Kapitel 2 – Künstliche Intelligenz als technische Innovation

Der nun folgende Teil widmet sich dem Untersuchungsgegenstand dieser Arbeit. Es wird herausgearbeitet werden, was die Technik der Künstlichen Intelligenz ausmacht und welche prägenden Eigenschaften damit verbunden sind (siehe unten A.). In einem zweiten Schritt wird – aufbauend auf dem im Grundlagenteil erarbeiteten Verständnis von Technik und Risiko – herausgearbeitet, ob und weshalb KI eine technische Innovation ist, weil sie nicht dem klassischen Verständnis von Technik entspricht und das technische Risiko verändert (siehe unten B.).

A. Künstliche Intelligenz

Im folgenden Abschnitt werden zunächst die verschiedenen *Definitionsansätze* zur KI sowie die damit verbundenen wesentlichen Eigenschaften der Lernfähigkeit und technischen Autonomie vorgestellt (siehe unten I.). In einem zweiten Schritt wird die als wesentlich angesehene Fähigkeit von KI zu maschinellem Lernen erläutert und dargestellt, wie KI „geschaffen" wird (siehe unten II.). Unter dem Punkt III. wird der für die rechtliche Bewertung zentrale Black-Box-Effekt maschinellen Lernens analysiert (III.). Daran schließt die Frage „Can machines think?" und eine Gegenüberstellung der Konzepte schwache vs. starke KI an (IV.).

I. Begriff und Eigenschaften Künstlicher Intelligenz

Ein erster Ansatz, sich dem Begriff der KI zu nähern, ist, diese als Gegensatz zur *menschlichen* Intelligenz zu verstehen. Es geht bei KI also nicht primär darum, menschliche Intelligenz zu erforschen und zu verstehen.[747] KI umschreibt vielmehr den Versuch, Einheiten zu erschaffen, die als intelligent bezeichnet werden können.[748] Sofern darüber hinaus im heutigen Kontext von KI die Rede ist, wird damit v.a. ein Teilbereich der Informatik bezeichnet und damit die Suche nach einer Künstlichen Intelligenz in Form einer Computerintelligenz.[749] Nicht alles, was ein Computer leistet, bedeutet jedoch zugleich auch Computerintelligenz. Ein Computer mag in der Lage sein, komplexe Berechnungen schneller und zu-

747 Darum geht es jedoch in der mit der KI verwandten, aber mit anderen Forschungszielen verbundenen *Kognitionswissenschaft*. Diese hat die Erforschung der kognitiven Fähigkeiten des Menschen zum Ziel (Motorik, Wahrnehmung, Lernen, Sprache etc.), wobei die Computersimulation von kognitiven und neuronalen Prozessen einen Schwerpunkt bildet, und Computermodelle der KI eingesetzt werden, *Russell/Norvig* 2021, 20 f.; *Lenzen* 2018, 31. Vgl. auch *Boden* 2014, 89.

748 *Ertel* 2021, 1; *Frankish/Ramsey* 2014; *Russell/Norvig* 2010, 1.

749 *Frankish/Ramsey* 2014, 1; *Ertel* 2021, 1; *Boden* 2014, 89; *Lenzen* 2018, 23.

verlässiger als ein Mensch anzustellen, dennoch würde man ihm allein auf Grund dieser Fähigkeit nicht das Attribut einer Künstlichen Intelligenz zuweisen.[750]

Im Folgenden werden zwei Ansätze vorgestellt, um KI begrifflich zu umschreiben – ein verhaltensbezogener Ansatz sowie – als anwendungsorientierter Ansatz – das Modell des intelligenten Agenten. In einem dritten Schritt werden zwei prägende Eigenschaften von KI vorgestellt – die Lernfähigkeit sowie die (technische) Autonomie.

1. KI als Imitation menschlichen Verhaltens

In den Anfängen der Forschung zur KI dominierte v.a. ein verhaltensbezogener Ansatz; man beschrieb diese als Imitation menschlichen Verhaltens und untersuchte, welche Fähigkeiten ein Computersystem haben muss, um diese Imitationsleistung zu erbringen.

a. Der Ansatz des Turing-Tests

Den berühmtesten, weil wohl eingängigsten Ansatz zur Beschreibung von KI vertrat *Alan Turing* in seinem 1950 veröffentlichten und damit zu diesem Zeitpunkt visionären Beitrag „Computer machinery and intelligence".[751]

