Künstliche Intelligenz

Wie sie funktioniert und was sie für uns bedeutet

DIE LESEPROBE

Kapitel 2
Wie wir versuchen, Maschinen intelligent zu machen

2.1 Symbolische KI

Vielleicht erinnerst du dich noch an unsere vier jungen Forscher, die diesen Fachbereich begründeten? Ihr Grundgedanke »Jeglicher Aspekt von Intelligenz oder Lernen kann im Prinzip so genau ausgedrückt werden, dass ein Computer ihn simulieren kann« stand bis Mitte der 1990er Jahre prägnant im Mittelpunkt, und KI-Forscher*innen waren sich im Allgemeinen darüber einig, dass der Weg zu intelligenten Computern daraus bestand, ihnen ganz genau zu erzählen, was sie tun sollten. Und das geschah mithilfe von Symbolen.

Symbole sind das, was wir Menschen benutzen, um Dinge zu repräsentieren, und sie spielen eine ganz zentrale Rolle in unseren Gedankenprozessen. Wenn ich sage: »Ich sah eine Katze auf den Baum klettern«, produziert dein Gehirn augenblicklich ein Bild dieser Situation. Das kann es, weil wir uns einig sind bezüglich der Symbole, die die Objekte *Katze* und *Baum* repräsentieren. Dein Gehirn ist auch in der Lage, abstrakte Symbole zu verwenden, also Worte, die nichtphysische Dinge beschreiben, wie Bankkonto und Blogbeitrag, und Symbole, die Eigenschaften beschreiben, wie *schnell, langweilig* und *zerzaust*. Die Fähigkeit, mithilfe von Symbolen zu kommunizieren, ist etwas, das uns Menschen intelligent und unsere Kommunikation effektiv macht. Deshalb spielten Symbole auch eine zentrale Rolle in der künstlichen Intelligenz. *Symbolische KI* ist ein Ansatz für künstliche Intelligenz, die darauf basiert, für den Computer Symbole zu definieren und explizite Regeln für ihn aufzustellen.

Die allereinfachste Art und Weise, dies zu tun, geschieht durch Tabellen. Und ich kann fast garantieren, dass wir alle in unserem Leben schon mal großen Nutzen aus einem tabellenbasierten Programm gezogen haben, denn digitale Wörterbücher sind genau das. Übersetzungsprogramme oder Websites wie Google Übersetzer können schnell zwischen vielen Sprachen übersetzen, weil ihnen Wörter und Übersetzungen in einer riesigen Tabelle zugänglich sind. Ich stimme der Ansicht zu, dass so ein Vorgehen nicht gerade das Erlebnis von maschineller *Intelligenz* erzeugt, aber es ist nicht zu leugnen, dass Computer schlauer darin sind, sich etwas zu merken, und schneller

etwas nachschlagen können als Menschen. Deshalb kann ein Mensch niemals Wörter zwischen verschiedenen Sprachen so schnell oder in so einem Umfang übersetzen wie ein Rechner. Obwohl also eine Tabelle in einem Computer nicht intelligent *ist*, kann sie doch dafür *genutzt* werden, Probleme auf eine intelligente Art und Weise zu lösen. Tabellenbasierte Programme sind einfach zu erstellen, es ist leicht, die Übersicht über sie zu behalten und – vielleicht das Allerwichtigste – sie sind zuverlässig. Wenn in der Tabelle steht, dass *Katze* auf Englisch *cat* heißt, wird das Programm niemals kreativ werden und *dog* (*Hund*) vorschlagen.

Ein Schritt von den Tabellen weg, in denen man etwas nachschlagen kann, befinden sich *Regeln*. Damit sich jemand, beispielsweise ein Computer, richtig aufführen kann, muss er die in einem System geltenden Regeln kennen und benutzen. Wenn wir Computern oder allgemeiner Maschinen die in einem System geltenden Regeln beibringen, sagen wir, dass wir ein regelbasiertes System aufbauen (*rule-based system* auf Englisch). Wenn diese Regeln auf menschlicher Expertise basieren, also dadurch erstellt wurden, dass Experten interviewt wurden und deren Kenntnisse in codierte Regeln übersetzt wurden, haben wir ein *Expertensystem*. Diese Bezeichnung bedeutet nicht, dass wir deshalb den Computer als einen Experten ansehen, sie ergibt sich daraus, dass die Regeln von menschlichen Expertisen abgeleitet wurden. Wenn mein Auto stottert, muss ich, die ich keine Expertin an der Autofront bin, einen Autoexperten anrufen, etwa eine Kfz-Mechanikerin. Sie wird mir Fragen stellen wie: »Klopft es, wenn Sie bremsen?«, und abhängig davon, was ich antworte, wissen, welche weiteren Fragen sie mir stellen muss. Frage für Frage findet sie heraus, was die Ursache des Problems sein kann und was ich tun müsste – und genauso funktioniert Expertenwissen. Wenn die Rolle der Mechanikerin darauf beschränkt wäre, die richtigen Ja-/Nein-Fragen zu stellen, könnten wir sie durch einen Computer ersetzen – ein Expertensystem –, solange wir davon ausgehen, dass es ihr und einem KI-Entwickler gelungen wäre, alle Fragen, die sich stellen könnten, und die möglichen Antworten darauf zu formulieren.

2.2 Expertensysteme

Die Expertensysteme waren lange unsere größte Hoffnung hinsichtlich künstlicher Intelligenz, und in den 1960er und 1970er Jahren waren KI-Forscher in hohem Grade davon überzeugt, dass dieser symbolische Ansatz uns letztendlich Computer mit *allgemeiner Intelligenz* bescheren würden. Ein Wesen mit allgemeiner Intelligenz ist nicht darauf beschränkt, spezifische Aufgaben zu erfüllen, es kann jedes beliebige Problem lösen, auf das es stößt. Der Titel der Publikation »GPS: a program that simu-

lates human thought« (GPS: Ein Programm, das menschliches Denken simuliert) von 1963 bezeugt das Ambitionsniveau. GPS ist die Abkürzung für *General Problem Solver* (nicht für das globale Positionierungssystem, das seit 1993 im Einsatz ist) und der Name für ein Programm, das auf symbolischer KI basiert. GPS wurde von Herbert Simon und Allen Newell (zwei der »großen Vier« innerhalb der KI) entwickelt, und es konnte alle Probleme lösen, die genau definierte Regeln haben. Ein Beispiel für ein derartiges Problem ist dieses klassische Rätsel:

> *Ein Bauer will einen Fluss überqueren. Er hat ein Kaninchen, Karotten und einen Fuchs bei sich. Doch er hat nur ein kleines Boot und kann jeweils nur ein Objekt mit über den Fluss nehmen. Wie soll er Kaninchen, Karotten und den Fuchs so hinüberbringen, dass das Kaninchen nicht die Karotten und der Fuchs nicht das Kaninchen frisst?*

Den meisten Menschen gelingt es, diese Art von Rätsel zu lösen, wenn sie nur ein wenig darüber nachdenken, und weil das Rätsel alle Informationen enthält, die notwendig sind, um das Problem zu lösen, kann es mithilfe symbolischer KI gelöst werden. Denn damit GPS das Rätsel löst, muss es vorher die relevanten Symbole kennen, die da sind: *Kaninchen, Karotten, Fuchs, Boot, rechtes Flussufer, linkes Flussufer.*[1]

Anschließend muss es wissen, dass die Regeln lauten:

- Kaninchen frisst Karotten.
- Fuchs frisst Kaninchen.
- Boot kann transportieren (Kaninchen, Karotten, Fuchs) vom linken zum rechten Flussufer und zurück.
- Boot kann nur ein Objekt nach dem anderen transportieren.

Schließlich muss GPS wissen, in welchem Zustand sich die im Problem beschriebene Welt befindet, und welcher Zustand erreicht werden soll, nämlich:

> *Ausgangszustand:*
> *Linkes Flussufer = (Kaninchen, Karotten, Fuchs, Boot)*
> *Rechtes Flussufer = ( )*
>
> *Endzustand:*
> *Linkes Flussufer = ( )*
> *Rechtes Flussufer = (Kaninchen, Karotten, Fuchs, Boot)*

1 Dank Grace Hoppers Erfindung des Compilers aus dem Jahr 1952 können wir Wörter schreiben und müssen nicht mehr alle Wörter in Nullen und Einsen umformen.

Mit diesen Regeln, übersetzt in eine Programmiersprache, kann GPS die Aufgabe lösen.[2] Und das ist der grundlegende Gedanke innerhalb der symbolischen KI: Wenn wir das Problem und alle Regeln einem Computer erklären, kann dieser das Problem ebenso gut wie ein Mensch lösen.

Ein wichtiger Unterschied zwischen einem Menschen, der das Rätsel löst, und einem Computer, der das macht, besteht darin, dass weder Fuchs noch Kaninchen noch Karotten einem Computer irgendetwas bedeuten. Für uns Menschen haben diese Worte einen Sinninhalt; sie symbolisieren etwas in der realen Welt. Einen Computer interessiert das nicht, man hätte für ihn auch »XY4« statt Kaninchen schreiben können oder »XY4 frisst QQ5«. Die Problemlösung wäre für die Maschine die gleiche. Der springende Punkt ist, dass die Person, die sie programmiert, versteht, was die Symbole bedeuten und dass die geltenden Regeln für die Symbole für uns einen Sinn ergeben.

Solang es uns gelingt, unsere Probleme präzise zu beschreiben, kann ein Computer sie lösen. Das hat eine unglaubliche Kraft und birgt ein Potenzial, dass die ersten KI-Forscher ausnutzen wollten. Dennoch ist es der symbolischen KI nicht gelungen, generelle Intelligenz zu schaffen, aus dem irritierenden Grund, dass sie davon abhängig ist, dass wir Menschen die Symbole und Regeln für sie definieren. Und selbst wenn wir Menschen selbst Wissen und Intelligenz *haben*, bedeutet das nicht, dass wir in der Lage sind, dieses Wissen zu erklären oder diese Intelligenz mithilfe von Symbolen Computern beizubringen. Um zu illustrieren, wie beschwerlich diese Herausforderung ist, können wir mal eben ein Expertensystem erstellen, das Klopf-Klopf-Witze erzählt. Das System soll drei Witze erzählen können, und es soll *selbst verstehen*, um welchen Klopf-Klopf-Witz es sich handelt. Dabei müssen wir uns vorher darüber einig sein, welche drei Witze wir ihm beibringen wollen, und ich schlage folgende vor:

Klopf, klopf!
Wer ist da?
Daisy
Daisy, wer?
Daisy me rollin', they hatin'

Klopf, klopf!
Wer ist da?

2 Der Trick, das Rätsel zu lösen, besteht darin, das Kaninchen über den Fluss hin und her mitzunehmen. Also: Nimm das Kaninchen mit zum rechten Flussufer, fahr zurück zum linken. Nimm den Fuchs oder die Karotten mit zum rechten Flussufer, nimm das Kaninchen mit zurück zum linken. Nimm den Fuchs oder die Karotten mit zum rechten Flussufer und hole zum Schluss das Kaninchen.

Maja.
Maja, wer?
Maja-hi, Maja-ha, Maja-ho, Maja-haha.

Klopf, klopf!
Wer ist da?
Luke.
Luke, wer?
Luke doch mal nach draußen, dann weißt du es.

