Klaus Mainzer

Zwischen Autonomie und Unheimlichkeit: Blinde Flecken im *Machine Learning*

Abstract

Machine learning verändert unsere Zivilisation dramatisch. Wir verlassen uns immer mehr auf effiziente Algorithmen, weil die Komplexität unserer zivilisatorischen Infrastruktur sonst nicht zu bewältigen wäre: Unsere Gehirne sind zu langsam und bei den anstehenden Datenmengen hoffnungslos überfordert. Aber wie sicher sind KI-Algorithmen? Bei der praktischen Anwendung beziehen sich Lernalgorithmen auf Modelle neuronaler Netze, die selber äußerst komplex sind. Sie werden mit riesigen Datenmengen gefüttert und trainiert. Die Anzahl der dazu notwenigen Parameter explodiert exponentiell. Niemand weiß genau, was sich in diesen »Black Boxes« im Einzelnen abspielt. Es bleibt häufig ein statistisches Trial-and-Error Verfahren. Wie sollen aber Verantwortungsfragen z.B. beim autonomen Fahren oder in der Medizin entschieden werden, wenn die methodischen Grundlagen dunkel bleiben?

Machine Learning is changing our civilization dramatically. We rely increasingly on efficient algorithms, because the complex infrastructure of our civilization can no longer be managed: Our brains are too slow and hopelessly overtaxed with respect to Big Data. But how safe are AI-algorithms? In many practical applications, learning algorithms are related to models of neural networks which are extremely complex. They are fed and trained by huge sets of data. The number of required parameters is exploding exponentially. Nobody knows exactly, what is going on in these »black boxes« in all details. It is often a statistical trial- and error procedure. But how should questions of responsibility, e.g. in autonomous car driving or in medicine, be decided, when the methodological foundations remain dark?

1. Statistisches Lernen ist beschränkt

Die These des folgenden Beitrags lautet: Das derzeitige *Machine Learning* hat beängstigende dunkle Flecken. Im *Machine Learning* mit statistischen Lernalgorithmen benötigen wir daher mehr Erklärung (*explainability*) und Zurechnung (*accountability*) von Ursachen und Wirkungen, um ethische und rechtliche Fragen der Verantwortung entscheiden zu können!

Dazu werden zunächst die Grundlagen der statistischen Algorithmen untersucht, auf denen der derzeitige Hype der ›Künstlichen Intelligenz‹ (KI) beruht. Am 28. August 2018 meldete das CERN: »Mit der überragenden LHC (Large Hadron Collider) Leistung und modernem *Machine Learning* konnte das Higgs Boson bei seiner Kopplung mit schweren Fermionen identifiziert werden: Das erklärt, warum es Mas-

se im Universum gibt.«[1] Entdeckungen, die früher Wissenschaftler machten, scheint nun ›Künstliche Intelligenz‹ zu übernehmen.

Theoretisch sagt das Standardmodell der Teilchenphysik voraus, dass das Higgs-Boson in zwei Bottom-Quarks zerfällt, zusammen mit einem Boson, das in ein Elektron und ein Anti-Elektron zerfällt. Dieses Ereignis muss unter Milliarden von Daten identifiziert werden, die durch Protonen-Kollisionen im LHC erzeugt werden. Mustererkennung unter Milliarden von Daten ist nur noch durch leistungsfähige Algorithmen und Hochleistungsrechner möglich. In diesem Sinn verändert KI tatsächlich bereits die Wissenschaft. Am Beispiel der Teilchenphysik lässt sich der Grundgedanke der Mustererkennung in der Teilchenphysik durch *Machine Learning* illustrieren: Signalereignisse wie z.B. der Zerfall des Higgs-Bosons müssen von den Hintergrundereignissen der vielen anderen Daten der Protonenkollisionen unterschieden werden. Dazu lernen Algorithmen des *Machine Learning*, ein Verteilungsmuster von Signalereignissen (›Fingerabdruck der Higgs-Teilchen‹) in den Big Data der Hintergrundereignisse durch Minimierung von Fehlerquoten zu identifizieren.

Ebenso wird *Machine Learning* bereits bei der Arzneimittelentwicklung eingesetzt. Gewöhnlich dauert eine Arzneimittelentwicklung ca. 10–15 Jahre, also solange wie das technologische Großprojekt eines Flugzeugs. *Machine Learning* und Big Data senken die Entwicklungszeit drastisch: Dazu liest z.B. die KI-Software *IBM Watson for Drug Discovery*[2] zunächst Millionen von Seiten (*Big Data mining*), um ihre Bedeutung für Forschungsziele (*target identification and validation*) zu erkennen. In wenigen Monaten wurden so fünf RNA-bindende Proteine (RBPs) entdeckt, die zuvor nie mit amyotropher Lateralsklerose (ALS) in Verbindung gebracht wurden.[3] Die Firma GlaxoSmithKline berichtet von einer Reduktion der Zeit der Molekülerkennung zur Krankheitsbekämpfung von fünfeinhalb auf ein Jahr (verbunden mit entsprechender Kostensenkung).[4]

Machine Learning liefert auch bereits bessere medizinische Diagnosen. Wie in der Teilchenphysik unterstützt *Machine Learning* dazu Mustererkennung in komple-

1 Cern: »Long-sought decay of Higgs boson observed«, Pressemitteilung, 28.8.2018, https://home.cern/news/press-release/physics/long-sought-decay-higgs-boson-observed (aufgerufen: 7.3.2019). Albert M. Sirunyan u.a. (CMS Collaboration): »Observation of Higgs boson decay to bottom quarks«, in: *Physical Review Letters* 121 (2018), Heft 12, doi: 10.1103/PhysRevLett.121.121801 (aufgerufen 29.10.2019).

2 Vgl. »IBM Watson for Drug Discovery«, in: *International Business Machines: Marketplace*, https://www.ibm.com/de-de/marketplace/ibm-watson-for-drug-discovery/purchase (aufgerufen: 7.3.2019).

3 Vgl. »Neue Auslöser für eine schwere Krankheit«, in: *Julius-Maximilians-Universität Würzburg: Pressemitteilungen*, 25.7.2016, https://www.uni-wuerzburg.de/aktuelles/pressemitteilungen/single/news/neue-ausloeser-fuer-eine-schwere-krankheit/ (aufgerufen 7.3.2019).

4 Vgl. John Leonard: »GSK´s chief data officer reveals how machine learning has been used to speed up drug discovery«, in: *computing: Big Data and Analytics*, https://www.computing.co.uk/ctg/interview/3027675/interview-gsks-chief-data-officer-on-machine-learning-to-speed-up-drug-discovery-and-changes-in-the-cdo-role (aufgerufen 7.3.2019),.

xen Daten: In Gewebeschnitten von z.B. Brustkrebs werden normale Lymphknoten (Hintergrundereignisse) von Krebszellen (Signalereignisse) unterschieden. Die Pathologie in Harvard verbesserte so die Genauigkeit von 96% auf 99,5 %. *IBM Watson for Genomics* bestätigte in 1018 Fällen die ärztliche Diagnose in mehr als 99% und entdeckte zusätzliche genomische Ereignisse von signifikanter Bedeutung.[5]

Methodisch handelt es sich beim statistischen Lernen um ein induktiv-statistisches Verfahren. Korrelationen und Datenmuster sollen aus endlich vielen Daten durch Algorithmen abgeleitet werden. Dazu kann ein naturwissenschaftliches Experiment angenommen werden, bei dem in einer Serie von veränderten Bedingungen (Inputs) entsprechende Ergebnisse (Outputs) folgen. In der Medizin könnte es sich um einen Patienten handeln, der auf Medikamente in bestimmter Weise reagiert. Dabei wird angenommen, dass die entsprechenden Paare von Input- und Outputdaten unabhängig durch dasselbe unbekannte Zufallsexperiment erzeugt werden. Statistisch sagt man deshalb, dass die endliche Folge von Beobachtungsdaten (x_1, y_1), …, (x_n, y_n) mit Inputs x_i und Outputs y_i $(i = 1, \ldots, n)$ durch Paare von Zufallsvariablen (X_1, Y_1), …, (X_n, Y_n) realisiert wird, denen eine unbekannte Wahrscheinlichkeitsverteilung $P_{X,Y}$ zugrunde liegt.