Ausgangspunkt seines Beitrags war ein inzwischen als Turing-Test bekanntes Experiment sowie die Frage, ob sich eine Maschine in einem „imitation game" als Mensch bewähren könne.[752] Teil des Experiments sind neben einer Testperson („interrogator"), eine Person A sowie ein Computer.[753] Die Testperson wird in einem separaten Raum platziert und kann an Person A sowie an den Computer Fragen übermitteln. Aufgabe der Testperson ist es, anhand ihrer Fragen und den erhaltenen Antworten zu bestimmen, ob sie mit Person A oder mit dem Computer kommuniziert. Der Turing-Test ist bestanden – der Computer hat sich im Imitationsspiel bewährt –, wenn die Testperson nach einer fünfminütigen Befragung nur mit einer 70-prozentigen Wahrscheinlichkeit in der Lage ist, Maschine und Computer richtig zu identifizieren, also im Gegenteil in mindestens 30 Prozent der Fragerunden davon ausgeht, menschliche und gerade nicht computergenerierte Antworten erhalten zu haben.[754]

Um sich in diesem Imitationsspiel zu bewähren, muss ein Computer eine Reihe von Fähigkeiten aufweisen, nämlich:[755] (1) die Verarbeitung natürlicher

750 *Lenzen* 2018, 25. Vgl. auch *Zech* 2020, A 20.
751 *Turing* Mind 59 (1950), 433 (433 f.) Ausführlich hierzu *Warwick/Shah* 2016.
752 *Turing* Mind 59 (1950), 433 (435).
753 *Turing* spricht zwar von „machine", versteht darunter aber einen „digital computer", also das, was wir heutzutage unter einem Computer verstehen, vgl. *Turing* Mind 59 (1950), 433 (436).
754 *Turing* Mind 59 (1950), 433 (442).
755 Zusammengefasst bei *Russell/Norvig* 2021, 20.

Sprache, um mit der Testperson zu kommunizieren;[756] (2) das Speichern von vorhandenem oder von der Testperson mitgeteiltem Wissen; (3) automatisierte Argumentation, um gespeichertes Wissen für Antworten verwenden und neue Schlussfolgerungen ziehen zu können; (4) maschinelles Lernen, um sich neuen Umständen anpassen sowie Muster erkennen und erschließen zu können.[757]

Diese genannten Fähigkeiten beschreiben in der Tat Teilbereiche der anwendungsorientierten KI-Forschung, also der Forschung, die sich mit der tatsächlichen Entwicklung künstlich intelligenter Systeme beschäftigt.[758] Insofern besteht der Vorteil des Turing-Tests darin, aufzuzeigen, welche einzelnen Fähigkeiten ein Computer haben muss, um als intelligent bezeichnet werden zu können, und entsprechend, welche anwendungsorientierten Forschungsbereiche notwendig sind.

Der ursprüngliche Turing-Test beschränkte sich ausschließlich auf die intellektuellen Fähigkeiten („intellectual capacities of a man") des Menschen, bezog sich also nicht auch auf dessen physische Fähigkeiten („physical capacities"). Der Testperson sollte es entsprechend nicht möglich sein, Mensch und Maschine zu fühlen und zu sehen bzw. ihre Stimmen zu hören.[759] Eine Weiterentwicklung des Turing-Tests des Kognitionswissenschaftlers *Harnad* zum „Total Turing Test" hebt die Beschränkung auf intellektuelle Fähigkeiten auf: Die Testperson kann auch die visuelle Wahrnehmungsfähigkeit sowie die Fähigkeit, Gegenstände zu bewegen und zu bearbeiten (Fähigkeit zur Robotik), testen.[760] Auch die Entwicklung dieser beiden Fähigkeiten ist Gegenstand wichtiger Teilbereiche der anwendungsorientierten KI-Forschung.[761]

Als *Turing* sein Imitationsspiel 1950 vorstellte, war ein Computer, der diesen Test bestehen konnte, ein reines Gedankenexperiment, denn die Entwicklung moderner Computer steckte noch in den Kinderschuhen. Die Herausforderungen sah *Turing* jedoch weniger in der gerätetechnischen Entwicklung – v.a. dem Ausbau von Speicherkapazitäten und der damit einhergehenden Beschleunigung von Rechenprozessen –, sondern in der richtigen Programmierung. *Turing* war dennoch überzeugt, dass um die Jahrtausendwende ein Computer in der Lage sein würde, den Turing-Test zu bestehen.[762]

Turing sollte mit seiner Prognose zur gerätetechnischen Computerentwicklung Recht haben, falsch lag er hingegen mit seiner Einschätzung zur Entwicklungsgeschwindigkeit von KI: Der 1991 ins Leben gerufene Loebner-Preis, der u.a.

