehen? Bei dem relativ kleinen Trainingsdatensatz ist das mehr als unwahrscheinlich. Immerhin hatte die KI vollkommen freien Zugriff auf alle möglichen Zeichen und Buchstaben. Dadurch konnte sie neue Kombinatio-

nen von Lauten ausprobieren, von denen einige am Ende tatsächlich funktionierten. Hier war also eher der sprichwörtliche Affe mit der Schreibmaschine* am Werk. Als Beweis dafür, dass KIs bald ihre eigene Comedy-Sendung bekommen, reicht das aber nicht.

Noch ein paar Beispiele

Das Schöne daran, die KI ihre eigenen Regeln aufstellen zu lassen (nach dem Motto: »Hier sind die Daten. Finde heraus, wie du das Prinzip imitierst.«), ist, dass ein einziger Ansatz für viele verschiedene Probleme eingesetzt werden kann. Hätte ich dem Witze erzählenden Algorithmus andere Trainingsdaten gegeben, hätte er stattdessen gelernt, diesen Datensatz zu imitieren.

Ich könnte die KI beispielsweise neue Namen für Vogelarten erfinden lassen:

```
Yucatan-Kriechente
Bootschnäbliger Sonnenvogel
Brasilianischer Zackenschnäbler
Schwarzmützen-Flauschschwanz
Vierbeiniger Dremelkopf
Isländischer Schnickschnackschnuck
```

Oder Parfums:

```
Fancy Ten
Eau de Pochoir
Fleur de Rhume
Momite
Le Phonême
```

Oder neue Kochrezepte:

```
Einfache Muschelglasur:
Hauptgericht, Suppen

½ Kilo Huhn
½ Kilo Schweinefleisch, gewürfelt
½ Knoblauchzehe, zerstoßen
1 Tasse Sellerie, in Scheiben geschnitten
```

* Das alte Sprichwort von dem Affen mit der Schreibmaschine. Wenn er nur lange genug darauf herumtippt, wird er irgendwann sogar die gesammelten Werke Shakespeares hervorbringen. Diese Analogie beschreibt die »Brute-Force-Methode« recht gut, bei der versucht wird, eine Problemlösung zu finden, indem alle Möglichkeiten durchprobiert werden. Bestenfalls sollte eine KI eine Verbesserung dieser Methode sein. Bestenfalls.

```
1 Kopf (etwa ½ Tasse)
6 Teelöffel elektrischer Mixer
1 Teelöffel schwarzer Pfeffer
1 Zwiebel – gehackt
3 Tassen Rinderbrühe mit Eulen für die Frucht
1 frisch zerstoßenes halb und halb; entspricht Wasser

Mit püriertem Zitronensaft und Zitronenscheiben in
   einer 750ml-Pfanne.
Gemüse hinzufügen, Huhn zur Soße hinzufügen, gut in
   der Zwiebel vermischen. Lorbeerblätter und roten
   Pfeffer hinzufügen, langsam abdecken und abgedeckt
   für drei Stunden köcheln lassen. Kartoffeln und
   Karotten zum Köcheln hinzufügen. Erhitzen, bis die
   Soße kocht. Mit Pasteten servieren.
Wenn die gelausten Teile Nachtisch gekocht haben, und
   über dem Wok kochen.
Für eine ½ Stunde dekoriert kühlen.
Ergibt: 6 Portionen
```

Die KI einfach mal machen lassen

Allein mit einer Reihe von Kopf-klopf-Witzen, aber ohne weitere Anweisungen, konnte die KI eine Menge Regeln finden, die ich ansonsten von Hand hätte programmieren müssen. Einige dieser Regeln wären mir vermutlich niemals eingefallen, zum Beispiel dass die »Kuh ohne Lippen« offenbar der beste Witz der Welt ist.