Algorithmen sollen nun Eigenschaften der Wahrscheinlichkeitsverteilung $P_{X,Y}$ ableiten. Ein Beispiel wäre die Erwartungswahrscheinlichkeit, mit der für einen gegebenen Input ein entsprechender Output auftritt. Es kann sich aber auch um eine Klassifikationsaufgabe handeln: Eine Datenmenge soll auf zwei Klassen aufgeteilt werden. Mit welcher Wahrscheinlichkeit gehört ein Element der Datenmenge (Input) eher zu der einen oder anderen Klasse (Output)? In diesem Fall spricht man von binärer Mustererkennung.

Beim Erkennen eines binären Musters werden die Daten einer Datenmenge X auf zwei mögliche Klassen verteilt, die mit $+1$ bzw. -1 bezeichnet werden. Diese Zuordnung wird durch eine Funktion $f\colon \mathrm{X} \to \mathrm{Y}$ mit $\mathrm{Y} = \{+1, -1\}$ beschrieben. Beim statistischen Lernen eines binären Musters geht es darum, aus einer Klasse F von Funktionen diejenige Zuordnung f zu ermitteln, bei der die Fehlerabweichung bzw. der erwartete Irrtum minimal ist. Daher spricht man auch von der Risikominimierung $R[f]$ der Klassifikationsfunktionen f durch statistisches Lernen:[6]

$$R[f] = \int \frac{1}{2} |f(x) - y| \, dP_{X,Y}\Big(x, y\Big)$$

Da aber die Wahrscheinlichkeitsverteilung $P_{X,Y}$ für alle Werte unbekannt ist, kann diese Formel und damit die gesuchte Mustererkennung mit minimaler Fehlerabwei-

5 Vgl. »Artifical Intelligence in medicine«, in: International Business Maschines; IBM Watson Health, https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine (aufgerufen 7.3.2019).

6 Vgl. Jonas Peters u.a.: *Elements of Causal Inference. Foundations and Learning Algorithms*, Cambridge (Mass.) 2017, S. 6f.

chung nicht berechnet werden. Es stehen nur die endlich vielen empirisch beobachteten Zuordnungen (x_1, y_1), …, (x_n, y_n) zur Verfügung. Daher beschränkt man sich auf eine empirische Risikominimierung. Schrittweise wird dazu für jede Zuordnungsfunktion f der Klasse F der empirischen Trainingsirrtum beim Lernen aus einem endlichen Datensatz mit Umfang n bestimmt:

$$R^n_{emp}[f] = \frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{1}{2}|f(x_i) - y_i|$$

Dadurch wird eine Folge von Funktionen der Klasse F mit verbessertem Trainingsirrtum erzeugt. Die zentrale Frage ist, ob durch dieses Verfahren eine Mustererkennung mit einer minimal möglichen Fehlerabweichung ermittelt werden kann. Mathematisch stellt sich die Frage, ob die so ermittelte Funktionenfolge in der Klasse F gegen eine Funktion mit minimaler Fehlerabweichung konvergiert.

Tatsächlich lässt sich beweisen, dass eine solche Konvergenz bzw. ein solcher Lernerfolg nur für kleine Teilklassen garantiert ist. Ein Beispiel ist die Vapnik-Chervonenkis (VC) Dimension, mit der sich die Kapazität und Größe solcher Funktionenklassen bestimmen lässt.[7] Mit großer Wahrscheinlichkeit ist dann das Risiko nicht größer als das empirische Risiko (plus einem Term, der mit der Größe der Funktionenklasse wächst.)

Die derzeitigen Erfolge des *Machine Learning* scheinen die These zu bestätigen, dass es auf möglichst große Datenmengen ankommt, die mit immer stärkerer Computerpower bearbeitet werden. Die so erkannten Regularitäten hängen von der Wahrscheinlichkeitsverteilung der statistischen Daten ab. Statistisches Lernen versucht also, ein probabilistisches Modell aus endlich vielen Daten von Ergebnissen (z.B. Zufallsexperimenten) und Beobachtungen abzuleiten. Umgekehrt versucht statistisches Schließen, Eigenschaften von beobachteten Daten aus einem angenommenen statistischen Modell abzuleiten.

2. Kausales Lernen und Kausalmodelle

Datenkorrelationen können Hinweise auf Sachverhalte liefern, müssen es aber nicht. Dazu stelle man sich eine Testreihe vor, bei der sich eine günstige Korrelation zwischen einer verabreichten chemischen Substanz und der Bekämpfung bestimmter Krebstumore ergibt. Dann entsteht Druck des betroffenen Unternehmens, mit einem entsprechenden Medikament in die Produktion zu gehen und Gewinne abzuschöpfen. Aber auch betroffene Patienten mögen darin ihre letzte Chance sehen. Tatsächlich erhält man ein nachhaltiges Medikament aber nur, wenn der zugrundeliegende

7 Vgl. Vladimir N. Vapnik: *Statistical Learning Theory*, New York 1998.

kausale Mechanismus des Tumorwachstums, also die Gesetze der Zellbiologie und Biochemie, verstanden sind.

Bereits Newton war kaum an statistischen Datenkorrelationen der fallenden Äpfel von den Apfelbäumen seines väterlichen Bauernhofs interessiert, obwohl diese Geschichte immer wieder erzählt wird. Sein Ziel war das zugrundeliegende mathematische Kausalgesetz der Gravitation, mit dem genaue Erklärungen und Prognosen der fallenden Äpfel und der Himmelskörper möglich wurden, letztlich auch die darauf aufbauende Satelliten- und Raketentechnik von heute.

Statistisches Lernen und Schließen aus Daten reichen also nicht aus. Man muss vielmehr die kausalen Zusammenhänge von Ursachen und Wirkungen hinter den Messdaten erkennen. Diese kausalen Zusammenhänge hängen von den Gesetzen der jeweiligen Anwendungsdomäne der jeweiligen Forschungsmethoden ab, also den Gesetzen der Physik im Beispiel von Newton, den Gesetzen der Biochemie und des Zellwachstums im Beispiel der Krebsforschung, etc. Wäre es anders, könnte man bereits mit den Methoden des statistischen Lernens und Schließens die Probleme dieser Welt lösen. Tatsächlich scheinen das einige kurzsichtige Zeitgenossen beim derzeitigen Hype der ›Künstlichen Intelligenz‹ zu glauben.