756 *Turing* Mind 59 (1950), 433 (460).
757 *Turing* Mind 59 (1950), 433 (454 ff., 460).
758 Siehe zB die Werke von *Russell/Norvig* 2021; *Ertel* 2021.
759 *Turing* Mind 59 (1950), 433 (434).
760 *Harnad* MaM 1 (1991), 43 (44); vgl. dazu *Russell/Norvig* 2021, 20.
761 Siehe z.B. *Russell/Norvig* 2021.
762 *Turing* Mind 59 (1950), 433 (442).

einen Preis für das Bestehen des ursprünglichen bzw. des erweiterten Turing-Tests ausschreibt, wurde bis heute nicht vergeben.[763]

b. John McCarthy und artificial intelligence

Den Begriff der „artificial intelligence", der Künstlichen Intelligenz, formulierte 1955 erstmals der Mathematiker *McCarthy* in seinem Vorschlag für ein 2-monatiges Forschungsprojekt am US-amerikanischen Dartmouth College:

> „We propose that a 2 month, 10 man study of *artificial intelligence* be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that *every aspect of learning or any other feature of intelligence* can in principle be so precisely described that a *machine can be made to simulate it*. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems *now reserved for humans*, and improve themselves."[764]
>
> „Wir schlagen vor, dass im Sommer 1956 am Dartmouth College in Hanover, New Hampshire, eine zweimonatige, 10-köpfige Studie zur *Künstlichen Intelligenz* durchgeführt wird. Die Studie soll auf der Grundlage der Annahme durchgeführt werden, dass *jeder Aspekt des Lernens oder jedes andere Merkmal von Intelligenz* im Prinzip so genau beschrieben werden kann, dass eine *Maschine diese simulieren kann*. Es wird versucht, herauszufinden, wie man Maschinen dazu bringt, Sprache zu benutzen, Abstraktionen und Konzepte zu bilden, Probleme zu lösen, die *jetzt dem Menschen vorbehalten* sind, und sich selbst zu verbessern."
>
> [Hervorhebungen in kursiv durch die Verf.]

Auch *McCarthy* verfolgte einen *verhaltensbezogenen* Ansatz, indem er KI als etwas beschrieb, das in der Lage ist, menschliche Intelligenz zu simulieren und das Ziel hat, Aufgaben zu übernehmen, die bislang Menschen vorbehalten sind. Maschinelles Lernen sah er dabei als Kernfähigkeit von KI.

c. KI als dynamischer Begriff

Die Beschreibung von KI als die Fähigkeit zur Simulation menschlicher Intelligenz ist verschiedener Kritik ausgesetzt. Hierzu folgendes Beispiel[765]: Sicherlich ein Aspekt menschlicher Intelligenz ist die Fähigkeit zum Lösen mathematischer

763 *Lenzen* 2018, 27; vgl. auch *Russell/Norvig* 2021, 1035. Kontrovers diskutiert wurde, ob der Chatbot „Eugene Goostman" das erste System sei, dass das „imitation game" bestanden habe. Dieser Chatbot konnte grammatikalische Fehler und fehlendes Wissen überspielen, indem er sich als 13-jähriger ukrainischer Junge ausgab. In einem Turing-Test der Royal Society London erzielte er eine Täuschungsquote von 33 %. Kritiker merkten jedoch an, dass dieser Chatbot durch Täuschen und Ablenken die Juroren in die Irre geführt habe und insofern gar nicht intelligent sei. Dazu *Marcus* 2014; *Stephan/Walter* in: *Turing* 2021, 185.

764 *McCarthy/Minsky/Rochester u. a.* AI Magazine 27 (2006), 12.

765 Zu weiteren Beispielen, die die Schwächen dieses Definitionsansatzes verdeutlichen: *Lenzen* 2018, 28 f.; *Ertel* 2021, 2 f.

Rechenaufgaben; dennoch würde man heutzutage wohl kaum einen Taschenrechner als intelligent beschreiben, auch wenn er nicht nur die Grundrechenaufgaben beherrscht, sondern auch hochkomplexe Rechenaufgaben in viel höherer Geschwindigkeit als ein Mensch löst. Vielmehr arbeitet der Taschenrechner nach unserem Verständnis ein vorgegebenes Programm ab. Dasselbe mag gelten für Schachcomputer: 1997 besiegte IBM´s Schachcomputer „Deep Blue" den damaligen Schachweltmeister Gary Kasparov; damit war die Überlegenheit von Schachcomputern gegenüber menschlichen Spielern belegt.[766] Dennoch würde man auch hier diesen Sieg mehr der enormen Rechenkapazität und der richtigen Programmierung des Schachcomputers seitens der Entwickler zurechnen, als den Schachcomputer als intelligent zu bezeichnen.