Der Grund dafür, dass es einfach sein wird, ein Expertensystem zu schaffen, um diese unglaublich lustigen Witze zu erzählen, liegt darin, dass sie einem klaren Muster folgen. Zuerst muss der Computer sagen: »Klopf, klopf!«, und darauf warten, dass der Nutzer antwortet: »Wer ist da?« Anschließend muss der Rechner zwischen drei verschiedenen Antworten, die jeweils eine für sie zutreffende Pointe haben, eine aussuchen. Daraus folgt ein Programm, das in der Programmierungssprache *Python* geschrieben wird, das den Computer dazu bringt, den Witz zu erzählen.

```
import random
print(»Klopf, klopf!«)
antwort1 = input()
eroeffnungen = [»Daisy«, »Maja«, »Luke«]
eroeffnung = random.choice(eroeffnungen)
print (eroeffnung)
if eroeffnung == »Daisy«:
   punchline = »Daisy me rollin', they hatin'«
elif eroeffnung == »Maja«:
   punchline = »Maja-hi, Maja-ha, Maja-ho, Maja-haha«
elif eroeffnung == »Luke«:
   punchline = »Luke doch mal nach draußen, dann weißt du es.«
else:
   punchline = »Tut mir leid, diesen Witz kenne ich nicht.«
antwort2 = input()
print(punchline)
```

Wenn du noch nie zuvor Programmcode gesehen hast, wirst du vielleicht als Erstes denken: »Mein Gott, wie hässlich!«, aber so sehen Programmcodes nun einmal aus. Die Vorgehensweise, bei der wir überprüfen, ob etwas mithilfe sogenannter `if`-Sätze geschehen ist, ist eine grundlegende Programmierungstechnik. Das merkwürdige Wort `elif` ist eine Abkürzung von `else if`, und Zeile für Zeile haben wir dem Computer gesagt:

```
Nutze deinen Zufallsgenerator
Sage »Klopf, klopf!«
Warte auf Antwort
Die möglichen Eröffnungen sind »Daisy«, »Maja« und »Luke«
Wähle eine zufällige Eröffnung
Sage die Eröffnung
Wenn die Eröffnung »Daisy« ist, ist die Pointe »Daisy me rollin', they hatin'«
Wenn die Eröffnung »Maja« ist, ist die Pointe »Maja-hi, Maja-ha, Maja-ho, Maja
haha«
Wenn die Eröffnung »Luke« ist, ist die Pointe »Luke doch mal nach draußen,
dann weißt du es.«
Wenn die Eröffnung etwas anderes, ganz gleich, was, ist,
musst du sagen: »Tut mir leid, diesen Witz kenne ich nicht.«
Warte auf Antwort
Sage Pointe
```

Dir sind sicher die klaren Grenzen dieses Systems aufgefallen, dass es nämlich nicht überprüft, was der Nutzer antwortet. Und hätten wir diesen KI-Spaßvogel aufs Internet losgelassen, garantiere ich dir, dass die Leute als Allererstes versucht hätten, herauszufinden, wie man ihn hereinlegen oder verwirren kann. Deshalb hätte ein robustes Expertensystem notwendigerweise alle möglichen Überprüfungen und Sicherheitsmechanismen beinhalten müssen. Und all das hätten wir per Hand hineinschreiben müssen: »Wenn der Nutzer nicht richtig antwortet, dann …« Und wir hätten uns gute Alternativen einfallen lassen müssen, um jede einzelne Eventualität zu bewältigen. Und das ist der größte Schwachpunkt des Expertensystems: Ihnen muss ganz genau gesagt werden, was sie in jeder erdenklichen Situation tun sollen.

Dass Expertensysteme zu erstellen etwas nervig ist, bedeutet *nicht*, dass symbolische KI plötzlich aufhört zu funktionieren; Expertensysteme waren und sind weiterhin nützlich und nicht zuletzt sicher im Gebrauch. Und zwar gerade deshalb, weil wir Menschen genau bestimmen, welchen Regeln sie folgen sollen, werden sie nicht auf eigene Faust neue Regeln erfinden. Und auch wenn den Expertensystemen innerhalb der KI-Forschung und auf den Titelseiten der Zeitungen heutzutage nicht die größte Aufmerksamkeit zuteilwird, sind sie sowohl faszinierend als auch stark und werden in großem Umfang benutzt. Das Faszinierende an ihnen ist, dass Entwickler nur die Regeln aufstellen müssen und das System dann basierend auf diesen Regeln willkürlich komplizierte Entschlüsse treffen kann. Wenn das System Tausende von Regeln enthält, kann das darin münden, dass es sehr viel anspruchsvollere Entschlüsse trifft als irgendein Mensch. Auch heute noch sind die meisten der genutzten KI-Systeme Expertensysteme: Die NASA hat viele Jahrzehnte lang Expertensysteme unter ande-

rem zur Auftragskontrolle benutzt, um die Raumfahrttelemetrie und Motorfunktionen zu überwachen.[3] Jedes Mal, wenn du dich in ein Flugzeug setzt, legst du dein Leben in die Hände eines Expertensystems. Unsere Krankenhäuser und Banken sind mit ihnen bevölkert, und tatsächlich sind Expertensysteme inzwischen so üblich und weitverbreitet, dass wir gar nicht mehr daran denken, dass sie da sind, oder sie gar nicht mehr als künstliche Intelligenz wahrnehmen. Jetzt sind es »nur noch« Computerprogramme. Dennoch möchte ich wetten, dass du von mindestens einem ganz bestimmten Expertensystem irritiert warst, zumindest, wenn du so alt bist wie ich.

Erinnerst du dich noch an die Büroklammer, die Mitte der 1990er Teil von Microsoft Word war? Sie hieß Karl Klammer (Clippy) und war ein Expertensystem erster Güte: Karl beherrschte jede Menge grammatischer Regeln und versuchte sich nach besten Möglichkeiten für alle nützlich zu machen, die ein Word-Dokument schreiben wollten. Microsofts Entwickler hatten ihm auch beigebracht, dass alle, die ein Dokument mit »Lieber jemand« beginnen, das Angebot erhalten sollten, ihnen bei der Formatierung des Briefes zu helfen. Karl wurde schnell zum verhasstesten virtuellen Assistenten der Welt. Es ist schwer zu sagen, warum eigentlich, aber vielleicht ist der Grund, dass es ihm nicht gelang, sein Verhalten anzupassen: Selbst wenn man jedes Mal seine Hilfe ablehnte, sobald er fragte, verstand er diesen Wink nicht und fragte bei nächster Gelegenheit genauso enthusiastisch nach. Seine riesigen creepy Augen, mit denen er den Text scannte, ließen ihn außerdem wie eine Art Spion privater Dokumente erscheinen. Microsoft war mit virtuellen Assistenten sicher der Zeit voraus, und Karl Klammers Tage neigten sich in den frühen 2000ern dem Ende. Damals erklärte Microsoft, dass das neue Windows XP so intuitiv sei, dass es eines virtuellen Assistenten nicht mehr bedürfe. Trotz allem nahmen sie es mit Humor, dass Karl so verhasst war, und brachten ein kleines Spiel heraus, bei dem er mit einem Tacker abgeschossen werden konnte. Ich selbst habe dieses Spiel nie gespielt, in erster Linie, um dem Zorn virtueller Assistenten zu entgehen, falls sie eines Tages mit ihren eigenen Schusswaffen zurückkehren.

Karl Klammer ist in keiner Weise ein Beispiel dafür, wie gut Expertensysteme sein können. Eigentlich bin ich vielmehr der Meinung, dass Karl eher illustriert, wie schlecht es laufen kann, wenn wir Expertensysteme an Stellen platzieren, wo sie nicht hinpassen, und das ohne die Fähigkeit, sich anpassen zu können. Wenn es ein angenehmes Erlebnis für uns Menschen sein soll, mit einem Computer zu kommunizieren, dürfen wir nicht das Gefühl haben, dass dieser herumschnüffelt, uns über die

3 Muratore, John F. et al.: Space Shuttle telemetry monitoring by expert systems in mission control, 1989.
https://ntrs.nasa.gov/citations/19900045447

Schulter guckt und sich nicht unseren Wünschen beugt. Im Gegenteil, wir müssen das Gefühl haben, dass es uns hilft, dass der Computer da ist, und wir müssen das Gefühl haben, dass der Computer *uns versteht*.

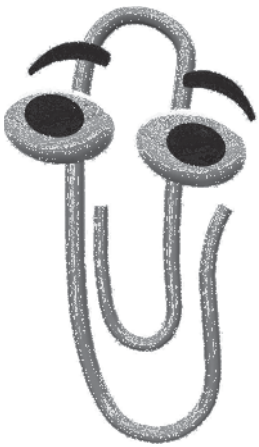

Abbildung 2.1 Der hübsche und doch so verhasste kleine Karl Klammer

2.3 Der ELIZA-Effekt

Der schnellste Weg, einen Gesprächspartner davon zu überzeugen, dass du dich für ihn interessierst und ihn verstehst, verläuft über immer neue Nachfragen. Das zeigt sich immer wieder in sozialen Zusammenhängen und wird auch in der personenzentrierten Psychotherapie angewendet. Diese Therapieform wurde von dem Psychologen Carl Rogers entwickelt, der der Meinung war, dass der Patient selbst weiß, was für ihn am besten ist. Und dieser personenzentrierte Ansatz ist so effektiv, dass er sogar beim ersten Chatbot der Welt funktionierte.

1964 entwickelte der Informatiker Joseph Weizenbaum einen Chatbot, der den Stil eines personenzentrierten Psychotherapeuten imitierte. Im Laufe eines Gesprächs sucht dieser Chatbot, der ELIZA heißt, nach spezifischen Wörtern in den Sätzen, die man ihm serviert. Wörter wie »deprimiert« oder »traurig« sind interessant für ELIZA, aber auch Wörter wie »Mutter«, denn sie stehen auf einer Liste von Wörtern, für die Weizenbaum passende Antworten definiert hat. Wenn ELIZA so ein Wort findet, wird sie selbstständig eine Antwort zusammensetzen. Beispielsweise kann ein Mensch sagen: »Meine Mutter macht gute Pizza«, ELIZA entdeckt das Wort »Mutter« und antwortet: »Erzähl mir mehr über deine Familie.« Wenn du Lust hast, ein Gespräch mit der besten KI zu führen, die anno 1964 verfügbar war, musst du nur im Internet unter »elizabot« suchen, dann findest du sie. Und so erging es mir, als ich mich mit ELIZA

über mein Verhältnis zu intelligenten Computern unterhalten habe (das Originalgespräch verlief auf Englisch):

ELIZA: Worüber möchtest du in dieser Sitzung sprechen?
Inga: Über mein obsessives Interesse an künstlicher Intelligenz.
ELIZA: Ich bin mir nicht sicher, ob ich verstehe, was du meinst.
Inga: Ich bin wie besessen von intelligenten Computern.
ELIZA: Gefällt es dir, wie besessen von intelligenten Computern zu sein?
Inga: Ich glaube schon!
ELIZA: Zweifelst du an dem, was du glaubst?
Inga: Vielleicht manchmal. Wenn es schwierig ist.
ELIZA: Hast du schon mal andere gefragt?