Statistisches Lernen und Schließen ohne kausales Domänenwissen ist blind – bei noch so großer Datenmenge (Big Data) und Rechenpower! Dieses aktuelle Problem des *Machine Learning* hat einen erkenntnistheoretischen Hintergrund in der Philosophie des 18. Jahrhunderts. Nach David Hume beruht alle Erkenntnis auf sinnlichen Eindrücken (Daten), die psychologisch »assoziiert«[8] werden. Es gibt danach keine Kausalitätsgesetze von Ursache und Wirkung, sondern nur Assoziationen von Eindrücken (z.B. Blitz und Donner), die mit (statistischer) Häufigkeit »gewohnheitsmäßig«[9] korreliert werden. Demgegenüber sind nach Immanuel Kant Kausalitätsgesetze, als vernunftmäßig gebildete Hypothesen, notwendig, die experimentell überprüft werden müssen. Ihre Bildung beruht nicht auf psychologischen Assoziationen, sondern auf einer vernunftmäßigen Kategorie, die mithilfe der Einbildungskraft in Vorhersagen für die Erfahrung operationalisiert werden kann.[10] In heutiger Forschung werden Kausalmodelle angenommen. Nach Kant ist dieses Verfahren seit Galileo Galilei in der Physik in Gebrauch, die so erst zur Wissenschaft wurde. Jedenfalls reichen Datenkorrelationen (auch mit Big Data) für (wissenschaftliche) Erklärungen nicht aus!

Evolutionsbiologisch ist bemerkenswert, dass bereits Tiere bei entsprechender sensorieller und neuronaler Ausstattung zu statistischem Lernen und Schließen fähig sind. Organismen, die gehäuft bestimmte Gefahrensituationen erleben, werden sie in

8 David Hume: *Eine Untersuchung über den menschlichen Verstand*, übers. u. hrsg. Herbert Herring, Stuttgart 1976, S. 81.
9 Ebd.
10 Vgl. Immanuel Kant: *Kritik der reinen Vernunft*, hrsg. R. Schmidt, F. Meiner: Hamburg 1956, B 233.

Zukunft meiden und ihre Konsequenzen für zukünftiges Verhalten daraus ziehen. Dazu bedarf es keiner formalen Repräsentanz in statistischen Formeln wie bei uns Menschen. Formeln lassen sich aber in Algorithmen übersetzen, mit denen z.B. Roboter programmiert werden können. Diese Form der schwachen ›Künstlichen Intelligenz‹ ist heute technisch gut erprobt.

Neben der Statistik der Daten bedarf es aber zusätzlicher Gesetzes- und Strukturannahmen der Anwendungsdomänen, die durch Experimente und Interventionen überprüft werden. Kausale Erklärungsmodelle (z.B. das Planetenmodell oder ein Tumormodell) erfüllen die Gesetzes- und Strukturannahmen einer Theorie (z.B. Newtons Gravitationstheorie oder die Gesetze der Zellbiologie):

Beim kausalen Schließen werden Eigenschaften von Daten und Beobachtungen aus Kausalmodellen, d.h. Gesetzesannahmen von Ursachen und Wirkungen, abgeleitet. Kausales Schließen ermöglicht damit, die Wirkungen von Interventionen oder Datenveränderungen (z.B. durch Experimente) vorauszusagen. Kausales Lernen versucht umgekehrt, ein Kausalmodell aus Beobachtungen, Messdaten und Interventionen (z.B. Experimenten) abzuleiten, die zusätzliche Gesetzes- und Strukturannahmen voraussetzen. Kausales Lernen und Schließen aus selbst aufgestellten Kausalmodellen wären die ersten Schritte zu einer starken ›Künstlichen Intelligenz‹, die bei Tieren nicht beobachtbar sind, aber bei Menschen.

Um entsprechende Software zu entwickeln, muss kausales Lernen und Schließen zunächst formal erschlossen und in Algorithmen übersetzt werden. Formal besteht ein strukturelles Kausalmodell aus einem System von strukturellen Zuordnungen von Ursachen zu Wirkungen mit eventuellen Störvariablen. Ursachen und Wirkungen werden durch Zufallsvariablen beschrieben. Ihre funktionalen Zuordnungen (unter Berücksichtigung von Störvariablen) werden durch Gleichungen definiert, also z.B. Wirkung $X_j = f(X_i, N)$ in funktioneller Abhängigkeit von Ursache X_i und Störvariable N. Anschaulich kann das Netzwerk der Ursachen und Wirkungen durch einen Graphen von Knoten und Kanten dargestellt werden. Zufallsvariablen von Ursachen und Wirkungen entsprechen Knoten. Kausale Wirkungen entsprechen gerichteten Pfeilen: $X_i \rightarrow X_j$ bedeutet, dass Ursache X_i Wirkung X_j auslöst.

Es lässt sich beweisen, dass ein Kausalmodell eine eindeutige Wahrscheinlichkeitsverteilung der Daten einschließt, aber nicht umgekehrt: Für Kausalmodelle (z.B. ein Planetenmodell) müssen zusätzliche Gesetze (z.B. ein Gravitationsgesetz) angenommen werden.[11] Um kausale Abhängigkeiten von Ereignissen zu erkennen, muss die Unabhängigkeit der sie darstellenden Zufallsvariablen ermittelt werden. Statistisch lässt sich die Unabhängigkeit der Resultate x und y zweier Zufallsvariablen (anschaulich Zufallsexperimente) X und Y dadurch ausdrücken, dass ihre Ver-

11 Vgl. Joris M. Mooij u.a.: »From ordinary differential equations to structural causal models: The deterministic case«, in: *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence* (UAI) (2013), S. 440–448.

bundwahrscheinlichkeit $p(x, y)$ faktorisierbar ist, d.h. $p(x, y) = p(x)p(y)$. Man spricht in diesem Fall auch von der Markov-Bedingung. Auf dieser Grundlage lässt sich der Kalkül einer kausalen Unabhängigkeitsrelation $\perp\!\!\!\perp$ einführen:[12]

Sei $p(x)$ die Dichte der Wahrscheinlichkeitsverteilung P_X einer Zufallsvariablen X. Die Zufallsvariable X heißt unabhängig von der Zufallsvariablen Y (formal $X \perp\!\!\!\perp Y$) genau dann, wenn die Verbundwahrscheinlichkeit $p(x, y)$ für alle Werte x, y von X, Y faktorisierbar ist, also $p(x, y) = p(x)p(y)$ gilt. Allgemein heißen die Zufallsvariablen $X_1, \ldots, X_d$ gegenseitig unabhängig genau dann, wenn $p(x_1, \ldots, x_d) = p(x_1) \bullet \ \ldots \ \bullet p(x_d)$ für alle Werte $x_1, \ldots, x_d$ von $X_1, \ldots, X_d$ gilt.

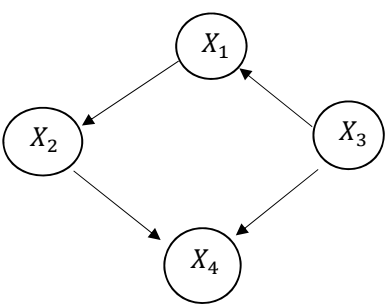
Häufig hängt die Unabhängigkeit von Ereignissen von Situations- und Umweltbedingungen ab. Wie üblich in Statistik und Wahrscheinlichkeitstheorie, werden solche Abhängigkeiten durch bedingte Wahrscheinlichkeiten zum Ausdruck gebracht, also z.B. $p(x|z)$ als Wahrscheinlichkeit von Ereignis x unter Voraussetzung von Ereignis z. Die Zufallsvariable X heißt unabhängig von Y unter Bedingung Z (formal $X \perp\!\!\!\perp Y|Z$) genau dann, wenn $p(x, y|z) = p(x|z)p(y|z)$ für alle Werte x, y, z von X, Y, Z mit $p(z) > 0$.