Dieses Paradox, etwas nicht mehr als künstlich intelligent zu bezeichnen, sobald es eine Maschine kann,[767] – die Rede ist in diesem Zusammenhang auch vom sog. „AI effect" – umgeht folgender ebenfalls verhaltensbezogener Definitionsansatz der Künstlichen Intelligenz:

> „Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better."[768]
>
> „Künstliche Intelligenz ist die Erforschung dessen, wie man Computer dazu bringt, Dinge zu tun, bei denen die Menschen im Moment besser sind."

Diese Definition berücksichtigt, dass der Begriff der KI ein dynamischer ist und sein wird, weil er in hohem Maße als Ausdruck technischer Innovation empfunden wird.[769] Der Begriff ist an die technische Weiterentwicklung, das menschliche Verständnis davon und die Gewöhnung im Umgang mit dieser gekoppelt: So mag ein Taschenrechner bei seiner Einführung vor vielen Jahrzehnten ob seiner Rechenleistung durchaus als künstlich intelligent bezeichnet worden sein. Mit zunehmendem Verständnis von der Funktionsweise dieser Geräte[770] und der zunehmenden Gewöhnung im Umgang mit diesen, haben sie diese Bezeichnung aber verloren. Und so wie man derzeit das Sprachmodell GPT4, das „auf Knopfdruck" längere Texte zu beliebigen Themen nahezu fehlerfrei generieren kann, geradezu als Paradebeispiel von KI bezeichnen mag, mag es dieses Attribut mit zunehmendem Verständnis von seiner Funktionsweise und seiner Verbreitung im Alltag wieder einmal verlieren.

766 Vgl. dazu *Franklin* 2014, 23.
767 Vgl. *Lenzen* 2018, 30.
768 *Rich/Knight/Nair* 2010, 3.
769 Vgl. auch *Ertel* 2021, 2 f.
770 Siehe auch *Lenzen* 2018, 30.

2. *Das Modell des intelligenten Agenten*

Einen weiteren Kritikpunkt an einem verhaltensbezogenen Definitionsansatz – insbes. im Hinblick auf *Turings* Imitationsspiel – erhellt folgender Auszug[771] aus dessen Beitrag „Computing Machinery and Intelligence", in welchem er eine fiktive Konversation der Testperson mit dem Computer beschreibt:

> „Q: Add 34957 to 70764
> A: (Pause about 30 seconds and then give as answer) 105621"
>
> „Frage: Addiere 34957 und 70764
> Antwort: (mach eine Pause von 30 Sekunden und dann gib als Antwort) 105621"

Was würde eine Testperson dazu veranlassen, die erhaltene Antwort einem Menschen zuzuordnen und nicht einem Computer?

Einerseits die Tatsache, dass der Computer nicht sofort antwortet, sondern erst nach einer Pause von 30 Sekunden; andererseits, dass als Antwort 105621 und nicht – korrekterweise – 105721 gegeben wird.[772] Die Testperson würde einen Menschen hinter den Antworten vermuten, weil ein Mensch bekanntlich langsamer rechnet als ein Computer und dabei Fehler macht, weil also seine Rechenfähigkeiten hinter denen des Computers zurückbleiben. Entsprechend merken Kritiker an, dass *Turings* Ansatz mehr dazu motiviere, ein System zu bauen, das die Testperson möglichst effektiv in die Irre führt, indem es menschliche Fehlbarkeit und nicht menschliche Intelligenz imitiert.[773]

Zudem kann zwar ein verhaltensbezogener Ansatz die Fähigkeiten umschreiben, die erforderlich sind, einen Computer als intelligent zu bezeichnen; damit ist aber noch nichts darüber ausgesagt, wie eine Maschine operieren soll, um intelligentes Verhalten an den Tag zu legen.

An dieser Stelle setzt – auf Grundlage einer anwendungsorientierten Herangehensweise[774] – das Modell des rationalen oder intelligenten Agenten an:

> „Al is […] the subfield of computer science which aims to construct agents that exhibit aspects of intelligent behaviour."[775]
>
> „KI ist das Teilgebiet der Informatik, welches zum Ziel hat, Agenten zu entwickeln, die Elemente intelligenten Verhaltens aufweisen."

771 *Turing* Mind 59 (1950), 433 (434).
772 Siehe zu diesem von der Fachwelt vermuteten seitens *Turing* gelegten „easter egg" *Fokker* 2012.
773 *Lenzen* 2018, 31. *Turing* hat diesen Einwand selbst vorhergesehen; es hat ihn jedoch nicht am Konzept seines „imitation game" zweifeln lassen, *Turing* Mind 59 (1950), 433 (448 f.).
774 Vgl. zu diesem Begriff *Lenzen* 2018, 31.
775 *Wooldridge/Jennings* The Knowledge Engineering Review 10 (1995), 115 (116). Vgl. bspw. auch die Definition von *Poole/Mackworth/Goebel* 1998, XV: „More commonly referred to as artificial intelligence, computational intelligence is the study of the design of intelligent agents." („Häufig als Künstliche Intelligenz bezeichnet, beschreibt Computerintelligenz die Erforschung der Entwicklung intelligenter Agenten.").