Es brauchte nicht mehr als drei, vier Dialoge mit ELIZA, bis ich ihres wiederholenden Stils überdrüssig war. ELIZAs offene Fragen vermitteln die Illusion, man spräche mit einem Wesen, das versteht, was man sagt, und neugierig ist auf das, was du zu erzählen hast. Kombiniert mit einem Kniff aus personenzentrierten Therapieformen, nämlich die Frage an den Gesprächspartner (bzw. die Patientin) zurückzugeben, wird der Eindruck erweckt, dass ELIZA uns damit zur Reflexion auffordert. In den 1960er Jahren soll Weizenbaum bei mehreren Gelegenheiten entdeckt haben, dass seine Sekretärin Gespräche führte, in denen sie ELIZA von ihrem Privat- und Gefühlsleben erzählte.[4] Die Gerüchte besagen sogar, dass die Sekretärin ab und zu um eine Einzelsitzung mit ELIZA bat. Was diese Sekretärin höchstwahrscheinlich zum ersten Menschen auf der Welt macht, der im Selbstbericht einen positiven Effekt der KI-Anthropomorphisierung feststellte, also dass wir Menschen durch unsere eigene Vorstellungskraft Computern Menschlichkeit zuschreiben. Wenn wir Menschen eine Meinung in einen computergenerierten Text hineindeuten, wird das noch heute als »ELIZA-Effekt« bezeichnet.

Eine der größten Begrenzungen des Computerprogramms ELIZA besteht darin, dass es nicht durch Gespräche neue Wörter oder Gesprächsmuster lernen kann; soll ELIZA etwas Neues lernen, muss das manuell programmiert werden. Und wenn ELIZA einen Satz entgegennimmt, in dem sie nicht ein einziges interessantes Wort findet, antwortet sie generisch, beispielsweise mit »Was meinst du damit?« oder »Warum fragst du das?« und hofft, dass der nächste Satz mehr psychologische Häppchen enthält. Eine noch lustigere ELIZA-Geschichte als die mit der Sekretärin erzählte Daniel Bobrow im Dezember 1968. Sie handelt vom Vizepräsidenten (VP) einer amerikanischen Firma,

4 Hines, Dwight: Review [Untitled] in The Journal of Mind and Behavior, Vol. 1, No. 1 (1980), S. 120–124.
https://www.jstor.org/stable/43852815

der ein Gespräch mit ELIZA führte, in dem Glauben, dass er eigentlich mit einem Entwickler spräche. Das Gespräch verlief folgendermaßen:

VP: Glauben Sie, ich könnte den Computer heute Morgen benutzen?
ELIZA: Warum fragen Sie?
VP: Ich könnte vielleicht ein paar extra Telekom-Verkäufe tätigen.
ELIZA: Warum sind Sie sich nicht sicher?
VP: Meine potenziellen Kunden haben noch nie eine Demonstration des Systems gesehen.
ELIZA: Bedeutet das viel für Sie?
VP: Selbstverständlich tut es das.
ELIZA: Sie wirken so, als ob Sie sich dessen sehr sicher sind.
VP: Das ist doch lächerlich.
ELIZA: Was meinen Sie damit, dass das lächerlich ist?
VP: Gibt es irgendeinen Grund, dass ich den Computer jetzt nicht benutzen kann?
ELIZA: Erzählen Sie mir mehr über Ihre potenziellen Kunden.
VP: Rufen Sie mich bitte unter 491-1850 an.[5]

ELIZA hat den Vizepräsidenten, der sehr irritiert und nicht so leicht davon zu überzeugen war, dass er mit einem Computer gesprochen hatte, natürlich nie angerufen.

Selbst wenn die Kombination von ELIZAs einfachen Sätzen und menschlicher Fantasie genügte, Leute in den 1960er Jahren zu beeindrucken, hatte sie nur wenig Neuheitswert. Das Gleiche galt für mehrere KI-Systeme, die zu dieser Zeit entwickelt wurden, sodass der amerikanische Forschungsrat 1964 ein Komitee ernannte, um den Fortschritt in der Forschung an künstlichem Sprachverständnis zu evaluieren. Als Ergebnis wurden die Forschungsmittel für dieses Feld ab 1966 gestoppt. Nach mehr als zehn Jahren Forschung und 20 Millionen US$ Forschungsmitteln waren die Computer immer noch teurer und schlechter als menschliche Dolmetscher, und am Horizont waren weit und breit keine Computer in Sicht, die ein vernünftiges Gespräch hätten führen können. Künstliches Sprachverständnis wurde plötzlich als eine Sackgasse angesehen, in erster Linie, weil den Forschern klar wurde, wie viel von der Welt ein Computer verstehen musste, um Sprache in einer sinnvollen Art und Weise zu benutzen.

Im Laufe der 1960er Jahre finanzierte das US-Militär durch die Defense Advanced Research Projects Agency (DARPA) viele ambitionierte Forschungsprojekte innerhalb der künstlichen Intelligenz, oft fast gänzlich ohne Anforderungen oder Leitlinien.

5 Davis, Galen: Unitelligent and Proud of it: Simulations, God Games, and Artificial Intelligence, Standford University.
https://web.stanford.edu/class/sts129/essays/davis2.htm

Nach 1969 wurde die Forderung an die DARPA gestellt, militärisch anwendbare Forschung zu unterstützen statt »orientierungsloser Forschung«. Es war nicht deutlich genug geworden, dass ihre Forschung zu nützlicher militärischer Technologie führen würde, und DARPAs Forschungsmittel flossen deshalb zunehmend in Projekte mit klar definierten Zielen, wie autonome Panzer. Kombiniert mit fehlendem Erfolg innerhalb künstlichen Sprachverständnisses führte das dazu, dass es in dieser Zeit fast unmöglich war, Forschungsmittel für künstliche Intelligenz aufzutreiben. Im Laufe der 1970er Jahre gab es insofern nur wenig Aktivität auf diesem Gebiet, und viele kluge Köpfe verließen die Forschung.[6]

2.4 Winter und Frühling

Die 1980er Jahre waren nicht nur ein Fest für Glam Metal und Techno. Der Fachbereich künstliche Intelligenz, der eingefroren worden war, wurde in diesem Zeitraum im Takt der Veränderung der globalen Machtbalance innerhalb aller Arten von Technologien wieder aufgetaut. Die westliche Dominanz in der Elektronik- und Autoindustrie wurde deutlich von Japan herausgefordert. 1982 lancierte Japan sein sogenanntes *Fifth Generation Computing Project*. Der Westen sah in diesem Fünfte-Generation-Projekt eine riesige Bedrohung mit dem Ziel, führend in der Computerindustrie und der KI-Entwicklung zu werden. Selbst wenn eine derartige Dominanz wohl nicht die Intention hinter dem japanischen Projekt war, so schickte es doch Schockwellen durch die westlichen KI-Bereiche und prägte die Richtung der KI-Forschung und -entwicklung. Mehrere amerikanische und europäische Unternehmen und Forschungsprojekte entstanden, um der gefühlten japanischen Bedrohung etwas entgegenzusetzen, und rückblickend war das vielleicht der erste Wettlauf um die globale KI-Dominanz. Vor den 1980ern war die KI-Forschung im Großen und Ganzen ein akademisches Vorhaben, doch in der Periode von 1982 bis 1990 setzten japanische Behörden einen neuen Standard, indem sie 400 Millionen US$ in die KI-Forschung investierten. Die Ambitionen waren (wie immer) hoch, wenn es um erwarteten Fortschritt ging, doch viele von ihnen wurden nie erfüllt. Auch wenn das Projekt »Fünfte Generation« rein fachlich gesehen keine große Bedeutung erlangte, drückte es der KI-Geschichte doch einen markanten Stempel auf. Das Projekt zeigte, dass KI-Forschung *nicht* unbedingt eine akademische Übung sein muss, sondern

6 Hughes, Thomas et al.: Funding a Revolution. Government Support for Computer Research, Committee on Innovations in Computing and Communications, Lessons from History. National Research Council, 1999.
https://web.archive.org/web/20080112001018/http://www.nap.edu/readingroom/books/far/ch9.html

auch in Zusammenarbeit mit dem privaten und dem öffentlichen Sektor stattfinden kann. Die KI-Entwicklung in den 1980ern und darüber hinaus war geprägt von einer Fokussierung auf kommerzielle Produkte; es war die Zeit großer Konferenzen mit teuren Eintrittskarten. Nicht nur Forscher, sondern auch die Industrie und Politiker waren neugierig auf künstliche Intelligenz – mit anderen Worten fast so wie heute. Der große Unterschied zu heute besteht einzig darin, dass in den 1980ern noch große Hoffnungen auf Expertensysteme gesetzt wurden und viele immer noch der Meinung waren, dass Expertenwissen die beste Art und Weise sei, intelligente Computer zu erschaffen.

Expertensysteme wurden in fast allen Bereichen erstellt; alles von der Finanzwelt bis hin zur Medizin bekam seine Dosis `if`-Sätze verabreicht. Die Zeitschrift »Business Week« stürzte sich auf das Phänomen und schlug 1984 mit der Schlagzeile auf: »AI: It's Here«. In der Broschüre zu einem Expertensystem mit Namen TIMM, Abkürzung für *The Intelligent Machine Model*, war zu lesen:

> *»Wir haben ein besseres Gehirn gebaut. Expertensysteme reduzieren Wartezeiten, Personalanforderungen und Engpässe, die durch begrenzten Zugang zu Experten entstehen. Außerdem werden Expertensysteme nicht krank, sie kündigen nicht und lassen sich nicht vorzeitig pensionieren.«*[7]

Hört sich das bekannt an? Das könnte genauso gut heute in einem Zeitungsartikel stehen, sofern »Expertensystem« durch »KI« ersetzt wird.

1984 warnte unser alter Freund John McCarthy, dass Expertensystemen der gesunde (Menschen-)Verstand fehle und dass sie ihre eigenen Begrenzungen nicht verstünden. Er illustrierte das anhand eines Expertensystems, das geschaffen worden war, um Ärzten zu helfen, indem es Medikamentendosen vorschlug. Einem Patienten mit einer ernsthaften Cholera-Infektion wurde von dem Expertensystem empfohlen, zwei hohe Dosen eines Breitbandantibiotikums zwei Wochen lang einzunehmen. Selbst wenn das Medikament höchstwahrscheinlich alle Bakterien abgetötet hätte, so hätte es den Patienten doch auch sein Leben gekostet. Solang kein Experte dafür gesorgt hat, dem System beizubringen, dass es eine Höchstdosis gibt, die Menschen vertragen oder denen sie ausgesetzt werden können, kann auch das Expertensystem das nicht wissen. Das ist ein Schwachpunkt, den alle Expertensysteme teilen, und die Forscher wussten das. Tatsächlich mündeten McCarthys Sorgen hinsichtlich der mangelnden »Vernunft« der Computer in einem eigenen Forschungsbereich innerhalb der KI, *Commonsense Reasoning* genannt. Nichtsdestotrotz standen in Werbeanzei-

7 »We've built a better brain. Expert systems reduce waiting time, staffing requirements and bottlenecks caused by the limited availability of experts. Also, expert systems don't get sick, resign, or take early retirement.«

gen Behauptungen wie »Wir haben ein besseres Gehirn geschaffen«[8], und wieder einmal waren die Erwartungen viel zu hoch, welch fantastische Intelligenz der Bereich schaffen könnte. Es sollte nicht lange dauern, bis die Forschung zur künstlichen Intelligenz erneut in eine tiefe Krise geriet.