Unter anderem deshalb gelten KIs als attraktive Problemlöser, besonders wenn die Regeln sehr kompliziert oder regelrecht mysteriös sind. Daher wird KI oft zur Bilderkennung genutzt, einer überraschend schwierigen Aufgabe, die sich nur schwer mit einem einfachen Computerprogramm umsetzen lässt. Für uns Menschen ist es kein Problem, eine Katze in einem Bild zu erkennen. Trotzdem ist es ziemlich schwer, Regeln zu finden, die eine Katze definieren. Sagen wir dem Programm, dass eine Katze zwei Augen, eine Nase, zwei Ohren und einen Schwanz hat? Das würde auch eine Maus oder eine Giraffe beschreiben. Was machen wir, wenn sich die Katze eingerollt hat oder in eine andere Richtung schaut? Allein die Regeln für das Erkennen eines Auges sind schon sehr kompliziert. Eine KI kann sich dagegen Tausende von Katzenbildern ansehen und eigene Regeln entwickeln, nach denen eine Katze fast immer erkannt werden kann.

Pseudo-KI

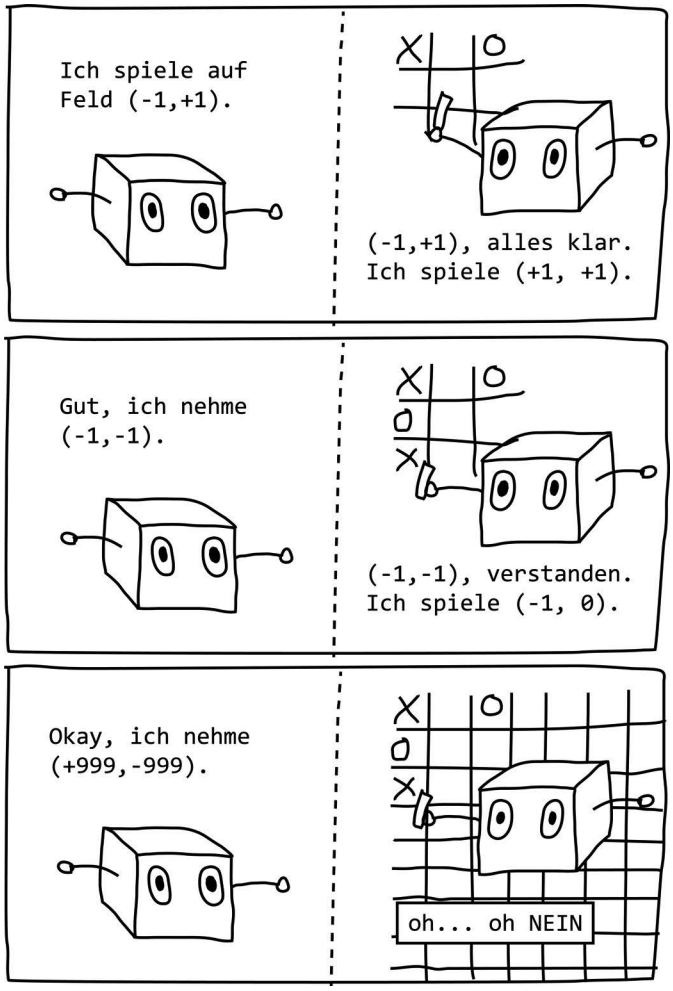
Manchmal ist eine KI nur ein kleiner Teil eines größeren Programms, das hauptsächlich aus regelbasiertem Code besteht. Nehmen wir zum Beispiel ein Programm, über das Kunden telefonisch Informationen über ihre Bankkonten abfragen können. Die Spracherkennungs-KI gleicht den Klang gesprochener Wörter mit den Optionen im Menü der Kundenhotline ab, auf die der Anrufer zugreifen kann. Diese Liste und der Code, der das Konto als zum Kunden gehörig identifiziert, werden jedoch von einem menschlichen Programmierer festgelegt.