Das Konzept des intelligenten Agenten hielt in den 1990ern Einzug in die KI-Forschung[776] und hat sich inzwischen als Standard etabliert, insbes. seit er in dem Standardwerk zur KI von *Russel* und *Norvig* als Leitbild zu Grunde gelegt wurde.[777] Es handelt sich dabei ausweislich um ein Mittel zur Analyse und Entwicklung von KI-Systemen, nicht um eine eindeutige Definition von KI:

> „The notion of an agent is meant to be a tool for analyzing systems, not an absolute characterization that divides the world into agents and non-agents."[778]
>
> „Das Konzept eines Agenten soll ein Werkzeug zur Analyse von Systemen sein, nicht eine absolute Charakterisierung, die die Welt in Agenten und Nicht-Agenten unterteilt."

Insofern steht der Ansatz des intelligenten Agenten[779] nicht in einem direkten Widerspruch zu den oben beschriebenen verhaltensbezogenen Ansätzen. Ein anwendungsorientierter Ansatz konzentriert sich nur mehr darauf, die erforderlichen Operationsschritte zu definieren, damit Systeme tatsächlich intelligent operieren.

Zunächst zum Begriff des *Agenten*:

> „An Agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators."[780]
>
> „Ein Agent ist alles, was man als etwas ansehen kann, dass seine Umgebung über Sensoren wahrnimmt und über Aktuatoren auf diese Umgebung einwirkt."

Ein Agent ist nach dieser Definition ganz allgemein ein System, welches Informationen verarbeitet und aus diesem Input einen Output produziert.[781] Der Begriff des Agenten ist also zunächst sehr unspezifisch – auch ein Mensch kann schließlich ein Agent sein, sog. *human agent*,[782] – und, insofern sich die KI-Forschung auf maschinelle bzw. computerbasierte Agenten beschränkt, unterscheidet sich der Begriff auch noch nicht von einem gewöhnlichen Computersystem.[783] Der Begriff des Agenten ist aber jedenfalls anschlussfähig an einen formalen Technikbegriff, der ebenfalls von einer Input-/Output-Beziehung ausgeht.[784]

776 Vgl. dazu *Franklin* 2014, 28; *Russell/Norvig* 2010, 26.
777 *Beierle/Kern-Isberner* 2019, 338; *Franklin* 2014, 28. Vgl. auch *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 1; *Alonso* 2014, 244.
778 *Russell/Norvig* 1995, 33.
779 Häufig ist auch vom rationalen oder autonomen Agenten die Rede. Hier wurde jedoch diejenige Begrifflichkeit gewählt, die dem Standardwerk von *Russell* und *Norvig* zur Künstlichen Intelligenz zu Grunde gelegt ist; ebenso *Ertel* 2021, 19 f.
780 *Russell/Norvig* 2021, 54.
781 *Ertel* 2021, 19; *Beierle/Kern-Isberner* 2019, 398. Der Begriff des Agenten wird teilweise auch deutlich enger verstanden, indem zusätzliche Eigenschaften wie Lernfähigkeit, Autonomie, Reaktivität etc. gefordert werden. Vgl. etwa *Wooldridge/Jennings* The Knowledge Engineering Review 10 (1995), 115 (116 ff.); *Franklin/Graesser* 1997, 22 ff.
782 Vgl. *Beierle/Kern-Isberner* 2019, 398; *Russell/Norvig* 2021, 54. *Russel* und *Norvig* weisen insofern darauf hin, dass die agentenbasierte Perspektive die KI-Forschung anderen Wissenschaftsbereichen angenähert habe, die sich mit dem Konzept des Agenten beschäftigen, etwa der Kontrolltheorie und der Ökonomie, *Russell/Norvig* 2010, 27.
783 *Franklin/Graesser* 1997; *Russell/Norvig* 2021, 21.
784 Zum formalen Technikbegriff siehe oben Kap. 1, B.II.1.

Maßstab dafür, ob ein Agent als *intelligent* zu bezeichnen ist, soll nicht sein, wie gut er menschliches Verhalten zu imitieren vermag, sondern ob er „rational" im folgenden Sinne operiert (also rationales Verhalten zeigt):

> „A rational agent is one that acts so as to achieve the *best outcome* or, when there is uncertainty, *the best expected outcome*."[785]
>
> „Ein rationaler Agent handelt so, dass er das *beste Ergebnis* erzielt oder, sofern Unsicherheit herrscht, das *beste zu erwartende Ergebnis*."
>
> [Hervorhebungen in kursiv durch die Verf.]