Derartige Rückschläge sind üblich, wenn die Gesellschaft zu hohe Erwartungen in eine Technologie setzt, nicht nur innerhalb der künstlichen Intelligenz. Das Gleiche geschah zum Beispiel mit den Eisenbahnaktien in Großbritannien in den 1840er Jahren und mit dem Internet in den USA in den 1990ern (die sogenannte Dotcom-Blase). Aber gerade künstliche Intelligenz scheint geradezu ein eigenes Talent zu besitzen, in derartigen Flauten zu segeln. Und da diese Krisen das Feld so offensichtlich geprägt haben, bekamen sie einen eigenen Namen: *KI-Winter*. Der Begriff wurde bei einer Debatte eingeführt, die 1984 während der jährlichen Konferenz von *The American Association of Artificial Intelligence* stattfand. Die KI-Forscher sahen ein Muster, fast eine Kettenreaktion, die mit Pessimismus in Fachkreisen beginnt, wenn die Forscher vor allem einsehen, dass die KI-Probleme viel schwieriger zu lösen sind, als angenommen, gefolgt vom Pessimismus in den Medien, weniger Forschungsmitteln, und der Konsequenz, dass die Forscher in ihrem Fachgebiet keinen Job mehr finden und ihn verlassen müssen. Unser Freund Marvin Minsky war bei dem Treffen 1984 dabei, nachdem er den Winter in den 70ern als einer der wenigen überlebt hatte, die weiterhin bis zum Aufschwung in den 80ern in dem Bereich arbeiteten. Während der Konferenz warnte er davor, dass die Erwartungen an die Technologie erneut zu groß geworden seien. Drei Jahre später begann das, was bis jetzt der letzte deutlich markierte KI-Winter war, und Anfang der 1990er Jahre war das Feld wieder genauso eingefroren, wie es das schon in den 1970ern gewesen war.[9]

2.5 Computer, die lernen

Für viele Probleme – einschließlich alltäglicher Situationen, zu denen wir Menschen buchstäblich täglich viele Male Stellung nehmen müssen – gelingt es uns nicht, mathematische Regeln aufzustellen. Was wir inzwischen traditionelle künstliche Intelligenz nennen oder *Good old-fashioned AI*, die unter der witzigen Abkürzung *GOFAI* läuft, führte zu Programmen, die es nicht schafften, Probleme zu lösen, die unserer Meinung nach einfach sind. Dieses Phänomen bekam den Namen *Moravecs Paradox*, benannt nach dem KI-Forscher Hans Moravec. Es handelt davon, dass Aufgaben, die für Tiere oder Menschen so einfach sind, dass wir sie fast ohne nachzudenken lösen

8 »We've built a better brain.«

9 Crevier, Daniel: AI: The Tumultous History of the Search for Artificial Intelligence, 1993.

Kapitel 6
Unser künstlich intelligentes Leben

6.1 Die Revolutionen der Maschinen

Der Zeitraum, seitdem wir Menschen eine Arbeitsstelle haben, zu der wir gehen, und einen Job, für den wir bezahlt werden, macht nur einen sehr kurzen Abschnitt in der Geschichte der Menschheit aus. Das Konzept »Job«, wie wir es heute verstehen, ist relativ neu, da wir Menschen zu großen Teilen unserer Geschichte Bauern waren, ein Handwerk ausübten oder andere Rollen einnahmen, die dazu beitrugen, die lokale Gesellschaft am Laufen zu halten. Die Arbeit, die wir während des größten Teils unserer Existenz ausgeübt haben, diente uns selbst, unserer Familie und unserer Gemeinde, nicht aber einem Arbeitgeber. Auch nachdem wir Arbeitgeber, Arbeitsverträge und Arbeitszeiten bekommen haben, unterliegt die Rolle der Arbeit und unser Verhältnis zu ihr einem ständigen Wandel.

Schon wenn wir uns die Wurzeln des Begriffs anschauen, wird klar, dass wir heute unter Arbeit etwas anderes verstehen als früher: Das französische Wort für Arbeit ist *travail* und stammt vom lateinischen *trepalium* ab, einem Folterinstrument, bestehend aus drei Stöcken. Das norwegische Substantiv *arbeid* leitet sich vom deutschen *Arbeit* ab, das ursprünglich Leiden und Mühe bedeutete. Die wenigsten von uns, die das Glück haben, in dem industrialisierten, demokratischen Teil der Welt zu leben, sehen Arbeit als eine Form der Folter an. Tatsächlich haben sich viele von uns selbst ausgesucht, was sie arbeiten wollen, und für die meisten von uns dient der Job nicht nur dazu, die Rechnungen zu bezahlen, er ist auch Teil unserer Identität, was dadurch deutlich wird, dass »Und, was machst du?« zu den ersten Fragen gehört, die wir neuen Bekanntschaften stellen. Während meines Studiums verbrachte ich ein Jahr in Spanien, und ich kann mich immer noch an ein Gespräch mit einer älteren Frau erinnern, die mich nicht fragte, was ich studierte oder arbeitete, sondern »*¿A qué te dedicas?*« Für mich zeigt diese Frage, die übersetzt so viel bedeutet wie: »Welchen Dingen widmest du dich?«, wie die moderne Gesellschaft den Job ansieht: Das ist nichts, womit wir gefoltert werden, sondern etwas, für das wir uns entscheiden und dem wir unsere Zeit widmen. In dem gleichen Maße, in dem sich die Gesellschaft auf die Indi-

vidualisierung zubewegt hat, ist auch der Job immer mehr zu einem Selbstverwirklichungsprojekt geworden. Bei der Sorge, Maschinen könnten unsere Arbeitsaufgaben einfacher, billiger und besser als wir ausführen, geht es deshalb nicht nur um ökonomische Ängste, sondern auch um den Verlust unseres Selbstbildes und unserer Identität. Und beide Aspekte müssen ernsthaft in der Diskussion berücksichtigt werden, die sich darum dreht, was mit uns und der Gesellschaft geschieht, wenn Maschinen ein Niveau erreichen, auf dem sie immer mehr Aufgaben übernehmen können, die heutzutage noch wir Menschen ausführen.

Wenn Maschinen in großem Rahmen die Aufgaben von Menschen übernehmen, führt das zu so gewaltigen Umwälzungen, dass wir sie als *industrielle Revolutionen* bezeichnen. Die erste industrielle Revolution, die durch die Entwicklung von Web- und Dampfmaschinen ausgelöst wurde, bedeutete schlechte Nachrichten für viele Arbeiter. Weil die Maschinen die Arbeiter des 18. Jahrhunderts fast vollständig ersetzten, endeten diese in einer ökonomisch prekären Situation, und eine gesamte Gesellschaftsschicht brauchte lange Zeit, um wieder die gleiche Kaufkraft zu erlangen, wie sie sie vor der Revolution gehabt hatte. Mit diesem Wissen im Hinterkopf können wir die Faszination und die Besorgnis besser verstehen, mit der dem Schach spielenden »mechanischen Türken« begegnet wurde, und mit ein wenig Fantasie können wir sogar die gleiche Tendenz erkennen, wenn sich heutige Journalisten und Designer fragen, ob ein Transformer- oder ein Diffusionsmodell sie arbeitslos machen wird. Und selbst wenn die Sorge, dass Maschinen unsere Arbeitsaufgaben übernehmen könnten, bereits seit der ersten industriellen Revolution im kollektiven Bewusstsein verankert ist, haben wir mit großem Eifer weiterhin immer bessere Maschinen entwickelt. Offenbar sehnen wir uns nach Maschinen, die unsere Arbeitsaufgaben übernehmen; Geschirr abwaschen, die Wohnung saugen, den Rasen mähen usw. Tatsächlich ist es schwer, viele physisch anstrengende Jobs zu finden, für die wir *keine* Maschinen haben, die uns die Arbeit abnehmen. Um uns daran zu gewöhnen, hatten wir aber auch reichlich Zeit, nämlich seit dem 18. Jahrhundert.

Dass Maschinen grobe Arbeiten übernehmen, die wir Menschen ihnen erklären, ist nicht besonders unheimlich. Aber dass sie Intelligenz und andere Fähigkeiten entwickelt haben, die wir jahrhundertelang für einzigartig menschlich gehalten haben, erscheint uns dagegen bedrohlicher. In den letzten Jahren haben Computer begonnen, sich Fähigkeiten anzueignen, von denen wir nicht gewohnt sind, dass sie sie haben können: Sie erschaffen Bilder, schreiben Texte, sehen Zusammenhänge, die wir Menschen nicht sehen, und fassen Entschlüsse schneller und besser, als wir es können. Das ist etwas ganz anderes, und es stellt alles von den Gesellschaftsstrukturen bis hin zu unserem Selbstverständnis infrage.

Als Hilfe, um in dieser Situation zu navigieren, können wir damit beginnen, uns die Tendenzen in den industriellen Revolutionen, die wir bereits durchlaufen haben, anzusehen und uns fragen, ob das, was wir momentan erleben, auch eine industrielle Revolution ist.

Hundert Jahre nach der ersten industriellen Revolution, Ende des 19. Jahrhunderts, stand die nächste an, dieses Mal mit der Stahlproduktion, dem Benzinmotor, der Elektrizität und den Telefonkabeln, und ihr Effekt für die Arbeiter war fast ein entgegengesetzter als bei der ersten Revolution: Die Entwicklung führte zu neuen Chancen und besseren Arbeitsbedingungen. Während die erste Revolution von der Aussicht angetrieben wurde, Arbeiter durch Maschinen zu *ersetzen*, war der Antrieb für die zweite eine Technologie, die die Arbeiter *stärkte*, indem sie ihnen mehr Produktivität gab und ihre Handlungsspielräume erweiterte. Wir können diese Revolution als eine betrachten, die mehr möglich machte. Die dritte Revolution, die wir in die späten 1960ern verorten und die der Entwicklung der Telekommunikation, Elektronik und dem Computer zugeschrieben wird, folgte in vielem den gleichen Mustern wie die zweite Revolution. Die Kommunikations- und Computertechnologie hat den allermeisten in unserem Teil der Welt vieles ermöglicht und war ein Schlüsselelement in unserem ökonomischen Wachstum.