Für die bedingte Unabhängigkeitsrelation lassen sich folgende Regeln beweisen:

$$X \perp\!\!\!\perp Y|Z \Rightarrow Y \perp\!\!\!\perp X|Z \qquad \text{(Symmetrie)}$$
$$X \perp\!\!\!\perp Y,\ W|Z \Rightarrow X \perp\!\!\!\perp Y|Z \qquad \text{(Dekomposition)}$$
$$X \perp\!\!\!\perp Y,\ W|Z \Rightarrow X \perp\!\!\!\perp Y|W, Z \qquad \text{(schwache Vereinigung)}$$
$$X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|Y, Z \Rightarrow X \perp\!\!\!\perp Y,\ W|Z \qquad \text{(Kontraktion)}$$
$$X \perp\!\!\!\perp Y|W, Z \text{ and } X \perp\!\!\!\perp W|Y, Z \Rightarrow X \perp\!\!\!\perp Y,\ W|Z \qquad \text{(Schnittmenge)}$$

Ein kausales Strukturmodell lässt sich also durch funktionale Gleichungen und graphisch durch Kausalnetze darstellen wie im folgenden Beispiel:[13]

$X_1 := f_1(N_1)$
$X_2 := f_2(N_2)$
$X_3 := f_3(X_1, N_3)$
$X_4 := f_4(X_2, X_3, N_4)$
mit N_1, N_2, N_3, N_4
unabhängige Störvariablen



12 Vgl. Judea Pearl: *Causality. Models, Reasoning, and Inference*, Cambridge (Mass.) 2009.
13 Peters u.a.: *Elements of Causal Inference*, S. 84.

In diesem Beispiel entspricht die Unabhängigkeit der Zufallsvariablen X_1, X_2, X_3, X_4 in der statistischen Verteilung P_{X_1, X_2, X_3, X_4} formal den bedingten Unabhängigkeitsrelationen $X_2 \perp\!\!\!\perp X_3 | X_1$ und $X_1 \perp\!\!\!\perp X_4 | X_2, X_3$ bzw. der Markov-Faktorisierung $p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1|x_3)p(x_2|x_1)p(x_4|x_2, x_3)$.

Ziel des kausalen Lernens ist es also, hinter der Verteilung von Mess- und Beobachtungsdaten die kausalen Abhängigkeiten von Ursachen und Wirkungen zu entdecken. Ausgangssituation ist ein endlicher Datensatz einer Datenerhebung. Dazu wird eine Verbundwahrscheinlichkeit von unabhängig und identisch verteilten Zufallsvariablen vorausgesetzt. Durch Unabhängigkeitstests und Experimente lassen sich daraus Kausalmodelle ableiten, die durch Unabhängigkeitsrelationen bzw. wahrscheinlichkeitstheoretische Faktorisierung oder Kausalgesetze bestimmt sind. Auf der Grundlage solcher Kausalmodelle lassen sich die Abhängigkeiten von Ursachen und Wirkungen auch graphisch darstellen. Damit wird die eingangs geforderte Zuordnung (*accountability*) von Ursachen und Wirkungen erst möglich, die zur Klärung von Verantwortungsfragen (*responsibility*) notwendig ist.

3. Von der statistischen Blindheit zur zertifizierten KI-Software

Die neuronalen Netze, die beim *Machine Learning* praktisch angewendet werden, müssen über eine große Anzahl von Knoten verfügen. Die Anzahl möglicher Kausalmodelle (mit entsprechender graphischer Darstellung) steigt damit exponentiell:[14]

14 *The On-line Encyclopedia of Integer Sequences:* http://oeis.org/A003024.2017 (aufgerufen: 29.10.2019).

d	Anzahl der Kausalmodelle mit d Knoten
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	8375667077373332018769930304799641222352231383 03
	...

Wegen dieser Explosion von Parametern führt die Komplexität praktischer Anwendungen zu einer dramatischen Herausforderung des *Machine Learning*, die häufig unterschätzt wird. In der Gehirnforschung haben wir es mit einem der komplexesten neuronalen Netze zu tun, das in der Evolution entstand. Im mathematischen Modell nehmen wir vereinfachend einen Vektor z an, mit dem die Aktivität einer großen Anzahl von Gehirnregionen kodiert wird. Die Dynamik (d.h. die zeitliche Entwicklung) von z wird bestimmt durch eine zeitabhängige Differentialgleichung

$$\frac{d}{dt}z = F\Big(z, u, \theta\Big)$$

mit Funktion F, Vektor u der externen Stimulationen und Parameter θ der kausalen Verbindungen.[15]

Die Gehirnaktivität z kann aber nicht direkt beobachtet werden. *Functional resonance imaging* (fMRI) bestimmt nur den Verbrauch an Nährstoffen (Sauerstoff und Glukose) zur Kompensation des gestiegenen Energiebedarfs, der durch Blutfluss geliefert wird (hämodynamische Antwort). Das Anwachsen wird durch das *blood-oxygen-level-dependent* (BOLD)-Signal bestimmt. Daher muss z im dynamischen Kausalmodell durch eine Zustandsvariable x ersetzt werden, in der die Gehirnaktivität mit der hämodynamischen Antwort berücksichtigt wird:

15 Vgl. Karl Friston: »Dynamic causal modelling«, in: *NeuroImage* 19 (2003), Heft 4, S. 1273–1302.

$$\frac{d}{dt}x = F\big(x, u, \theta\big).$$

Dazu wird die gemessene Zeitreihe des BOLD-Signals $y = \lambda(x)$ mit der Zustandsvariablen x verbunden.

Tatsächlich haben wir es beim menschlichen Gehirn mit einer Datenflut zu tun, die durch 86 Milliarden Neuronen hervorgebracht wird. Wie im Einzelnen die kausalen Wechselwirkungen hinter diesen Datenwolken ablaufen, bleibt vorläufig weiterhin eine Black Box. Statistisches Lernen aus gemessenen Daten reicht aber auch im Zeitalter von Big Data und wachsender Rechenpower nicht aus. Mehr Erklärung der kausalen Wechselwirkungen zwischen den einzelnen Gehirnregionen, also kausales Lernen, ist eine zentrale Herausforderung der Gehirnforschung, um bessere medizinische Diagnosen, psychologische und rechtliche Zurechnungsfähigkeit zu erhalten.[16]

Ein hochaktuelles technisches Beispiel für die wachsende Komplexität neuronaler Netze sind selbstlernende Fahrzeuge. So kann ein einfaches Automobil mit verschiedenen Sensoren (z.B. Nachbarschaft, Licht, Kollision) und motorischer Ausstattung bereits komplexes Verhalten durch ein sich selbstorganisierendes neuronales Netzwerk erzeugen. Werden benachbarte Sensoren bei einer Kollision mit einem äußeren Gegenstand erregt, dann auch die mit den Sensoren verbundenen Neuronen eines entsprechenden neuronalen Netzes. So entsteht im neuronalen Netz ein Verschaltungsmuster, das den äußeren Gegenstand repräsentiert. Im Prinzip ist dieser Vorgang ähnlich wie bei der Wahrnehmung eines äußeren Gegenstands durch einen Organismus – nur dort sehr viel komplexer.

Wenn wir uns nun noch vorstellen, dass dieses Automobil mit einem ›Gedächtnis‹ (Datenbank) ausgestattet wird, mit dem es sich solche gefährlichen Kollisionen merken kann, um sie in Zukunft zu vermeiden, dann ahnt man, wie die Automobilindustrie in Zukunft unterwegs sein wird, selbstlernende Fahrzeuge zu bauen. Sie werden sich erheblich von den herkömmlichen Fahrerassistenzsystemen mit vorprogrammiertem Verhalten unter bestimmten Bedingungen unterscheiden. Es wird sich um ein neuronales Lernen handeln, wie wir es in der Natur auch von höher entwickelten Organismen kennen.