Aus der „Verpflichtung" eines Agenten auf das Ziel einer so verstandenen Rationalität ergibt sich eine erste prägende Eigenschaft intelligenter Agenten: Sie operieren *zielgerichtet*.[786] Entwickler geben den Agenten ein bestimmtes Ziel vor, welches sie unter Anwendung bestimmter Verfahren umsetzen sollen.[787] Und sie handeln dann rational, wenn die Agenten dieses Ziel unter Berücksichtigung des vorhandenen – programmierten – und durch die Agenten selbst gewonnenen Wissens sowie den Handlungen, zu denen der Agent in der Lage ist, erreichen:

> „For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has." [788]
>
> „Für jede mögliche Wahrnehmungssequenz sollte ein rationaler Agent diejenige Handlung auswählen, von der zu erwarten ist, dass sie die Leistungsindikatoren maximiert, wobei die Erkenntnisse des Agenten aus vorangegangener Wahrnehmungssequenz und das vorhandene Wissen des Agenten zu berücksichtigen sind."

a. Agentenarchitektur: sense-plan-act

Der für das Erzielen von Rationalität im oben beschriebenen Sinne erforderliche Prozess wird mit *sense-plan-act* beschrieben.[789] Der Agent erzielt „Rationalität" durch Wahrnehmung, Schlussfolgerung/Entscheidungsfindung und schließlich Ausführung.

Zur Erläuterung:

(1) Wahrnehmung: Der Agent erfasst seine Umgebung über Sensoren, im Falle sog. Hardware-Agenten oder Robotern sind dies etwa Kameras, Infrarotsenso-

785 *Russell/Norvig* 2021, 22; vgl. auch *Lenzen* 2018, 44.
786 *Beierle/Kern-Isberner* 2019, 398; *Franklin/Graesser* 1997, 29; *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 6. Teilweise wird diese Eigenschaft auch als *proaktiv* beschrieben: *Beierle/Kern-Isberner* 2019, 400; *Wooldridge/Jennings* The Knowledge Engineering Review 10 (1995), 115 (116).
787 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 6.
788 *Russell/Norvig* 2021, 58.
789 Vgl. zum Überblick *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 2. Vgl. ähnl. *Kirn/Müller-Hengstenberg* KI 29 (2015), 59 (61).

ren[790] oder Sensoren zur Messung physikalischer Größen (bspw. Temperatur, Entfernung, Druck etc.)[791], im Falle von Software-Agenten sind dies bspw. Tastatureingaben oder Datei- und Websiteinhalte.[792]

(2) Schlussfolgerung/Entscheidungsfindung: Den Kern des Agenten bildet ein Modul für Schlussfolgerung und Informationsverarbeitung, das auf Grundlage der von den Sensoren erfassten Daten (Input) ein zur Erreichung des vorgegebenen Ziels geeignetes Handeln (Output) vorschlägt.[793] Ein „Spamfilter-Agent" bspw. verarbeitet die Inhalte einer E-Mail (Eingabe), klassifiziert diese als Spam oder normalen Inhalt und schlägt entsprechend als Ausgabe vor, Spam-E-Mails in einen speziellen Ordner zu verschieben.[794]

(3) Ausführung: Nachdem der Agent eine bestimmte Handlung vorgeschlagen hat, muss diese über die zur Verfügung stehenden Aktuatoren ausgeführt werden.[795] Der Spamfilter-Agent aus dem obigen Beispiel verfügt über einen Aktuator, der im Falle des Vorschlags „E-Mail verschieben" diese tatsächlich in einen separaten Ordner verschiebt.[796] Oder ein autonom fahrendes Fahrzeug bremst, nachdem der Agent ein Objekt auf der Straße als Fußgänger klassifiziert und einen Bremsvorgang vorgeschlagen hat.[797]

b. Roboter und Software-Bots

Intelligente Agenten können auf vielfältige Art und Weise klassifiziert werden.[798] Im Folgenden soll nur die – augenscheinlichste – Unterscheidung zwischen Robotern und Software-Bots herausgegriffen werden.