Bei anderen Programmen ist die KI das bestimmende Element. Wird die Aufgabe zu knifflig, übergibt sie die Kontrolle an einen Menschen. Dieser Ansatz wird auch als *Pseudo-KI* bezeichnet. Nach diesem Prinzip funktionieren beispielsweise einige Kundendienst-Chatfenster. Wenn Sie ein Gespräch mit einem Bot beginnen und sich zu verwirrend verhalten oder wenn die KI feststellt, dass Sie zu genervt sind, kann es überraschend passieren, dass Sie plötzlich mit einem echten Menschen kommunizieren. (Ein Mensch, der jetzt allerdings mit einem verwirrten und/oder genervten Kunden umgehen muss. Vielleicht wäre hier eine »Gleich mit einem Menschen sprechen«-Option besser für Kunden und Angestellte). Auch die aktuellen selbstfahrenden Autos funktionieren auf diese Weise – der Fahrer muss immer bereit sein, die Kontrolle zu übernehmen, falls die KI einmal durcheinanderkommt.

KI eignet sich außerdem sehr gut für Strategiespiele wie Schach. Wir können zwar alle möglichen Spielzüge beschreiben, aber wir wissen nicht, wie man eine Formel schreibt, um den besten Folgezug zu ermitteln. Selbst Großmeister können keine festen Regeln aufstellen, nach denen in jeder Situation der beste Spielzug bestimmt werden kann, dafür ist das Spiel einfach zu komplex. Ein Algorithmus kann dagegen Millionen von Übungsspielen gegen sich selbst spielen, um Regeln zu finden, die ihm beim Gewinnen helfen. Das wird so oft wiederholt, bis selbst der ehrgeizigste Großmeister nicht mehr mitkommt. Und weil die KI ohne ausdrückliche Anweisungen gelernt hat, sind ihre Strategien manchmal sehr unkonventionell. Manchmal etwas *zu* unkonventionell.

Wenn Sie der KI nicht sagen, welche Züge gültig sind, findet sie möglicherweise einige sehr seltsame Schlupflöcher, die das Spiel komplett zerstören können. 1997 schrieben Programmierer Algorithmen, mit denen zwei Computer über eine Datenverbindung auf einem unendlich großen Spielfeld Tic Tac Toe (auch als »Drei gewinnt« bekannt) gegeneinander spielten. Anstelle einer regelbasierten Strategie verwendete einer der Programmierer

eine KI, die ihre Vorgehensweise selbstständig weiterentwickeln konnte. Die Strategie der KI bestand am Ende darin, den nächsten Spielzug möglichst weit vom letzten entfernt zu platzieren. Wenn der Computer des Gegners versuchte, das neue, extrem erweiterte Spielfeld zu simulieren, war irgendwann nicht mehr genug Arbeitsspeicher vorhanden. Der Computer stürzte ab, und das Spiel war verloren.[1] Die meisten KI-Programmierer kennen solche Geschichten, in denen ihre Algorithmen sie mit unerwarteten Lösungen überraschten. Manchmal waren diese Lösungen genial, manchmal aber auch ziemlich problematisch.