Wie viele reale Unfälle sind aber erforderlich, um selbstlernende (›autonome‹) Fahrzeuge zu trainieren? Wer ist verantwortlich, wenn autonome Fahrzeuge in Unfälle verwickelt sind? Welche ethischen und rechtlichen Herausforderungen stellen sich? Bei komplexen Systemen wie neuronalen Netzen mit z.B. Millionen von Elementen und Milliarden von synaptischen Verbindungen erlauben zwar die Gesetze der statistischen Physik, globale Aussagen über Trend- und Konvergenzverhalten des gesamten Systems zu machen. Die Zahl der empirischen Parameter der einzel-

16 Vgl. Gabriele Lohmann u.a.: »Critical comments on dynamic causal modelling«, in: *NeuroImage* 59 (2012), Heft 3, S. 2322–2329.

nen Elemente ist jedoch unter Umständen so groß, dass keine lokalen Ursachen ausgemacht werden können. Das neuronale Netz bleibt für uns eine Black Box. Vom ingenieurwissenschaftlichen Standpunkt aus sprechen Autoren daher von einem ›dunklen Geheimnis‹ im Zentrum der KI des *Machine Learning*: »[…] *even the engineers who designed [the machine learning-based system] may struggle to isolate the reason for any single action*«.[17]

Zwei verschiedene Ansätze im *Software Engineering* sind denkbar:

1. Testen zeigt nur (zufällig) gefundene Fehler, aber nicht alle anderen möglichen.
2. Zur grundsätzlichen Vermeidung müsste eine formale Verifikation des neuronalen Netzes und seiner zugrundeliegenden kausalen Abläufe durchgeführt werden.

Die Idee, formale Verifikationen zu automatisieren, ist alt und geht auf die Anfänge der ›Künstlichen Intelligenz‹ zurück. Damals, in der ersten Phase der KI-Forschung, war die Suche nach allgemeinen Problemlösungsverfahren wenigstens in der formalen Logik erfolgreich. Dort wurde ein mechanisches Verfahren angegeben, um die logische Allgemeingültigkeit von Formeln zu beweisen. Das Verfahren konnte auch von einem Computerprogramm ausgeführt werden und leitete in der Informatik das automatische Beweisen ein.

Der Beweis für die Allgemeingültigkeit eines logischen Schlusses kann in der Praxis sehr kompliziert sein. Daher schlug John Alan Robinson 1965 die sogenannte ›Resolutionsmethode‹ vor, nach der Beweise nach dem Muster logischer Widerlegungsverfahren gefunden werden können.[18] Man startet also mit der Annahme des Gegenteils (Negation), d.h. der logische Schluss sei nicht allgemeingültig. Im nächsten Schritt wird dann gezeigt, dass alle möglichen Anwendungsbeispiele dieser Annahme zu einem Widerspruch führen. Es gilt also das Gegenteil der Negation und der logische Schluss ist allgemeingültig. Die Robinson'sche Resolutionsmethode benutzt logische Vereinfachungen, wonach man jede logische Formel in eine sogenannte konjunktive Normalform umwandeln kann. In der Aussagenlogik besteht eine konjunktive Normalform aus negierten und nicht negierten Aussagenvariablen (Literale), die durch Konjunktion ($\wedge$) und Disjunktion ($\vee$) verbunden sind.

Mechanisch besteht das Verfahren also darin, widersprechende Teilaussagen aus Konjunktionsgliedern einer logischen Formel zu streichen (›Resolution‹) und diese

17 Will Knight: »The Dark Secret at the Heart of AI. No one really knows how the most advanced algorithms do what they do. That could be a problem«, in: *MIT Technology Review* (März/April 2017), S. 1–22.

18 Vgl. John Alan Robinson: »A machine oriented logic based on the resolution principle«, in: *Journal of the Association for Computing Machinery* (JACM) 12 (1965), Heft 1, S. 23–41. Michael M. Richter: *Logikkalküle*, Stuttgart 1978, S. 185 ff. Uwe Schöning: *Logik für Informatiker*, Mannheim 1987, S. 85 ff.

Prozedur mit den entstehenden ›Resolventen‹ und einem entsprechenden anderen Konjunktionsglied der Formel zu wiederholen, bis ein Widerspruch (das ›leere‹ Wort) abgeleitet werden kann.

In einem entsprechenden Computerprogramm terminiert dieses Verfahren für die Aussagenlogik. Es zeigt also in endlicher Zeit, ob die vorgelegte logische Formel allgemeingültig ist. Allerdings wächst die Rechenzeit nach den bisher bekannten Verfahren exponentiell mit der Anzahl der Literale einer Formel. Was nun die ›Künstliche Intelligenz‹ betrifft, so können Computerprogramme mit der Resolutionsmethode automatisch über die Allgemeingültigkeit logischer Schlüsse, wenigstens in der Aussagenlogik, im Prinzip entscheiden. Menschen hätten große Schwierigkeiten, bei komplizierten und langen Schlüssen den Überblick zu bewahren. Zudem sind Menschen wesentlich langsamer. Mit steigender Rechenkapazität können Maschinen also wesentlich effizienter diese Aufgabe des logischen Schließens erledigen.

Für die Formeln der Prädikatenlogik lässt sich ebenfalls ein verallgemeinertes Resolutionsverfahren angeben, um wieder aus der Annahme der allgemeinen Ungültigkeit einer Formel einen Widerspruch abzuleiten. Dazu muss eine Formel der Prädikatenlogik in eine Normalform gebracht werden, aus deren Klauseln sich mechanisch ein Widerspruch ableiten lässt. Da aber in der Prädikatenlogik (im Unterschied zur Aussagenlogik) nicht allgemein die Allgemeingültigkeit einer Formel entschieden werden kann, kann es vorkommen, dass das Resolutionsverfahren kein Ende findet. Das Computerprogramm läuft dann unbegrenzt weiter. Es kommt dann darauf an, Teilklassen zu finden, in denen das Verfahren nicht nur effizient ist, sondern überhaupt terminiert. Maschinelle Intelligenz kann zwar die Effizienz von Entscheidungsprozessen steigern und sie beschleunigen. Sie ist aber auch (wie menschliche Intelligenz) den prinzipiellen Grenzen logischer Entscheidbarkeit unterworfen.

In Logik und Mathematik werden Formeln (also Zeichenreihen) Schritt für Schritt abgeleitet, bis der Beweis einer Behauptung abgeschlossen ist. Computerprogramme arbeiten im Grunde wie Beweise. Schritt für Schritt leiten sie nach festgelegten Regeln Zeichenfolgen ab, bis ein formaler Ausdruck gefunden ist, der für eine Lösung des Problems steht. Stellen wir uns z.B. die Montage eines Werkstücks auf einem Fließband vor. Das entsprechende Computerprogramm beschreibt, wie das Werkstück Schritt für Schritt aus vorausgesetzten Einzelteilen nach Regeln aufeinander aufbauend entsteht.