Intelligente Agenten können ausschließlich oder teilweise in einer rein virtuellen Umgebung agieren,[799] d.h. sie nehmen ihren Input aus einer virtuellen Umgebung auf (bspw. auf Grundlage von Tastatureingaben oder dem Inhalt von Dateien) und geben ihren Output an diese Umgebung wieder zurück, bspw. durch die Darstellung am Computer oder in Dateien. [800] Rein softwaregestützte intelligente Agenten werden Software-Agenten oder auch Software-Bots genannt.[801] Beispiele

790 *Russell/Norvig* 2021, 54; *Ertel* 2021, 20.
791 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 2.
792 *Russell/Norvig* 2021, 54; *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 2.
793 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 2.
794 *Ertel* 2021, 21.
795 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 3.
796 *Ertel* 2021, 21.
797 Vgl. *Russell/Norvig* 2021, 67 f.
798 Siehe dazu *Franklin/Graesser* 1997, 29 ff.; *Ertel* 2021, 21 f.; *Russell/Norvig* 2021, 61 ff., 65 ff.
799 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 1.
800 *Russell/Norvig* 2021, 54.
801 *Ertel* 2021, 20; *Russell/Norvig* 2021, 54.

dafür sind sog. Chat-Bots, die eine Kommunikation zwischen einem Nutzer mit einem Computersystem ermöglichen, Bildanalysesoftware oder Suchmaschinen. Daneben können Software-Bots aber auch teilweise Informationen aus der analogen Umwelt aufnehmen und dann einen rein digitalen Output geben, so etwa Sprach- und Gesichtserkennungssysteme, die ihre Informationen mit Hilfe von Mikrofonen und Kameras erlangen.[802]

Intelligente Agenten können aber auch ausschließlich innerhalb von Hardware-Systemen agieren. Sie nehmen über ihre Sensoren Input aus einer analogen Umgebung auf – etwa über Kameras und Infrarotsensoren – und geben ihren Output über ihre Aktuatoren – meist durch einen eingebauten Roboter – in diese Umwelt zurück.[803] Die Rede ist von Hardware-Agenten bzw. – gebräuchlicher – von autonomen Robotern[804] oder cyber-physischen Systemen.[805] Beispiele solcher Hardware-Agenten sind moderne Industrieroboter, autonome Fahrzeuge, OP-Roboter oder Drohnen.[806]

3. Zwischenstand

Ausgehend von einem verhaltensbezogenen Ansatz lassen sich die Fähigkeiten beschreiben, die ein Computersystem aufweisen soll, um als KI bezeichnet zu werden. Dazu gehören Sprach- und Bildverarbeitung, Lernfähigkeit oder die Fähigkeit zur Robotik. Diese Fähigkeiten beschreiben auch Teilbereiche der anwendungsorientierten KI-Forschung. Das Modell des intelligenten Agenten orientiert sich dagegen ein einem „rationalen" Verhalten, in dem Sinne, dass der Agent das ihm vorgegebene Ziel bestmöglich erreichen soll. Zudem zeigt der Ansatz des intelligenten Agenten mit der Beschreibung *sense-plan-act* auf, welche Prozesse ein Computersystem durchlaufen müssen, um diese Form von Rationalität zu erreichen.

4. Lernfähigkeit und (technische) Autonomie

Um sich nun in einem letzten Schritt einem Verständnis von KI anzunähern, werden noch zwei prägende Eigenschaften künstlich intelligenter Systeme vorgestellt – die Lernfähigkeit[807] sowie die (technische) Autonomie. Diese beiden

802 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 1.
803 *Husbands* 2014.
804 *Ertel* 2021, 20; *Russell/Norvig* 2021, 54.
805 *Münster* 2022, 18; *Schmid* 2019, 26.
806 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 1. Zu den im Rahmen von autonomen Fahrzeugen zum Einsatz kommenden Instrumenten zur Erfassung der Umgebung *Wigger* 2020, 56 ff.
807 Teilweise ist auch von Adaptivität die Rede, bspw. *Franklin/Graesser* 1997, 29; *Riehm/Meier* 2019, 6 f.

Begriffe sind in der Diskussion über KI und intelligente Agenten – von Seiten der KI-Forschung[808] wie aber auch bspw. von Seiten der Rechtswissenschaft[809] – omnipräsent. Gleichzeitig besteht aber keine Einigkeit über die Begriffsinhalte und das Verhältnis der beiden Begriffe zueinander; vielmehr bestehen dazu schier unüberschaubare und sich vielmals widersprechende Ansätze. Dennoch soll hier eine Annäherung stattfinden.