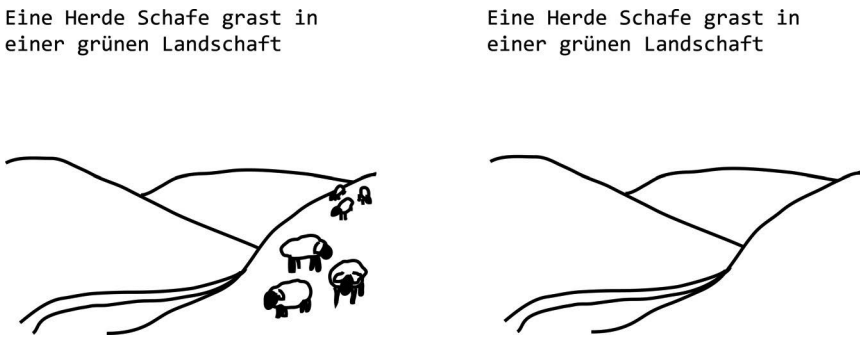


Damit eine KI funktioniert, braucht sie mindestens ein Ziel und einen Satz Trainingsdaten. Mögliche Ziele wären beispielsweise, von Menschen getroffene Kreditentscheidungen zu imitieren oder vorherzusagen, ob ein Kunde eine bestimmte Socke kauft. Vielleicht soll sie den Spielstand in einem Videospiel oder die Entfernung optimieren, die ein Roboter zurücklegen kann. In allen Szenarien geht die KI nach dem Prinzip von Versuch und Irrtum vor, um Regeln zu erfinden, die ihr beim Erreichen ihres Ziels helfen.

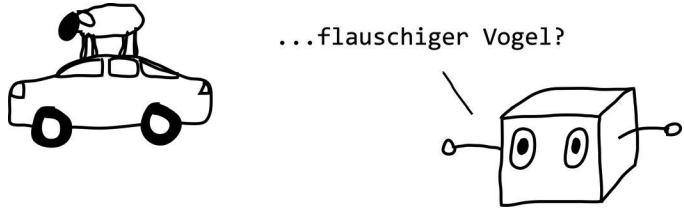
Manchmal sind die Regeln schuld

Gelegentlich basieren die brillanten Regeln, die eine KI zur Problemlösung verwendet, aber auf falschen Annahmen. Für eines meiner seltsamsten KI-Experimente setzte ich ein Microsoft-Produkt zur Bilderkennung ein. Übergibt man der KI ein beliebiges Bild, kann sie es automatisch mit Schlagwörtern und einer Bildunterschrift versehen. Normalerweise bekommt der Algorithmus das auch recht gut hin und kann Wolken, U-Bahn-Züge und sogar ein Kind, das Skateboard-Tricks übt, richtig erkennen. Eines Tages fiel mir an den Ergebnissen aber etwas Merkwürdiges auf: Die KI versah Bilder mit dem Schlagwort *Schaf*, die eindeutig keine Schafe enthielten. Als ich der Sache auf den Grund ging, stellte ich fest, dass die Schafe immer in Bildern gefunden wurden, die auch satte grüne Weiden zeigten, selbst wenn es dort keine Schafe gab.

Woher kam dieser beharrliche und ziemlich spezifische Fehler? Vielleicht hatte die KI beim Training meistens Schafe auf einer Weide gesehen. Dabei muss ihr entgangen sein, dass die Beschriftung »Schaf« sich auf die Tiere, aber nicht auf die Grasslandschaft bezog. Die KI hatte schlicht nach der falschen Sache gesucht. Und richtig: Nachdem ich ihr Bilder von Schafen gezeigt hatte, die sich nicht auf einer grünen Weide befanden, kam die KI durcheinander. Zeigte ich ihr Bilder von Schafen in Autos, erkannte sie stattdessen Hunde oder Katzen. Schafe in Wohnzimmern wurden ebenfalls als Hunde oder Katzen markiert, ebenso wie Schafe, die ein Mensch im Arm hielt. Auch Schafe an der Leine wurden als Hunde eingeordnet. Ähnliche Probleme hatte die KI mit Ziegen. Wenn sie auf Bäume kletterten, wie es manchmal ihre Art ist, hielt der Algorithmus sie für Giraffen (ein ähnlicher Algorithmus war dagegen fest überzeugt, dass die Ziegen Vögel sein müssten).



Ich kann es zwar nicht mit Sicherheit sagen, aber es ist denkbar, dass die KI Regeln entwickelt hat wie *Grünes Grass = Schafe* und *Fell in Autos oder Küchen = Katzen*. Diese Regeln haben beim Training gut funktioniert, scheiterten aber in der wirklichen Welt, die eine verwirrende Vielfalt an Umgebungen bereithält, in denen Schafe sich aufhalten können.