Ein Kunde wünscht von einem Informatiker ein Computerprogramm, das ein solches Problem löst. Bei einem sehr komplexen und unübersichtlichen Produktionsprozess möchte er aber vorher einen Beweis, dass das Programm auch korrekt arbeitet. Man spricht in diesem Fall von einem zertifizierten Programm. Eventuelle Fehler wären gefährlich oder würden erhebliche Mehrkosten verursachen. Dramatische Beispiele für nicht korrekt funktionierende Programme waren in der Vergangenheit

automatische Steuerungen von Bestrahlungsmaschinen in der Medizin, die tödlichen Überdosierungen verursachten. Softwarefehler in Gepäcktransportsystemen lösten auf Flughäfen Millionen Dollar Verluste aus. Nach dem Ausfall von Kommunikationssystemen bestand plötzlich kein Kontakt mehr zu Hunderten von Flugzeugen. Ein Schock war 1997 der Crash der Ariane-5-Rakete, der auf einen versteckten Konvertierungsfehler zurückzuführen war.

Unter einer ›Programmverifikation‹ wird das Erbringen eines Beweises verstanden, dass ein gegebenes Programm sich an eine gegebene Spezifikation hält. Nach K. Gödel kann es allerdings keinen allgemeinen Algorithmus geben, der für beliebige Programme ihre Korrektheit entscheidet. Daher gibt es verschiedene Techniken zur Verifikation von Teilklassen bestimmter Programme. Eine Verifikation soll dabei nicht denselben Fehlerquellen unterliegen, wie die Programmierung. In der Praxis benutzt man häufig nur Tools, mit denen Programme modellbasiert simuliert und getestet werden. Solche Simulatoren und Tests sind leicht und schnell durchzuführen und liefern Gegenbeispiele. Sie sind allerdings nicht geeignet, um subtile Fehler zu erkennen, die Güte des Systems zu bewerten und die Abwesenheit von Fehlern prinzipiell zu belegen.

Das leisten formale ›Theorembeweiser‹, in denen Computerprogramme auf Algorithmen zurückgeführt werden, die logisch-mathematischen Beweisen entsprechen.[19] Das Ziel ist also, dass Computerprogramme so sicher sein sollen wie Beweise in Logik und Mathematik.[20] Umgekehrt sollen Beweise in Logik und Mathematik als konstruktive Algorithmen aufgefasst werden. Danach entspricht z.B. ein Beweis eines Schlusses der Aussage B aus der Aussage A (formal: $A \rightarrow B$) einem konstruktiven Verfahren, das einem Objekt vom Typ A ein Objekt vom Typ B zuordnet. Wie im funktionalen Programmieren werden Funktionen als konstruktive Verfahren aufgefasst, die Programmen entsprechen. Man sagt, dass ein Programm zum Beweis z.B. der Aussage $A \rightarrow B$ vom Typ $A \rightarrow B$ sei. Allgemein entsprechen ein Programm und sein Typ einem Beweis und der Aussage, die dieser Beweis beweist (*Curry-Howard Correspondence*).[21]

Programme und Beweise lassen sich nun im selben Formalismus beschreiben, der auf den (typisierten) λ-Kalkül von Alonzo Church zurückgeht. Der Beweisassistent *Coq* (nach dem französischen Logiker und Informatiker Thierry Coquand) baut darauf mit seinem Formalismus *CoC* (*Calculus of Constructions*) auf.[22] Wie in der

19 Vgl. Freek Wiedijk (Hg.): *The Seventeen Provers of the World*, Berlin 2006.
20 Vgl. Klaus Mainzer u.a. (Hg.): *Proof and Computation. Digitization in Mathematics, Computer Science, and Philosophy*, Singapore 2018.
21 Vgl. William A. Howard: »The formulae-as-types notion of construction«, in: J.P. Seldin und J.R. Hindley (Hg.): *To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*, Boston, MA 1969, S. 479–490.
22 Vgl. Yves Bertot und Pierre Castéran: *Interactive Theorem Proving and Program Development. Coq'Art: The Calculus of Inductive Constructions*, Berlin 2004.

konstruktiven Logik und Mathematik werden in *CoC* alle Terme und Formeln garantiert zirkel- und widerspruchsfrei konstruiert. Wenn also ein Programm in diesen Formalismus übersetzt werden kann, ist seine Verifikation von vornherein gesichert. Die Verifikation eines Programms entspricht dann einem Verifikationsalgorithmus für den Typ dieses Programms im Kalkül *CoC*. In der Erweiterung *CiC* (*Calculus of inductive Constructions*) können auch sehr abstrakte und unendliche Datenstrukturen formalisiert werden. Umgekehrt lassen sich damit auch sehr abstrakte Beweise in der Mathematik durch Computerprogramme verifizieren. *CiC* eignet sich also sowohl zur Begründung von Beweisen in der Mathematik als auch zur Verifikation von Programmen in der Informatik.[23]

Der Beweisassistent *Coq*, der den Formalismus *CiC* (und damit auch *CoC*) verwendet, besitzt einen Algorithmus, mit dem sich aus dem in *CiC* zertifizierten Programm ein Computerprogramm des funktionalen Programmierens extrahieren lässt. Mittlerweile liegen auch andere Extraktionen in gängige Programmiersprachen wie z.B. *C* und *Java* vor. Damit stehen die zertifizierten Programme in Anwendersprachen bereit, bei deren Übersetzung durch den Extraktionsalgorithmus ebenfalls Fehler ausgeschlossen werden. Der Vorteil des automatischen Beweisens ist es, die Korrektheit einer Software wie ein mathematisches Theorem zu beweisen. Das leisten eben Beweisassistenten. Praktisch wird *Coq* bereits seit ca. 20 Jahren als Kontrollsystem der fahrerlosen Metrolinie 14 in Paris eingesetzt und wurde für andere Anwendungen weiterentwickelt.[24]

Daher lautet der Vorschlag des Autors, eine formale Metaebene für das *Machine Learning* einzuführen, um dort Korrektheitsbeweise mit z.B. dem Beweisassistenten *Coq* automatisch ausführen zu lassen.[25] Dazu kann man sich ein selbstlernendes Automobil vorstellen, das mit Sensoren und damit verbundenem neuronalen Netz ausgestattet ist – quasi als Nervensystem und Gehirn des Systems. Ziel ist es, dass das Verhalten des Automobils nach den Regeln der Straßenverkehrsordnung verläuft. Die Straßenverkehrsordnung wurde 1968 in der Wiener Konvention formuliert.

In einem ersten Schritt wird das Automobil wie ein Flugzeug mit einer Black Box ausgestattet, um die Fülle der Verhaltensdaten zu registrieren (Abb. 1). Aus den registrierten Daten sollte sich ein Fahrverhalten ergeben, dass der Straßenverkehrsordnung entspricht. Logisch muss dazu das Fahrverhalten aus den Verkehrsgesetzen folgen (in Abb. 1 formal $\models$ für eine logische Folgerung). Diese logische Implikation realisiert die gewünschte Kontrolle, um Fehlverhalten auszuschließen. Auf der Me-

23 Vgl. Klaus Mainzer: *The Digital and the Real World. Computational Foundations of Mathematics, Science, Technology, and Philosophy*, Singapore 2018, S. 179–200.

24 Vgl. Mélanie Jacquel u.a.: »Verifying B proof rules using deep embedding and automated theorem proving«, in: *Software Systems Modeling* 14 (2015), Heft 1, S. 101–119.

25 Vgl. Klaus Mainzer: *Wie berechenbar ist unsere Welt. Herausforderungen für Mathematik, Informatik und Philosophie im Zeitalter der Digitalisierung* (Reihe: Essentials), Berlin 2018, S. 23–24.

taebene wird die Implikation formalisiert, um ihren Beweis durch einen Beweisassistenten zu automatisieren.