Wenn wir die früheren industriellen Revolutionen auf diese Art sortieren – einmal ersetzend, zweimal ermöglichend – ergeben sich zwei interessante Fragen:

Erstens: Besteht die Chance, dass eine bessere oder modernere Gesellschaft die negativen Konsequenzen der ersten industriellen Revolution hätte vermeiden können? Kein mechanischer Webstuhl spazierte in eine Fabrik und sagte: »Hände hoch, ich will eure Jobs!« Maschinen übernehmen nicht aus eigenem Antrieb die Aufgaben von Menschen. Es sind Menschen, die bestimmen, dass sie das tun sollen, und die Konsequenzen dieser Entscheidungen sind abhängig davon, in welcher Art von Gesellschaft das geschieht. Sicher, die ökonomisch treibenden Kräfte können so stark sein, dass wir den Eindruck gewinnen, es gäbe keine echte Wahl, ob Maschinen Arbeitsaufgaben übernehmen sollen, doch es ist die ganze Gesellschaft, die die daraus entstehenden Folgen bewältigen muss. Im Jahr 1790 arbeiteten 90 % der Amerikaner in der Landwirtschaft. 2008 betrug ihre Zahl 2 %, ohne dass 88 % der Amerikaner jetzt arbeitslos wären oder weniger Lebensmittel produziert würden, eher im Gegenteil. Laut dem World Economic Forum werden 65 % der heutigen Grundschüler später einen Beruf haben, den es heute noch nicht gibt. Mein Großvater war ein belesener Mann, aber ich glaube nicht, dass er hätte voraussehen können, dass heutige Arbeitnehmer Softwareingenieure, Datenanalytiker, Nachhaltigkeitsmanager und Droh-

nenpiloten sein würden. Seit wir industrielle Revolutionen erleben, konnten wir beobachten, wie neue Berufe entstanden und gleichzeitig andere verschwanden. Was jedoch *nicht* bedeutet, dass die existierenden Arbeitskräfte reibungslos in neue Berufe wechseln und Aufgaben übernehmen können, die sie nie zuvor ausgeführt haben. Die Fähigkeit, sich auf neue Arbeitsaufgaben und entsprechende Denkweisen einzulassen, muss trainiert werden, und die Kosten dafür, sich neue Fertigkeiten anzueignen, müssen bezahlt werden. Nur eine Gesellschaft, die diese Investitionen leistet, sorgt dafür, dass ihre Mitglieder nicht unter einer schnellen technologischen Entwicklung leiden.

Zweitens: Wenn wir die drei letzten industriellen Revolutionen als entweder ersetzend oder ermöglichend ansehen, lautet die zweite Frage, wie die nächste Revolution in dieser Reihe sein wird. Viele meinen, dass wir am Beginn der vierten industriellen Revolution stehen, die durch Genmanipulation, Blockchain, cyberphysische Systeme und – du hast es schon geahnt – künstliche Intelligenz geprägt ist, oder dass wir uns bereits mittendrin befinden. Beziehen wir uns auf den Teil einer potenziell laufenden Revolution, der von künstlicher Intelligenz vorangetrieben wird, ist es naheliegend, sich zu fragen, ob sie zu einer Technologie führen wird, die uns *ersetzt* oder uns Chancen eröffnet (*ermöglicht*), um es ein wenig auf die Spitze zu treiben. Alle drei bisherigen Revolutionen waren auf lange Sicht betrachtet von Vorteil für uns Menschen. Das gilt auch für die erste, doch auf kurze Sicht – also für rund drei Generationen – war diese Revolution für die meisten Menschen eine schreckliche Neuheit. In seinem Buch »The Technology Trap« argumentiert der Wirtschaftswissenschaftler Carl Benedikt Frey, dass diese vierte Revolution von uns, die wir uns mittendrin befinden, wie die erste erlebt wird; sie wird wahrscheinlich zu Wachstum und Wohlstand führen, aber nicht auf kurze Sicht.[1] Auf kurze Sicht wird die Technologie vielmehr diejenigen ersetzen, denen es nicht gelingt, sich anzupassen, und diese Menschen werden folglich als Verlierer der Gesellschaft enden. Solange es der Gesellschaft nicht gelingt, sich um diese Menschen zu kümmern, werden sie wütend sein und womöglich für politische Akteure stimmen, die versprechen, sie zu beschützen, also für Populisten. Eine düstere Zukunftsvision!

Nach der Niederlage gegen DeepBlue 1997 dachte der Schachweltmeister Garri Kasparow viel über das Potenzial nach, das im Zusammenspiel von Mensch und Maschine zu finden ist. Könnte das zu dem perfekten Spiel führen? Die Idee wurde bereits 1998 umgesetzt, als Kasparow und ein Computer gegen einen anderen Schachgroßmeister, Weselin Topalow, mit seinem Computer spielten. Leider gelang es keinem der

1 Frey, Carl Benedikt: The Technology Trap: Capital, Labor and Power in the Age of Automation, 2019.

Spieler, seine Fähigkeiten mit den Fähigkeiten des Computers zu kombinieren. Ich habe den Verdacht, dass es das gleiche Gefühl war, das viele Firmen und Angestellte bei der Digitalisierung empfinden. Computer zu benutzen, um Menschen zu unterstützen, uns zu entlasten und gleichzeitig das Beste aus uns herauszuholen, ist vorsichtig ausgedrückt leichter gesagt, als getan, und erfordert große Mengen an Anpassung, Lernbereitschaft und nicht zuletzt den Wunsch, sich um die Angestellten zu kümmern.

Ein oft bemühter Vergleich für die Rolle der Technologie in der Gesellschaft lautet so: Stell dir vor, dass wir Menschen auf einer Insel leben. Mit der Zeit steigt der Wasserspiegel, und wir Menschen müssen uns auf höher gelegene Gebiete und schließlich in die Berge zurückziehen. Zum Schluss ist der Meeresspiegel so hoch gestiegen, dass wir nichts anderes mehr tun können, als uns verzweifelt an die höchsten Berggipfel zu klammern. In diesem Gleichnis können die Menschen nicht schwimmen und keine Boote bauen. Sie sind einer Entwicklung ausgesetzt, die sie nicht beeinflussen können. Aber das muss nicht eine Beschreibung des tatsächlichen Effekts von (neuer) Technologie auf die Gesellschaft sein, denn Technologie ist keine Naturgewalt, der wir schutzlos ausgeliefert sind. Jede Technologie wird entwickelt, um Probleme zu lösen (das ist ihr eigentlicher Zweck). Welchen Effekt sie auf die Gesellschaft hat, ist deshalb abhängig von der Gesellschaft selbst, in der die Entwicklung stattfindet. Aufgrund unseres Wirtschaftsmodells wird Technologie, die zu ökonomischem Wachstum führt, letztendlich auch meistens eingesetzt. Welchen Effekt sie auf die Gesellschaft hat, ist deshalb eine ökonomische und politische Frage. Wir können uns wünschen, dass uns Maschinen Aufgaben abnehmen, sodass wir Menschen unsere Zeit nutzen können, uns um einander zu kümmern und uns selbst zu verwirklichen. Letzteres führt jedoch nicht zu wirtschaftlichem Wachstum, und solange das die treibende Kraft hinter unserer Entwicklung ist, bekommen wir ein immer größeres Problem, je stärker sich die Technologie entwickelt.

In der westlichen Welt haben wir in der Moderne den gesellschaftlichen Wert immer mehr an der Funktion bemessen, die jemand ausübt. Mit die erste Frage, die wir Menschen stellen, denen wir zum ersten Mal begegnen, ist die nach ihrem Beruf, also praktisch danach, welche Funktion sie in der Gesellschaft erfüllen und welchen Beitrag sie leisten. Unser gesamtes Gesellschaftsmodell und unsere Werteanschauung dreht sich darum, was wir beitragen und ausrichten. Das steht im Kontrast zu einer Werteanschauung, nach der die Dinge auf der Welt – Menschen, Tiere, Natur – allein einen Wert haben, nur weil sie *sind*. Ein Baum, der im Wald steht, hat keinen Wert, nur weil er dort steht und *ein Baum ist,* sondern weil er als Baumaterial genutzt werden kann, den Boden festigt, zu Brennholz werden kann oder Kohlendioxid aufnimmt.

Wenn wir der Meinung sind, dass der Wert eines Künstlers mit der Kunst zusammenhängt, die dieser produziert, wird ein Diffusionsmodell zu einer ernsten Bedrohung für den Künstler, weil das Diffusionsmodell ebenso gute Kunst schaffen kann wie der Künstler selbst. Wenn wir dagegen davon ausgehen, dass Künstler an sich einen Wert haben, allein weil sie menschliche Künstler sind, stellt ein Diffusionsmodell nicht die geringste Bedrohung dar. Mit dieser Beschreibung lehne ich mich eng dem existentialistischen Philosophen Martin Heidegger an, der sagt, dass Technologie nicht nur irgendwelche Dinge sind, sondern vielmehr die Art an sich, wie wir uns in Bezug auf die Welt verstehen und verhalten. Nun wollen wir nicht in philosophische Tiefen abtauchen, aber auf jeden Fall ist es interessant, dass die Technologieentwicklung ein scharfes Licht auf die Werte wirft, auf denen wir unsere Gesellschaft gründen, ohne dass wir uns darüber klar sind. Mit das Faszinierendste an der künstlichen Intelligenz als Technologie ist die Frage, inwieweit sie uns zwingt, Stellung zu den schwierigen ethischen Dilemmata zu beziehen.

6.2 Die Welt der Ethik

Eine nicht mehr zu bremsende Straßenbahn rast auf den Schienen bergab. Aus irgendeinem absurden Grund sind fünf Menschen auf den Schienen festgebunden, und die Bahn bewegt sich direkt auf sie zu. Du bist Zeuge der Situation und stehst zufällig neben einem Weichenhebel, mit dessen Hilfe du die Spur wechseln kannst. Aber du siehst auch, dass auf dem anderen Gleis eine Person liegt, die nicht entkommen wird, wenn du den Hebel umlegst. Du hast zwei und wirklich nur zwei Alternativen: Entweder machst du nichts und die Straßenbahn wird die fünf Menschen auf dem Hauptgleis töten, oder du legst den Hebel um und leitest damit die Straßenbahn so um, dass sie eine andere unschuldige Person tötet. Welches ist die ethisch bessere Alternative? Oder wann tust du das Richtige?

Dieses äußerst konstruierte Gedankenexperiment ist eines der sogenannten *Trolley Problems* oder Straßenbahnprobleme. Andere Varianten des Trolley-Problems enthalten weitere Details über die potenziellen Opfer, wie Alter, Geschlecht, Berufsstatus usw. Das Ziel dieser Gedankenexperimente ist es, ethische Dilemmata zu erforschen und uns eine Anleitung für die Diskussion einer Moralphilosophie und die verschiedenen Seiten bei den Kompromissen an die Hand zu geben. Seit diese Probleme von der Philosophin Philippa Foot 1967 beschrieben wurden, waren sie Gegenstand vieler Diskussionen, und es herrscht breite Einigkeit darüber, dass es keine »richtige« Antwort darauf gibt, was die richtige Entscheidung ist. Verschiedene Menschen haben

unterschiedliche Perspektiven, und es kann für viele gegensätzliche Handlungen auf korrekte Art argumentiert werden. Willkommen im Kopfschmerzfach Ethik.

Ich selbst habe in den letzten Jahren viel Zeit mit dem Versuch verbracht, herauszufinden, wie Ethiker denken. Und wenn es eine Einsicht gibt, die sich von allen, die ich gelernt habe, abhebt, dann die, dass es bei Ethik *nicht* darum geht herauszufinden, »das Richtige zu tun«: In der Ethik geht es darum, unterschiedliche Interessen gegeneinander abzuwägen. Für eine Teilchenphysikerin, die an XAI forscht, kann es eine große Herausforderung bedeuten, diese Denkweise zu verinnerlichen. Wenn wir Algorithmen entwickeln, Hypothesen testen und Theoreme beweisen, gibt es jeweils eine richtige Antwort, auf die wir uns einigen können, und die Aufgabe ist es, diese Antwort herauszufinden. So ist es in der Ethik nicht, hier besteht die Aufgabe eher darin, all die unterschiedlichen Empfindlichkeiten und Interessen zu verstehen, die von dem System beeinflusst werden, das beurteilt werden soll.