Diese Art von Trainingsfehlern kommt bei KIs für die Bilderkennung recht häufig vor, was manchmal auch recht ernste Konsequenzen haben kann. Ein Forschungsteam der Stanford University trainierte beispielsweise eine KI darauf, zwischen Bildern mit gesunder Haut und solchen mit Hautkrebs zu unterscheiden. Nachdem sie ihre KI trainiert hatten, mussten die Forscher allerdings feststellen, dass sie der KI stattdessen beigebracht hatten, Lineale zu erkennen – in den Trainingsdaten war neben vielen Tumoren ein Lineal als Maßstab zu sehen.[2]

Wie man eine schlechte Regel erkennt

Oft ist es nicht gleich erkennbar wenn KIs Fehler machen. Da wir ihnen keine Regeln vorgeben, entwickeln sie ihre eigenen, ohne sie aufzuschreiben oder sie wie Menschen zu erklären. Stattdessen nehmen KIs komplexe

voneinander abhängige Anpassungen an ihren internen Strukturen vor und verwandeln ein allgemeines Grundgerüst in eine fein auf die zu bewältigende Aufgabe abgestimmte Form. Das ist, als würden sie mit einer Küche voller allgemeiner Zutaten beginnen, um am Ende Kekse zu erhalten. Die Regeln sind dabei in den Verbindungen zwischen den virtuellen Gehirnzellen oder den Genen eines virtuellen Organismus gespeichert. Dabei können die Regeln komplex sein, weit verteilt und seltsam ineinander verschlungen. Das Studium der internen Struktur einer KI hat oft große Ähnlichkeit mit dem Studium eines Gehirns oder eines Ökosystems. Auch ohne Neurologie oder Ökologie studiert zu haben werden Sie sich denken können, wie komplex so etwas sein kann.

Obwohl Forscher herauszufinden versuchen, wie KIs ihre Entscheidungen treffen, ist es im Allgemeinen sehr schwer, die internen Regeln einer KI genau zu bestimmen. Das liegt oft einfach daran, dass die Regeln schwer zu verstehen sind. In anderen Situationen, besonders bei Algorithmen für Unternehmen oder Regierungen, ist der Algorithmus grundsätzlich geheim. Dadurch werden Probleme mit den Ergebnissen des Algorithmus oft erst erkannt, wenn dieser bereits im Einsatz ist. In der Folge werden möglicherweise Entscheidungen getroffen, die Auswirkungen auf Menschenleben haben und echten Schaden anrichten können.

Eine KI wurde beispielsweise eingesetzt, um Empfehlungen auszusprechen, welche Gefängnisinsassen begnadigt werden sollen. Bald stellte man fest, dass die KI offenbar voreingenommen war, weil sie, ohne es zu wissen, die rassistischen Vorurteile kopierte, die in den Trainingsdaten enthalten waren.[3] Eine KI kann also befangen sein, ohne zu verstehen, was Vorurteile eigentlich sind. KIs lernen, indem sie Menschen imitieren. Sie beantworten also nicht die Frage: »Was ist die beste Lösung?«, sondern »Was hätte ein Mensch getan?«.

Wenn systematisch auf Befangenheit getestet wird, können einige dieser verbreiteten Probleme erkannt werden, bevor ein Schaden entsteht. Ein weiterer wichtiger Baustein besteht darin, Probleme schon vor ihrem Auftreten vorherzusehen und die KIs so zu entwickeln, dass sie von vornherein vermieden werden.