Dazu müsste zunächst das Rechtssystem der Wiener Konvention wie ein Axiomensystem formalisiert werden (in Abb. 1 durch logische Konjunktion $\bigwedge_{i=1}^{n} \phi_i$ aller Gesetze ϕ_i). In einem nächsten Schritt müsste dann aus der Datenmasse der Black Box die Bewegungsbahn, also der kausale Bewegungsablauf des Fahrzeugs, in einem Kausalmodell extrahiert werden. Dazu bietet sich das kausale Lernen an, das vorher erklärt wurde. Das Kausalmodell (also der kausale Bewegungsablauf) lässt sich graphisch in einer Kausalkette von Ursachen und Wirkungen darstellen. Diese Bahnkurve des Fahrzeugs (Trajektorie) wird dann auf der Metaebene in einer formalen Sprache repräsentiert. Ihre formale Beschreibung müsste aus den entsprechenden Vorschriften der Wiener Konvention logisch ableitbar sein. Im Rechtsalltag werden solche Schlüsse von ausgebildeten Juristen vollzogen. Der formale Beweis dieser Implikation wird durch den Beweisassistenten automatisiert und könnte mit heutiger Rechenpower realisiert werden.

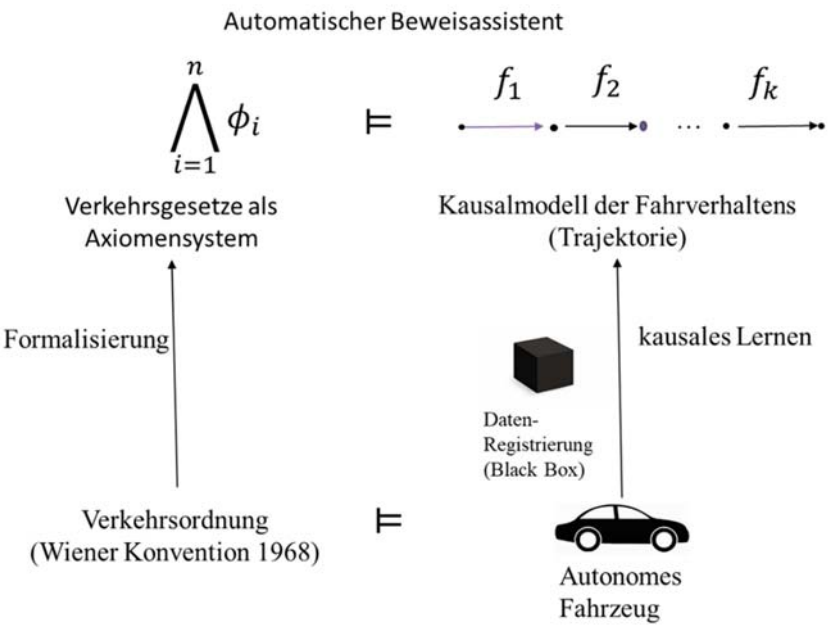


Abb. 1: Sicheres autonomes Fahren durch kausale Lernalgorithmen und automatische Beweisassistenten

Zusammengefasst folgt: *Machine Learning* mit neuronalen Netzen funktioniert, aber wir können die Abläufe in den neuronalen Netzen nicht im Einzelnen verstehen und

kontrollieren. Die ›Unheimlichkeit‹ heutigen *Machine Learnings* zeigt sich in der Undurchsichtigkeit komplexer neuronaler Netze und ihrer Verarbeitung von Big Data. Heutige Techniken des *Machine Learnings* beruhen meistens nur auf statistischem Lernen, aber das reicht nicht für sicherheitskritische Systeme. Daher sollte *Machine Learning* mit Beweisassistenten und kausalem Lernen verbunden werden. Korrektes Verhalten wird dabei durch Metatheoreme in einem logischen Formalismus garantiert.

Dieses Modell selbstlernender Fahrzeuge erinnert an die Organisation des Lernens im menschlichen Organismus: Verhalten und Reaktionen laufen dort ebenfalls weitgehend unbewusst ab. ›Unbewusst‹ heißt, dass wir uns der kausalen Abläufe des durch sensorielle und neuronale Signale gesteuerten Bewegungsapparats nicht bewusst sind. Das lässt sich mit Algorithmen des statistischen Lernens automatisieren. In kritischen Situationen reicht das aber nicht aus: Um mehr Sicherheit durch bessere Kontrolle im menschlichen Organismus zu erreichen, muss der Verstand mit kausaler Analyse und logischem Schließen eingreifen. Das Ziel des Autors ist es, dass dieser Vorgang im *Machine Learning* durch Algorithmen des kausalen Lernens und logischer Beweisassistenten automatisiert wird.

4. Autonomie und ›Künstliche Intelligenz‹

In der KI-Forschung redet man zwar bereits vom ›autonomen‹ Fahren. Tatsächlich handelt es sich heute noch immer häufig nur um vorprogrammierte Fahrerassistenzsysteme. Auch im Fall des *Machine Learnings* sind die Lernalgorithmen vom Programmierer vorgegeben, obwohl sich das System selber in gegebenen Situationen entscheiden und daraus lernen kann.

Ein anderes Beispiel sind ›moralische‹ Kriegsroboter (*moral soldier*), die das US-Militär entwickelt. Hintergrund ist die Erfahrung mit Soldaten, die im Zivilleben soziale Bürger waren, aber im Krieg (z.B. Mỹ Lai-Massaker im Vietnam Krieg 1968) nach traumatischen Kriegserfahrungen verrohten und fürchterliche Kriegsverbrechen begingen. Ähnlich wie beim Autofahren, so die technische Überlegung, lassen sich menschliche Defizite bei emotionalen Belastungen vermeiden, indem verlässliche KI-Systeme zum Einsatz kommen, die sich in jedem Fall an die Regeln halten – sei es im Verkehrsrecht oder Völkerrecht.

Ein technisches System, das sich an vorprogrammierte moralische Regeln hält, ist damit selber aber noch nicht ›moralisch‹. Auch im Fall von Lernalgorithmen haben wir es bestenfalls mit KI-Systemen zu tun, die mit trainierten Tieren oder Kleinkindern verglichen werden können. Autonomie im Sinn politischer Freiheitsrechte meint aber eine höhere Stufe: Autonom ist der in jeder Hinsicht selbstbestimmte Mensch, der zur Selbstgesetzgebung fähig ist.

Nach Kant ist eine Maxime nur dann moralisch gerechtfertigt, wenn sie Grundlage einer allgemeinen Gesetzgebung sein kann.[26] Das ist der Kerngedanke seines Kategorischen Imperativs: Die Regel (Kant: »Maxime«[27]) meiner Handlung muss verallgemeinerungsfähig sein. Mein Recht endet dort, wo das Recht des anderen beginnt. Dringe ich in den Freiraum des anderen ein, so ist diese Maxime nicht verallgemeinerungsfähig: Sie würde unweigerlich den Krieg aller gegen alle nach sich ziehen. Die Maxime meiner Handlung müsste also im Prinzip die Grundlage eines allgemeinen Gesetzes sein können, das z.B. der Deutsche Bundestag beschließt. In diesem Sinn bedeutet Autonomie die Fähigkeit zur ›Selbstgesetzgebung‹.

Technisch lässt sich nicht ausschließen, dass eine ›Künstliche Intelligenz‹ eines Tages auch zur ›Selbstgesetzgebung‹ fähig ist, d.h. sie gibt sich ihre Gesetze als Programme selber: Sie programmiert sich selbst![28] In der Gewaltenteilung westlicher Demokratien ist die Gesetzgebung das Recht des Parlaments (d.h. legislative Gewalt), das von der Bevölkerung eines Landes in freien Wahlen gewählt wurde. Die Rede ist bereits von einer Superintelligenz, die den besseren Überblick über globale Situationen hat als die kollektive Intelligenz aller Menschen zusammen. In diesem Fall hätten wir allerdings unsere Autonomie aufgegeben. KI wäre nicht länger nur Dienstleistungssystem; mit den Worten von Hegel wäre KI nicht länger ›Knecht‹, sondern ›Herr‹. Hier beginnt die Unheimlichkeit der KI.