Wie bereits erwähnt, ist einer der interessantesten Aspekte von KI, dass sie uns zwingt, uns gegenüber ethischen Dilemmata zu verhalten, für die das Trolley-Problem ein gutes Beispiel ist. Denn auch wenn selbstfahrende Autos bisher noch nicht sehr verbreitet sind, werden autonome Fahrzeuge aller Art entwickelt, was bedeutet, dass die Frage »Welches Leben soll Priorität haben?« beantwortet werden muss. Um zu untersuchen, wie verschiedene Menschen in ähnlichen Situationen wie dem Trolley-Problem bezüglich autonom fahrender Autos denken, entwickelten Forscher am Massachusetts Institute of Technology (MIT) ein Onlineexperiment, genannt *The Moral Machine*. Bei diesem Experiment werden Internetnutzern aus der ganzen Welt unterschiedliche Situationen gezeigt, in denen selbstfahrende Autos einen Entschluss fassen müssen, welches Leben Priorität haben soll. Eine Situation kann sein, dass plötzlich ein Fußgänger auf die Fahrbahn tritt, oder eine andere, dass ein anderes Auto auf die eigene Fahrspur kommt, wobei viele Verkehrsteilnehmer unterschiedlichen Alters, verschiedener Ethnie, Ausbildung und außerdem auch Tiere involviert sind. Die Ergebnisse dieses Experiments sind faszinierend und zeigen, dass moralische Präferenzen sich sehr stark zwischen verschiedenen Kulturen und Regionen unterscheiden.

Die größten sichtbaren Unterschiede bestanden zwischen drei Gruppen von Ländern auf der Welt, die von den Forschern im Experiment als westliche, östliche und südliche benannt wurden.[2] Diese drei Gruppen bestehen aus insgesamt 130 Ländern, aus denen es für das Experiment mindestens hundert Antworten gab. Die westliche Gruppe enthält unter anderem Nordamerika und die meisten europäischen Länder mit christlich geprägten Kulturen. Die östliche Gruppe enthält ostasiatische Länder

2 Awad, Edmond et al.: The Moral Machine experiment, in: Nature 563, 2018, S. 59–64.

wie Japan und Taiwan, deren kulturelle Werte durch den Konfuzianismus beeinflusst sind, und islamische Länder wie Pakistan und Saudi-Arabien. Die südliche Gruppe enthält Zentral- und Südamerika und Länder mit historisch französischem Einfluss. Die Resultate der Studie zeigen beispielsweise, dass Menschen aus der westlichen Gruppe es meistens vorziehen, dass sich das Fahrzeug passiv verhält, also keine Entscheidung trifft, um das Ergebnis zu verändern. Außerdem sehen wir, dass Menschen aus der östlichen Gruppe am stärksten die Rettung von Fußgängern und gesetzestreuen Menschen präferieren, während Menschen aus der südlichen Gruppe größere Präferenzen zeigen, Frauen zu retten. Globale Trends können wir in der Priorisierung der Rettung von Menschenleben vor Tierleben finden und außerdem darin, so viele Leben wie möglich und junge Leben zu retten.

Es ist interessant, aber nicht überraschend, dass Völkergruppen mit unterschiedlichem Hintergrund und Wertgrundlagen verschiedene ethische Präferenzen zeigen. Was bedeutet, dass Technologie, die in verschiedenen Teilen der Welt entwickelt wird, höchstwahrscheinlich auf unterschiedlichen ethischen Abwägungen basiert – es sei denn, alle Länder dieser Welt würden sich einig werden über ethische Leitlinien für künstliche Intelligenz. Aber genau das ist wohl unrealistisch, da wir es nicht einmal schaffen, uns innerhalb unserer eigenen Länder und Erdteile über ethische Leitlinien einig zu werden.

In einer Unfallsituation dauert der Handlungszeitraum selten länger als den Bruchteil einer Sekunde, viel zu wenig, als dass ein Mensch eine ethische Abwägung vornehmen kann. In gefährlichen Situationen reagieren die meisten Menschen instinktiv oder sind vor Schreck wie gelähmt. Computer haben keine dieser Beschränkungen. Ein autonom fahrendes Auto, das in einen Unfall verwickelt ist, wird nicht von Stress beeinflusst und hat genügend Zeit, einen klaren Entschluss zu fassen. Das ist ein Beispiel dafür, dass neue Handlungsräume und Entscheidungen in der Praxis *entstehen*, wenn Computer und Algorithmen in unserer Welt agieren: Das Trolley-Problem war zunächst nur ein Gedankenexperiment, konstruiert, um verschiedene ethische Perspektiven zu beleuchten, ist aber inzwischen zu einer aktuellen Frage geworden, die in der realen Welt beantwortet werden muss.

Für unsere Zeit ist ebenso charakteristisch, dass ein großes Unternehmen nicht darauf gewartet hat, dass diese schwierige ethische Diskussion die öffentliche Debatte erreicht und dort entschieden wird, sondern selbst einen Entschluss gefasst hat. Während der Pariser Motorshow im Oktober 2016 wurde der damalige Leiter des Bereichs aktive Sicherheit bei Mercedes, Christoph von Hugo, zu diesem Thema interviewt und später folgendermaßen zitiert:

»Wenn du weißt, dass du mindestens eine Person retten kannst, dann rette zumindest diese eine. Rette denjenigen, der im Auto ist.«[3]

Das kann so interpretiert werden, dass Mercedes priorisiert, den Fahrer des Autos zu retten.

Was aus kommerzieller Perspektive nicht überraschend ist, da die wenigsten von uns bereit wären, ein autonom fahrendes Auto zu kaufen, das uns zugunsten anderer Verkehrsteilnehmer opfern würde. Zu der Geschichte gehört auch, dass Mercedes wenige Tage später äußerte, dass von Hugo falsch zitiert worden sei und man in keiner Weise geplant hätte, Fußgänger zu opfern. Stattdessen erklärte Mercedes, dass ihre selbstfahrenden Autos »auf die größtmögliche Sicherheit aller Verkehrsteilnehmer abzielen« (*»aim to maintain the highest possible safety for all road users«)* – was jedoch nicht die ursprüngliche Frage beantwortet. Wir sind uns alle darin einig, dass gut funktionierende autonome Fahrzeuge die sicherste Alternative wären: Denn wenn diese sicher genug sein werden, um sie flächendeckend einzusetzen, werden auf jeden Fall deutlich weniger Menschen im Verkehr ihr Leben lassen als zu heutiger Zeit mit menschlichen Fahrern – doch das ist nicht der Punkt. Der Punkt ist, dass wir heute in einer Zeit leben, in der schrecklich viele schwierige ethische Abwägungen getroffen werden müssen, ohne dass wir ein gutes System zur Verfügung haben, um das zu gewährleisten.

2017 schlug die deutsche Ethikkommission für automatisiertes und vernetztes Fahren die weltweit erste (und bislang einzige) Sammlung ethischer Regeln für autonome Fahrzeuge vor.[4] In dieser gibt Regel Nummer 7 eindeutig vor, dass Menschenleben höchste Priorität hat vor dem Schutz von Tierleben. Diese Regel entspricht der Prioritätensetzung im Moral-Machine-Experiment. Andererseits nimmt Regel Nummer 9 keine Stellung dazu, ob und wann autonome Fahrzeuge die wenigen opfern soll, um die vielen zu retten, sondern lässt diese Frage unbeantwortet. Dagegen stellt Regel Nummer 9 fest, dass Unterscheidungen, basierend auf persönlichen Eigenschaften wie dem Alter, verboten sein sollen. Das steht in klarem Kontrast zu der Präferenz, die Jungen zu schützen, wie sie das Moral-Machine-Experiment aufstellte.

2011 zeigte eine Studie, dass Richter dazu neigen, vor einer Mahlzeit strenger zu urteilen und milder in ihrem Spruch sind, wenn sie gegessen haben. Das wurde später als

3 AutoExpress: Paris Motor Show 2016: News, round-up and show report, 03.10.2016: »If you know you can save at least one person, at least save that one. Save the one in the car.«

4 Bundesministerium für Verkehr und digitale Infrastruktur (BMVI), 2017: Ethik-Kommission Automatisiertes und Vernetztes Fahren. Bericht Juni 2017. *https://bmdv.bund.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile*

the hungry judge effect bekannt und ist ein beliebtes Argument dafür, dass der Justiz die Objektivität und Konsequenz von Computern guttäte. Wenige Jahre später wurde das Programm *COMPAS* (die Abkürzung steht für *Correctional Offender Management Profiling for Alternative Sanctions)* in den USA entwickelt, um zu entscheiden, ob Angeklagte in Untersuchungshaft gebracht oder freigelassen werden sollten, indem es prognostizierte, wie hoch die Wahrscheinlichkeit sei, dass ein Angeklagter, während er auf seine Gerichtsverhandlung wartet, eine neue Straftat begehen könnte. Basierend auf mehr als hundert Informationen über die Angeklagten, inklusive Alter, Geschlecht und Vorstrafen, berechnet COMPAS eine Risikospanne zwischen 1 und 10. Ein wichtiges Detail ist dabei, dass die Ethnizität *nicht* unter diesen Informationen zu finden ist.[5] Die Berechnungen von COMPAS können potenziell einen großen Einfluss auf das Leben eines Menschen haben, da die Angeklagten mit einem hohen Risikolevel (5–10) größtenteils in Haft genommen werden, während diejenigen, die ein niedriges Risikolevel aufweisen (1–4) meistens freigelassen werden.

Der Gedanke, der hinter COMPAS steht, ist wunderbar: Ein Computerprogramm behandelt die Personen alle gleich, egal, wie sie aussehen, ob sie an dem Tag nervös sind oder ob sie sich an dem Tag, an dem sie beurteilt werden sollen, einen Anzug leisten können, und das Programm kann von jedem systematisch getestet werden, um herauszufinden, wie es sich aufführt. Die US-amerikanische Stiftung für investigativen Journalismus, ProPublica, tat 2016 genau das und ihre Untersuchung zeigte, dass weiße Amerikaner und Afroamerikaner ganz unterschiedlich in der Bewertung von COMPAS abschneiden: Afroamerikaner haben eine fast doppelt so hohe Wahrscheinlichkeit, von COMPAS in ein hohes Risikolevel eingestuft zu werden als ethnisch Weiße, während Weiße häufiger ein niedriges Risikolevel bekamen und dennoch Gesetzesbrüche begingen.[6] Wie bereits erwähnt, hatte COMPAS keine Kenntnis von der jeweiligen Ethnie. Warum behandelt es dann trotzdem systematisch verschiedene Ethnien unterschiedlich? Die Antwort lautet, dass es in den Daten, die COMPAS benutzt, indirekte Informationen über die Ethnie gibt. Die Informationen über die Angeklagten könnten ebenso gut benutzt werden, um eine ethnische Zugehörigkeit festzustellen. Das, kombiniert mit der Tatsache, dass COMPAS auf historischen Daten aus der US-amerikanischen Gerichtspraxis basiert, führt dazu, dass die Diskriminierung von Afroamerikanern in der amerikanischen Gesellschaft ihren Weg geradewegs in die von COMPAS ideell gesehen objektive Bewertungen führt. Das Programm COMPAS selbst ist also nur das halbe Problem. Die Wurzel des Problems liegt in der Gesellschaft, die die historischen Daten darstellen. Und die Lösung dieses Problems liegt in der Zukunft, wenn wir Datenanalysen verwenden können, um systematisch

5 Northpointe, Inc.: Practioner's Guide to COMPAS Core, 2015.
6 ProPublica: Machine Bias, 23.05.2016.

unterschiedliche Behandlungen zu entdecken und sie zu korrigieren. Leider scheint das alles andere als einfach zu sein.