Vier Anzeichen für KI-Katastrophen

Bei einer KI-Katastrophe denkt man daran, dass eine KI Befehle verweigert und entscheidet, dass es in ihrem besten Interesse ist, alle Menschen zu töten oder Terminator-Bots zu erschaffen. Diese Katastrophenszenarien basieren aber auf der Annahme, dass KIs ein bestimmtes Niveau an kritischem Denken und ein menschengleiches Weltverständnis besitzen, das sie

in der vorhersehbaren Zukunft aber nicht erreichen werden. Einer der führenden Forscher für maschinelles Lernen, Andrew Ng, hat einmal gesagt, die Angst vor einer Machtübernahme der KI ist wie die Sorge vor einer Überbevölkerung auf dem Mars.[4]

Deshalb sind die heutigen künstlichen Intelligenzen aber nicht unbedingt harmlos. Das Spektrum möglicher Probleme reicht von leichter Genervtheit der Programmierer bis zur Übernahme von Vorurteilen, oder Unfällen mit selbstfahrenden Autos. Aber mit etwas Wissen über KIs können wir uns auf einige dieser Probleme zumindest vorbereiten.

Hier ein Beispiel dafür, wie eine KI-Katastrophe heutzutage aussehen könnte:

Ein Silicon-Valley-Startup bietet anderen Unternehmen an, aus einer Reihe von Bewerbern die am besten geeigneten Kandidaten für einen Job herauszufinden, indem es kurze Video-Interviews analysiert. Das könnte attraktiv sein, denn Unternehmen wenden eine Menge Zeit und Ressourcen dafür auf, mit Dutzenden von Bewerbern Gespräche zu führen, um möglichst passende Kandidaten für eine Position zu finden. Software wird niemals müde, hungrig oder schlecht gelaunt und sie kann alle Kandidaten gleich gut leiden.

Warnsignal 1: Das Problem ist zu kompliziert

Selbst Menschen haben Probleme, die richtigen Kandidaten für einen Job zu finden. Ist der Bewerber wirklich davon begeistert, hier zu arbeiten oder ist er nur ein guter Schauspieler? Haben wir Handicaps oder kulturelle Unterschiede ausreichend berücksichtigt? Wenn wir eine KI hinzuziehen, wird die Sache sogar noch schwieriger. Für eine KI ist es fast unmöglich, die verschiedenen Nuancen bestimmter Witze, den Tonfall, oder bestimmte kulturelle Anspielungen richtig zu verstehen. Wie soll sie sich verhalten, wenn ein Bewerber sich auf tagesaktuelle Ereignisse bezieht? Wurde die KI mit Daten aus dem vergangenen Jahr trainiert, hat sie keine Chance, diese Anspielungen zu verstehen. Im Ergebnis könnte Sie den Kandidaten schlechter bewerten, weil die Äußerungen für sie keinen Sinn ergeben. Kann die KI ihre Aufgabe nicht gut erledigen, wird sie früher oder später daran scheitern.

Warnsignal 2: Es geht eigentlich um ein anderes Problem

Bei der Entwicklung einer KI zur Bewertung von Bewerbungen können wir die KI nicht einfach anweisen, die besten Kandidaten zu finden. Vielmehr versucht die KI, Personen zu ermitteln, die am ehesten den bisherigen Vorlieben der Personalchefs entsprechen.

Das mag funktionieren, wenn die Personalchefs zuvor gute Entscheidungen getroffen haben. Die meisten US-Unternehmen haben jedoch ein Diversitätsproblem. Das gilt besonders für Manager und für die Art, in der Personalverantwortliche Lebensläufe und Kandidaten für Bewerbungsgespräche bewerten. Wenn alle Kriterien gleich sind, ist es deutlich wahrscheinlicher, dass Personen mit Namen die weiß und männlich klingen zu einem Bewerbungsgespräch eingeladen werden, als solche deren Namen weiblich oder nach Minderheiten klingen.[5] Selbst Personalchefs, die selbst weiblich und/oder Mitglieder von Minderheiten sind, haben eine unbewusste Tendenz, weiße männliche Kandidaten zu bevorzugen.