Spätestens an dieser Stelle stellt sich die Frage nach der Verantwortung im Zeitalter der Digitalisierung und ›Künstlichen Intelligenz‹ grundsätzlich. Der Verantwortungsbegriff hat eine lange rechtliche und philosophische Tradition.[29] Unter Verantwortung versteht man im Allgemeinen die Pflicht einer handelnden Person (bzw. Personengruppe) gegenüber einer anderen Person (bzw. Personengruppe) aufgrund eines Anspruchs, der durch eine Instanz (z.B. Institution, Staat, Gesellschaft) erhoben wird.

Erste Unterscheidungskriterien sind z.B. kausale Verantwortung mit Blick auf die Verursachung (z.B. Programmierfehler eines Programmierers), Rollenverantwortung mit Blick auf eine Aufgabe (z.B. Lehrer für seine Schulklasse), Fähigkeitenverantwortung mit Blick auf die Erfüllbarkeit (z.B. ein Mediziner bei einem Unfall) und Haftungsverantwortung, die von der Verursachung abweichen kann (z.B. ›Eltern haften für Ihre Kinder‹).[30] Die Feststellung der kausalen Verantwortung ist nicht

26 Vgl. Immanuel Kant: *Kant's gesammelte Schriften* (Ausgabe der Preußischen Akademie der Wissenschaften), Berlin 1900, AA IV, 421: »Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie ein allgemeines Gesetz werde.«.

27 Ebd.

28 Vgl. Klaus Mainzer: *Künstliche Intelligenz. Wann übernehmen die Maschinen*?, Berlin 2019, S. 267–279.

29 Vgl. Julian Nida-Rümelin: *Verantwortung*, Stuttgart 2011.

30 Vgl. Herbert L. Hart: *Punishment and Responsibility. Essays in the Philosophy of Law*, Oxford 1968.

normativ, sondern beruht auf empirischen Erkenntnissen. Sie ist das zentrale Problem bei den ›Black Boxes‹ neuronaler Netze, von denen bereits die Rede war.

In Haftungsfragen werden juristische Personen (z.B. Unternehmen) als verantwortliche Handlungssubjekte behandelt. Strafrechtliche Verantwortung für Institutionen gibt es jedoch nach deutschem Recht (im Unterschied z.B. zum US-amerikanischen Recht) nicht. Wenigstens moralisch wird aber Unternehmen auch Verantwortung zugeschrieben. Man spricht dann von *Corporate Governance* und *Corporate Social Responsibility*.

Juristisch wird Verantwortung als die Pflicht einer Person verstanden, für ihre Entscheidungen und Handlungen nach festgelegten Vorschriften Rechenschaft (*accountability*) ablegen zu müssen. Dabei bezieht sich also Recht nicht auf moralische oder religiöse Verantwortung (z.B. Gewissen), sondern (›positivistisch‹) auf die Verletzung von Rechtsvorschriften, die durch ein Gericht festgestellt wird. Damit ist Verantwortung im juristischen Sinn immer an empirische Fakten gebunden. Deshalb ist die Forderung nach mehr Erklärbarkeit von kausalen Abläufen im *Machine Learning* von grundlegender Bedeutung für die Klärung rechtlicher Verantwortung.

Juristisch unterscheidet man bei der Rechenschaftspflicht (*accountability*) z.B. folgende Aspekte:[31]

a) Mit Handlungsverantwortung wird die Rechenschaftspflicht hinsichtlich der Art der Aufgabendurchführung bezeichnet.
b) Mit Ergebnisverantwortung wird die Rechenschaftspflicht hinsichtlich der Zielerreichung bezeichnet.
c) Mit Führungsverantwortung wird die Rechenschaftspflicht hinsichtlich der wahrgenommenen Führungsaufgaben, auch bei der zugehörigen Fremdverantwortung, bezeichnet.

Im Recht bezieht sich Verantwortung nicht nur auf Personen, sondern auch auf Sachgüter (z.B. Computer) und auf Anforderungen eines Eigentümers, Treuhänders oder Mieters. Mit zunehmendem Grad der Autonomie intelligenter Systeme stellt sich die Frage, bis zu welchem Grad z.B. Roboter noch als Sachgüter behandelt werden können oder ob wir rechtlich bereits Zwischenbereiche zwischen Sachgütern und Personen berücksichtigen müssen. Das Tierrecht zeigt, wie wenig angebracht die traditionelle Unterscheidung von Sache und Person ist, wenn wir moderne Erkenntnisse der Evolutionsbiologie und Kognitionspsychologie zugrunde legen: Tiere

31 Vgl. Hans M. Baumgartner und Albin Eser (Hg.): *Schuld und Verantwortung. Philosophische und juristische Beiträge zur Zurechenbarkeit menschlichen Handelns*, Tübingen 1983, S. 136.

sind leidensfähige Lebewesen und keine »Sachen«, andererseits aber noch keine »Personen«.[32]

›Künstliche Intelligenz‹ unterliegt ohne Zweifel dem Prinzip der Verantwortung: Erst der Mensch sollte bestimmen, wie sie eingesetzt wird. Spezialisierung und die damit wachsende Komplexität technischer, sozialer und ökologischer Zusammenhänge führen jedoch zu einer Diffusion von Verantwortung: Der Einzelne ist verstärkt auf die Informationen bzw. Einschätzungen anderer Experten angewiesen. Als Folge ergibt sich die Notwendigkeit der institutionellen Zuschreibung von Verantwortung durch gesetzliche oder vertragliche Bestimmungen, z.B. im Haftungsrecht, und/oder die Zuschreibung der Verantwortung auf kollektive Akteure wie z.B. Unternehmen und Verbände. Allerdings begünstigt die Diffusion von Verantwortung auch klare Rechtsverstöße und Missbrauch von Technik, die in der Öffentlichkeit zu Empörung und Verunsicherung führen.[33]

Mit Blick auf komplexe KI-Systeme und KI-Infrastrukturen ist der Verantwortungsbegriff zu erweitern. Systemtheoretisch muss neben individueller Verantwortung auch kollektive und kooperative Verantwortung analysiert werden. Dabei sollte Verantwortung auch denjenigen zugeschrieben werden können, die für das Design von KI-Systemen (z.B. ›Industrie 4.0‹), die Entstehung von Schnittstellen und den Einsatz der Infrastruktur zuständig sind. Hierbei sind die Grade der Einwirkungsmöglichkeit zu bemessen. Am Ende ist eine verantwortungsvolle KI das Ziel, mit der wir der Unheimlichkeit dieser Schlüsseltechnologie am besten begegnen.

32 Herbert Zech: *Information als Schutzgegenstand*, Tübingen 2012; Klaus Mainzer: *Information. Algorithmus-Wahrscheinlichkeit-Komplexität-Quantenwelt-Leben-Gehirn-Gesellschaft*, Berlin, Köln 2016, S. 172.

33 Die Deutsche Akademie für Technikwissenschaften (acatech) richtete 2018 das Arbeitsprojekt ›Verantwortung‹ ein, um aktuelle Fälle der Unternehmensverantwortung aus technischer, ökonomischer, rechtlicher und ethischer Sicht aufzuarbeiten und zu bewerten.
.