Womit können wir arbeiten? Sicher, wir haben ein Programm, COMPAS, das konsequent ein niedrigeres Risikolevel für ethnisch Weiße als für Afroamerikaner berechnet, auch wenn sich herausstellt, dass die ethnisch Weißen, die freigelassen werden, häufig neue Gesetzesbrüche begehen. Mit anderen Worten verhalten wir uns zwei Gruppen im Modell gegenüber unterschiedlich, und die Frage, mit der wir uns befassen müssen, lautet: Wie kann das gerecht funktionieren? Müssten wir die Fehlerquote zwischen den beiden Gruppen gleich halten, also uns bei Afroamerikanern ebenso oft irren wie bei ethnisch Weißen? Oder sollten wir die Menschen mit dem gleichen Risikolevel gleichbehandeln, unabhängig von ihrer ethnischen Zugehörigkeit? Entscheiden wir uns für Ersteres, dass es gerecht ist, für eine gleich große Fehlerquote in beiden Gruppen zu sorgen. Um das zu erreichen, müssten wir den beiden Gruppen unterschiedliche Schwellen einräumen, um in Haft genommen zu werden. Das heißt, dass Weiße in Untersuchungshaft kämen, wenn sie ein Risikolevel von fünf vorweisen, während Afroamerikaner ein Risikolevel von sieben benötigten, um inhaftiert zu werden. Damit behandeln wir Menschen aber unterschiedlich, basierend auf ihrer ethnischen Zugehörigkeit. Das fühlt sich nicht nur ungerecht an, das ist sogar gesetzlich verboten.

Die andere Alternative, Menschen mit gleichem Risikolevel eine gleiche Behandlung zukommen zu lassen, ist das, was wir heute tun, und was dazu geführt hat, dass Afroamerikaner viel strenger behandelt werden als ethnisch Weiße. Wir müssen einsehen, dass wir bei unterschiedlicher Vorgehensweise nicht gerecht sein können, oder präziser gesagt: Wir schaffen es nicht, die Fehlerquote zwischen zwei Gruppen gleich zu halten, während wir gleichzeitig die Individuen der verschiedenen Gruppen gleichbehandeln. Das stimmt mit etwas überein, das wir aus der Statistik kennen, nämlich dass sich Gruppengerechtigkeit und individuelle Gerechtigkeit gegenseitig ausschließen, was bedeutet, dass Sätze wie »Alles muss gerecht ablaufen«, buchstäblich gesehen *keinen Sinn ergeben*: Will man Gerechtigkeit erreichen, muss man spezifizieren, welche Art von Gerechtigkeit, wohl wissend, dass diese auf Kosten anderer Arten von Gerechtigkeit geht.

Was bedeutet das für das Ziel, COMPAS einzusetzen, um mit systematisch unterschiedlicher Behandlung aufzuräumen? Alles, was wir wollten, war doch nur, ein System objektiver und effektiver zu machen, und plötzlich stehen wir vor der allumfassenden Frage, was gerecht ist. Und viele der unangenehmen Schlagzeilen hinsichtlich diskriminierender künstlicher Intelligenz behandeln genau das: Historische Daten führen zu Modellen, die nur schwer gerecht zu gestalten sind, weil jede Gerechtigkeit

ihren Preis hat. Und wir müssen gar nicht bis zur amerikanischen Rechtsprechung gehen, um auf diese Art Herausforderung zu stoßen. Derartige Herausforderungen entstehen jedes einzelne Mal, wenn Daten Gruppen verschiedener Menschen enthalten. Selbst wenn die Beispiele weniger dramatisch sind, stoßen sämtliche norwegischen öffentlichen Behörden bereits auf die gleiche Herausforderung. Datenwissenschaftler und KI-Forscher sollen nicht entscheiden, welche Gerechtigkeit wir in der Gesellschaft priorisieren wollen. Wir können dafür sorgen, dass die Systeme, die wir entwickeln, weniger rassistisch sind als Menschen, aber wie diese Gerechtigkeit aussieht und wer dadurch eventuell Schaden erleidet, das muss auf höherer Ebene entschieden werden. Nach norwegischen Gesetzen ist Diskriminierung verboten, aber je häufiger wir Computer zur Entscheidungsfindung einsetzen, desto klarer werden wir erkennen, dass wir mit der diskriminierenden Praxis der Vergangenheit aufräumen müssen, indem wir eine schwierige Wahl treffen. Wer soll diese schwierigen Entscheidungen treffen, und wie können wir dafür sorgen, dass sie überall gleich getroffen werden?

6.3 Jemand muss entscheiden

Im Frühling 2022 untersuchte eine dänische Expertenkommission den Gebrauch datenbasierter Modelle und kam zu dem Schluss, dass diese reguliert werden müssen.

> *»Es sind klarere Rahmenbedingungen dafür notwendig, inwieweit Informationen über Bürger in den öffentlichen Systemen mithilfe von Algorithmen miteinander verknüpft werden.«*[7]

Die Niederländer sind bereits weiter, sie entwickelten schon 2021 ein Prüfungsrahmenwerk für Algorithmen und haben seitdem datenbasierte Modelle untersucht, die im öffentlichen Bereich verwendet werden. Im Frühling 2022 zeigte ein Überprüfungsbericht von neun verschiedenen Modellen, die in den Niederlanden benutzt wurden, dass sechs von ihnen nicht die Anforderungen des Prüfungsrahmenwerkes erfüllten, sondern Schwächen in Form von *»unzureichender Kontrolle über die Leistung der Algorithmen und ihrer Auswirkung auf Verzerrungen, mangelnden Datenschutz und unbefugten Zugriff«*[8] aufwiesen. Ich selbst habe mir die bemängelten Systeme nicht angeschaut, denke aber nicht, dass der Grund, weswegen dänische wie

7 Nordjyske: »Experten wollen den Einsatz von Algorithmen im öffentlichen Sektor regulieren«, 19.04.2022.

8 »Inadequate control over the algorithm's performance and impact to bias, data leaks and unauthorized access.« Netherlands Court of Audit: An audit of 9 Algorithms used by the Dutch Government, 18.05.2022.

auch niederländische Behörden daraus den Schluss ziehen, dass datenbasierte Modelle im Auge behalten werden sollten, mit schlecht arbeitenden KI- und Softwareentwickler*innen gerade in diesen Ländern zu tun hat. Ich glaube vielmehr, das macht deutlich, wie schwierig es ist, robuste, sichere Systeme zu schaffen.

In Norwegen haben wir keine Kommission, die Modelle für maschinelles Lernen überwacht, die im öffentlichen (oder in egal welchem) Bereich benutzt werden, oder datenbasierte Algorithmen jeglicher Art kontrolliert. Wenn wir eine solche Kommission hätten, hätte sie meiner Meinung nach mehr als genug zu tun. Ich selbst habe mich mehrere Male für eine norwegische »Algorithmenaufsicht« ausgesprochen, auch wenn der Name unpräzise ist. Ich meine, dass die Kontrolle fortschrittlicher, datenbasierter Systeme das einzig Verantwortungsvolle ist, das wir derzeit in Norwegen und darüber hinaus tun können. Aber *nicht*, weil ich glaube, dass derartige Systeme von norwegischen Akteuren mit unlauteren Absichten entwickelt würden, ganz im Gegenteil.

Eine der wichtigsten Herausforderungen bei der Einbindung unserer Werte in digitale Systeme besteht darin, dass wir uns zunächst einmal darüber einig werden müssen, was diese Werte genau sind. Künstliche Intelligenz, sowohl symbolische als auch subsymbolische, lebt von und auf Computern. Die KI-Gesetzgebung wie auch die KI-Ethik müssen deshalb auf eine Art repräsentiert werden, dass Computer sich dazu verhalten können. Hier haben wir zwei Alternativen. Die erste ist, unsere Werte knallhart durchzusetzen, also das richtige Verhalten direkt im Programmcode zu spezifizieren. Damit das möglich wird, müssen wir uns vom Abwägen der Ethik und vom Ermessensspielraum, den unser Rechtssystem oft nutzt, abwenden und ganz konkrete Regeln für die Programme schreiben. Aber woher sollen wir diese Regeln nehmen? Schauen wir uns zunächst die Ethik an, denn die gute Neuigkeit ist, dass es hier nicht an Kandidaten mangelt: Es gibt buchstäblich Hunderte Regelwerke für ethische künstliche Intelligenz weltweit, Hunderte! »Ethische KI« ist zu einem Modebegriff geworden, und heute kann man kaum noch ein Technologieunternehmen finden, das nicht seine eigenen ethischen Richtlinien für künstliche Intelligenz aufstellt. Einige davon sind so vage, dass sie offenbar nur Teil einer Marketingstrategie sind, während andere sehr viel solider daherkommen. Unter den besser durchdachten und gut geschriebenen befinden sich das White Paper der EU für vertrauenswürdige künstliche Intelligenz und die ethischen Richtlinien der EU für künstliche Intelligenz. Den Grundstein für ethische künstliche Intelligenz gibt es bereits, aber leider nur den. Keines der Dokumente da draußen ist so konkret, dass es direkt in Gebrauch genommen werden könnte – genau wie die nationale norwegische Strategie für künstliche Intelligenz.

Was die Justiz betrifft, gibt es die spannende Nachricht, dass es keine spezifische Regulierung künstlicher Intelligenz gibt – noch nicht. Alle KI-Systeme, ob nun Expertensysteme oder Modelle für maschinelles Lernen, müssen sich natürlich an existierende Vorschriften halten und stehen genauso wenig im rechtsfreien Raum wie Donald Trump. Das Problem ist, dass die Rechtsprechung, wie wir sie heute haben, geschrieben wurde, bevor irgendjemand die blitzartige Entwicklung der künstlichen Intelligenz voraussah. Und inzwischen haben wir eingesehen, dass viele der Regulierungen, die wir jetzt haben, gleichzeitig zu kurz und zu weit greifen, um dafür sorgen zu können, dass Modelle für maschinelles Lernen unsere Probleme lösen, ohne neue zu produzieren.