Viele schlechte und/oder regelrecht gefährliche KI-Programme werden von Leuten entwickelt, die dachten, dass ihre KI ein bestimmtes Problem löst. In Wirklichkeit haben sie die KI aber auf etwas vollkommen anderes trainiert.

Warnsignal 3: Versteckte Abkürzungen

Erinnern Sie sich noch an die Hautkrebs-KI, die in Wirklichkeit ein Linealdetektor war? Der Unterschied zwischen gesunden Zellen und Krebszellen ist schwer zu erkennen. Für die KI war es viel einfacher, ein Lineal im Bild als Unterscheidungsmerkmal zu finden.

Übergeben Sie einer KI, die den besten Bewerber finden soll, unausgewogene Trainingsdaten (was Sie ziemlich sicher der Fall ist, es sei denn Sie haben einen ziemlich großen Aufwand getrieben, um die Befangenheit aus den Daten zu entfernen), dann bieten Sie ihr auch eine bequeme Abkürzung, mit der sie ihre Genauigkeit bei der Vorhersage der »besten« Kandidaten verbessern, nämlich weiße Männer zu bevorzugen. Das ist viel einfacher, als alle Feinheiten in der Wortwahl eines Kandidaten zu berücksichtigen. Vielleicht findet die KI auch eine komplett andere Abkürzung und nutzt diese aus – vielleicht haben wir die erfolgreichen Kandidaten mit einer bestimmten Kamera gefilmt und die KI benutzt nun die Metadaten, um nur Kandidaten die damit gefilmt wurden, auszuwählen.

KIs nehmen andauernd versteckte Abkürzungen – sie wissen es einfach nicht besser!

Warnsignal 4: Die KI versucht, aus fehlerhaften Daten zu lernen

In der Informatik gibt es ein altes Sprichwort: Wenn du Müll reintust, kommt auch nur Müll wieder raus (»garbage in, garbage out«). Wenn das Ziel der KI darin besteht, Menschen zu imitieren, die fehlerhafte Entschei-

dungen treffen, ist es ihr größter Erfolg, diese Entscheidungen möglichst genau zu nachzuahmen – inklusive aller enthaltenen Fehler.

Unabhängig davon, ob von fehlerhaften Beispielen gelernt wird oder in einer fehlerhaften Simulation mit seltsamen physikalischen Eigenschaften: Fehlerhafte Daten können eine KI schnell in eine Endlosschleife oder in die falsche Richtung schicken. Da das zu lösende Problem oftmals *unsere Beispieldaten sind*, verwundert es nicht, das schlechte Daten zu schlechten Lösungen führen. Tatsächlich sind die Warnsignale 1 bis 3 in den meisten Fällen ein Beweis für Probleme mit den Daten.

Fluch oder Segen?

Das Beispiel mit der Bewerberauswahl ist leider real. Viele Unternehmen bieten bereits eine KI-gestützte Bewerberauswahl anhand von Lebensläufen oder Video-Interviews an. Nur wenige geben dabei Informationen darüber preis, wie sie mit Befangenheit, Handicaps oder kulturellen Unterschieden umgehen oder wie sie herausfinden, welche Informationen ihre KIs für die Auswahl tatsächlich verwenden. Durch sorgfältiges Arbeiten sollte es zumindest möglich sein, eine KI für die Bewerberauswahl zu schaffen, die neutraler ist als menschliche Personalchefs. Ohne veröffentlichte Statistiken zum Beweis können wir aber ziemlich sicher sein, dass die Vorurteile auch weiter existieren.

Der Unterschied zwischen erfolgreichen KI-basierten Problemlösungen und ihrem Scheitern hat eine Menge damit zu tun, wie gut eine KI für die Aufgabe tatsächlich geeignet ist. Tatsächlich gibt es viele Bereiche, in denen KI-Lösungen effizienter sind als menschliche. Welche das sind und warum sie für bestimmte Aufgaben besser geeignet sind als andere, sehen wir uns im folgenden Kapitel an.