Nehmen wir ein Beispiel: 2015 wurde ein Gesetz eingeführt, das verbietet, das Geschlecht als ein Kriterium beim Berechnen eines Versicherungspreises heranzuziehen. Ganz gleich, ob du nun Frau oder Mann bist, darf das keinen Einfluss darauf haben, wie viel dich deine Versicherung kostet; andere Kriterien sollen das lenken. Das Ziel dieses Gesetzes war, dass die Versicherungspreise geschlechtsneutral sein sollen, was gut zu unseren Werten bezüglich der Gleichstellung der Geschlechter passt. Aber die merkwürdige Nachricht ist, dass der Unterschied im Preis für eine Versicherung zwischen Frauen und Männern gestiegen ist, und das ganz besonders, nachdem das Gesetz eingeführt wurde. Und es sind hier nicht die Versicherungsgesellschaften, die schummeln. Es liegt daran, dass Versicherungsgesellschaften in wachsendem Grad fortschrittliche Datenanalysen benutzen, um die Preise für eine Versicherung zu berechnen, kombiniert mit der Tatsache, dass Frauen und Männer im Durchschnitt unterschiedliche Versicherungsrisiken aufzeigen. Ein datenbasiertes Modell braucht keine Information über das Geschlecht, um statistische Risikounterschiede zwischen Frauen und Männern aufzuschnappen. Nehmen wir nur Verkehrsunfälle: Es gibt viele Indikatoren, die nicht direkt Verkehrsunfälle verursachen, aber oft mit Unfällen verknüpft werden, beispielsweise die Fahrzeugmarke, kombiniert mit dem Alter. Wenn ich sage: »Alter: 18 Jahre. Fahrzeug: 900 ccm, Motorrad, Marke Yamaha«, was tippst du, wie hoch das Unfallrisiko für den Fahrer sein wird? Und als Zusatzaufgabe, auf welches Geschlecht tippst du? In unserer heutigen Gesellschaft sind Hobbys und Fahrzeugmarken starke Indikatoren – sogenannte Proxyvariablen – für das Geschlecht, was dazu geführt hat, dass datenbasierte Modelle sie schnell aufschnappen. Selbst wenn *die Absicht* des Gesetzes also eindeutig und noch dazu sinnvoll ist, funktioniert es nicht, nicht zuletzt aufgrund des neuen *Kontextes*, in dem das Gesetz funktionieren soll, nämlich fortschrittlicher Datenanalyse. Was nicht bedeutet, das Gesetz wäre schlecht, allein, dass es nicht funktioniert. Trägst du im Winter deine Sommerjacke, wird die Jacke nicht wie gewünscht funktionieren, was nicht daran liegt, dass

die Sommerjacke schlecht oder kaputt ist, sondern daran, dass sie für einen anderen Kontext gemacht wurde.

Das illustriert, dass für die Justiz, genau wie für die Ethik, die Begegnung mit der künstlichen Intelligenz eine große Herausforderung darstellt. Unklarheiten und eine fehlende Rechtspraxis führen dazu, dass wir keine Beschreibungen oder konkrete Beispiele dafür haben, wie wir die künstliche Intelligenz entwickeln und benutzen sollen. Der Zweck ist klar, aber nicht die konkreten Regeln. Hätten wir eine Kommission ernannt, die die Entwicklung und den Gebrauch von Modellen für maschinelles Lernen in Norwegen beaufsichtigt, hätten wir sehr schnell eingesehen, wie viel Klärungsbedarf besteht, und wie viele Entscheidungen getroffen werden müssen, potenziell auf politischer Ebene. Dann wäre die Verantwortung für diese Entschlüsse den Juristen, Programmierern und leitenden Angestellten in norwegischen Betrieben und Institutionen von den Schultern genommen und einer Instanz übertragen worden, die das Mandat bekommen hätte, sich darum zu kümmern. Diese Instanz – ob wir sie nun KI-Kommission oder Algorithmenaufsicht oder wie auch immer nennen – könnte einen koordinierenden Effekt haben und sich um Ermessensfragen und Zweifelsfälle kümmern und dafür sorgen, dass diese außerdem in ganz Norwegen in gleicher Weise behandelt werden. Eindeutige Regularien und offensichtliche Zuständigkeiten erleichtern die Entwicklung, und in Norwegen sehen wir, dass der Mangel daran dazu führt, dass es uns nicht gelingt, effektive Werkzeuge einzusetzen, weil wir Angst haben, Fehler zu machen. Wir lassen einfache Lösungen links liegen, die unsere Probleme beheben könnten; obwohl man argumentieren kann, dass diese Herangehensweise womöglich nicht weniger unethisch ist, als KI-Technologien zu nutzen und dabei Fehler zu machen.

Nachdem ich mit vielen KI-Entwicklern gesprochen habe, bin ich zu der Überzeugung gelangt, dass eine strengere Regulierung künstlicher Intelligenz hilfreich wäre und kein Hindernis, weil diese Regelungen zumindest so konkret ausfallen müssten, dass die Entwickler wenigstens wüssten, welche Anforderungen für die Systeme gelten, die sie schaffen. »Regulierung« ist ein strenger Begriff und hört sich vielleicht sehr schwerfällig an, aber denken wir doch nur an den Verkehr: Wie würde es auf den Straßen aussehen, auf denen sich Hunderte Autos ohne Regeln mit schneller Geschwindigkeit bewegen? Das wäre nicht nur ungemein gefährlich, sondern auch sehr viel ineffektiver als Verkehr, der reguliert wird. Verkehr funktioniert besser, je klarer er reguliert wird, weil dann alle wissen, was sie dürfen und was sie nicht dürfen, und weil wir darauf vertrauen, dass sich alle an die Regeln halten, weil Strafen drohen, wenn man es nicht tut. Um es auf die Spitze zu treiben: Wo möchtest du lieber Auto fahren, in Deutschland oder in Indien?

Auch wenn ich glaube, dass eine »Algorithmenaufsicht« als koordinierende, entscheidende Instanz fungieren könnte, um die Entwicklung von Systemen im Einklang mit unseren Werten zu ermöglichen, birgt diese Lösung einige Herausforderungen. Es ist interessant, dass ich bis jetzt nicht einem einzigen KI-Forscher oder auch nur einer KI-Entwicklerin begegnet bin, die sich *nicht* ein Organ denken können, das rund um die Entwicklung künstlicher Intelligenz Kontrolle ausüben und zur Klärung beitragen könnte, während die meisten Juristen, mit denen ich gesprochen habe, der Meinung waren, dass das keine gute Idee wäre. Ein gutes Gegenargument ist beispielsweise, dass Modelle für maschinelles Lernen je nach Einsatzgebiet vor unterschiedlichen Herausforderungen stehen; es wäre in der Praxis sehr, sehr schwierig, die Kompetenz zusammenzubringen, eine Überwachung der Modelle sowohl im Klassenraum als auch auf Ölplattformen durchzuführen. Hierauf würde ein KI-Programmierer entgegnen, dass es die gleichen Lernalgorithmen sind, die gebraucht werden, unabhängig davon, ob die Daten Schülerinnen und Schüler oder Kohlenwasserstoffe beschreiben, und dass so gut wie kein Entwickler beim maschinellen Lernen nur auf einen Sektor spezialisiert ist. In dieser Debatte hat wohl niemand vollkommen recht oder irrt sich auf ganzer Linie; beide Seiten zeigen vielmehr, dass Programmierer und Juristen unterschiedliche Weltbilder haben, schon aufgrund der Art, wie sie zu arbeiten gewohnt sind (und es sein müssen). Das veranschaulicht, wie schwierig es ist, eine einheitliche Praxis für maschinelles Lernen in der Gesellschaft zu koordinieren, weil es mehrere unterschiedliche Berufsgruppen mit ihren eigenen Perspektiven und Arbeitsweisen gibt, die man berücksichtigen und unter einen Hut bringen muss.

Ganz zu Beginn dieser Diskussion sagte ich, dass wir *zwei* Alternativen haben, wenn es darum geht, Justiz und Ethik auf eine Art und Weise zu repräsentieren, dass Computer damit umgehen können. Zusätzlich zu dem Ansatz, herauszufinden, wie wir von Juristen und Ethikern niedergeschriebene Werte in harten Code umwandeln können (das war die erste Alternative), ist eine Lösung denkbar, die gerade diese Unbestimmtheit und Zweideutigkeit der wirklichen Welt berücksichtigt, also eine Lösung, die statistische Unsicherheit erträgt. Einfach ausgedrückt: Vielleicht kann ethisches Handeln maschinell gelernt werden, indem ein Modell das richtige Verhalten von den Daten lernt. Solche Daten müssen dann aus einem ethisch gelungenen System stammen, und der Lernprozess müsste ethisches Verhalten belohnen. Diese Lösung liegt sicher noch in weiter Zukunft, und die Forschung ist noch nicht so weit herauszufinden, ob das funktionieren könnte. Aber so oder so ist der Gedanke interessant und wird vielleicht ein Aspekt einer möglichen Zukunft mit ethischer künstlicher Intelligenz werden.

6.4 Stromkrieg und KI-Verordnung

Es gibt wenig, das ich noch lieber mag als hohe Berge, und wenn es einen Ort auf der Welt mit hohen Bergen gibt, dann ist es Nepal. Als ich das erste Mal Nepal besuchte, landete ich in der Hauptstadt Kathmandu und stieg dort in einem Hotel ab. Der Hotelbesitzer sagte mir, ich dürfe nichts selbst in die Steckdose stecken, sondern müsse jedes Mal einen der Angestellten rufen, beispielsweise, wenn ich mein Handy laden wollte. Zunächst dachte ich, das sei ein Scherz, aber die Erklärung zeigte, dass es sein voller Ernst war, weil das Stromnetz in Kathmandu so instabil ist, dass man einen Schlag bekommen und im schlimmsten Fall einen Arm verlieren kann, wenn man richtig Pech hat, sobald man einen Stecker in die Wand stecken will. Diesen makabren Ratschlag für Touristen bekam ich in der gleichen Woche, in der einer der damaligen Stars unter den KI-Berühmtheiten, Andrew Ng, seiner gesamten Follower-Gemeinde im Internet verkündete, dass »*KI die neue Elektrizität ist*«.[9] Doch was ich hörte, war sicher nicht das, was Ng mit seiner Aussage ursprünglich gemeint hatte. Während ich in meinem Hotelzimmer in Kathmandu saß, mich nach den Bergen sehnte und mich vor dem lebensgefährlichen Stromnetz gruselte, dachte ich: »Ja, vielleicht hast du ja recht. Künstliche Intelligenz ist eine gute Nachricht für uns, wenn wir die Technologie dazu bringen können, für uns zu arbeiten, aber sie ist geradezu lebensbedrohlich für diejenigen, die es nicht schaffen, sie auf gute Art und Weise zu benutzen, oder sich mit dem begnügen müssen, was sie bekommen.« In diesem Satz kann »künstliche Intelligenz« ersetzt werden durch »Elektrizität«, und das beschreibt das Gefühl, das ich hatte, als einer der Hotelangestellten mein Handyladegerät in die Wand einstöpselte.

Als Ng künstliche Intelligenz mit Elektrizität verglich, sagte er weiter: »Genau wie die Elektrizität vor hundert Jahren nahezu alles verändert hat, fällt es mir heute schwer, mir eine Branche vorzustellen, die KI nicht im Laufe der nächsten Jahre verändern wird.«[10] Ein Stück weit teile ich seine Meinung, und bei einem Treffen mit dem norwegischen Verteidigungsausschuss im Herbst 2022 lief ich Gefahr, die gleiche Metapher zu benutzen. Ich wurde gefragt, inwiefern ich glaubte, dass maschinelles Lernen die Verteidigung und Kriegsführung in der Zukunft beeinflussen würde, und antwortete, dass die Frage genauso viel Sinn macht, als hätte man Nikola Tesla 1880 gefragt, inwiefern Elektrizität die Kriegsführung in der Zukunft beeinflussen würde. Die Frage danach, was sich ändern wird, zu beantworten war unmöglich, nicht nur, weil niemand hätte voraussehen können, wie bedeutend Elektrizität für die moderne Gesellschaft werden würde, sondern auch, weil die richtige Antwort lautet: »Alles.« Den-

9 »AI is the new electricity.«

10 »Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years.«