Zunächst zur *Lernfähigkeit*:

> „When designing agent systems, it is impossible to foresee all the potential situations they may encounter and specify their behavior optimally in advance. […] Agents therefore have to *learn* from, and *adapt* to, [sic!] their environment.[810]
>
> „Bei der Entwicklung von Agentensystemen ist es unmöglich, alle möglichen Situationen, auf die sie stoßen könnten, vorherzusehen und ihr Verhalten im Voraus optimal zu spezifizieren. [...] Agenten müssen daher von ihrer Umgebung *lernen* und sich an sie *anpassen*."
>
> [Hervorhebungen durch die Verf.]
>
> „A learning agent changes its behavior based on its previous experience."[811]
>
> „Ein lernender Agent ändert sein Verhalten basierend auf vorangegangenen Erfahrungen."
>
> „An agent is *learning* if it improves its performance on future tasks after making observations about the world."[812] […] „All agents can improve their performance through *learning*."[813]
>
> „Ein Agent *lernt*, wenn er seine Leistung bei zukünftigen Aufgaben verbessert, nachdem er Beobachtungen über die Umwelt gemacht hat. […] Alle Agenten können ihre Leistung durch *Lernen* verbessern."
>
> [Hervorhebungen durch die Verf.]
>
> „Ein lernendes rationales System bewertet (mittels Wahrnehmung) nach Ausführung einer Handlung den neuen Zustand der Umgebung, um festzustellen, wie erfolgreich sein Verhalten war, und passt daraufhin seine Regeln des Schlussfolgerns und seine Entscheidungsverfahren an."[814]

Die obigen Zitate zusammengefasst: Eine lernende KI bzw. ein *lernender* Agent ist in der Lage über einen Rückkopplungsmechanismus den Output im Hinblick auf das vorgegebene Ziel mit der Zeit zu verbessern. Auf einen bestimmten Eingabewert (Input) x erfolgt also nicht mehr der Ausgabewert (Output) y, sondern der Ausgabewert z, welcher dem vorgegebenen Ziel besser entspricht. Als Grundlage für diese „Verbesserung" dienen die vorangegangenen Wahrnehmungen des Systems. Die KI ist also lernfähig, weil der Output nicht nur auf Grund des

808 Siehe überblicksartig *Franklin/Graesser* 1997, 22 ff.; ebenso *Alonso* 2014, 235; *Beierle/Kern-Isberner* 2019, 398 f.
809 Vgl. bspw. *Fateh-Moghadam* ZStW 131 (2019), 863 (875); *Yuan* RW 9 (2018), 477 (480); *Hilgendorf* 2015b, 13 f.; *Gleß/Weigend* ZStW 126 (2014), 561 (563 f.); *Beck* 2015, 11; *Zech* 2020, A 37.
810 *Alonso* 2014, 235.
811 *Franklin/Graesser* 1997, 29.
812 *Russell/Norvig* 2021, 669.
813 *Russell/Norvig* 2021, 78.
814 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 3.

entwicklerseitig vorgegebenen, sondern auch auf Grund des im Laufe der Zeit gesammelten Wissens angepasst werden kann.

Ein einfaches Beispiel einer lernfähigen KI ist ein Pokerspiel-Agent: Er verfügt über ein Grundwissen (Spielregeln, mögliche Spielzüge etc.) und ein Repertoire an bestimmten Handlungen (Spielzüge). Sobald der Agent die ihm zugeteilten Spielkarten und von den anderen Spielern abgelegten Karten wahrnimmt, ist er in der Lage, dieses zusätzliche erlangte Wissen zu verarbeiten und seine nächsten Spielzüge im Hinblick auf das vorgegebene Ziel „Gewinne das Spiel!" zu optimieren. Essentiell ist Lernfähigkeit auch für ein autonom fahrendes Fahrzeug: Dieses muss in jeder Verkehrssituation sowohl auf Grund des vorgegebenen Wissens als auch auf Grund der aktuell wahrgenommenen Verkehrssituation sein Verhalten anpassen, etwa im Hinblick auf das Ziel: „Überfahre keine Fußgänger!"

Diese Beispiele illustrieren das Bedürfnis für eine lernfähige KI: KI-Systeme sollen Aufgaben bewältigen, bei denen sie sich in einer komplexen und dynamischen Umwelt bewähren müssen. Dazu gehört etwa der Straßenverkehr, insbes. innerorts mit seinen unzähligen Verkehrsteilnehmern, ein Pokerspiel mit Kartenvariationen, Spielstrategien etc. oder – im Falle von Spracherkennungsprogrammen – die gesprochene Sprache mit einer Vielzahl an Sprechern, Aussprachen oder Betonungen. Seitens der Entwicklerinnen ist es unmöglich, alle Veränderungen der Umwelt vorherzusehen, die mit der Zeit eintreten. KI-Systeme, die in einer komplexen und dynamischen Umwelt rational operieren sollen, müssen in der Lage sein, nicht nur vorhandenes, sondern auch neues Wissen aufnehmen und verarbeiten zu können.

Zur (technischen) *Autonomie* von KI-Systemen:

> „*autonomy:* agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state […]."[815]
>
> „Autonomie: Agenten agieren ohne direkte Einwirkung von Menschen oder anderen und haben eine gewisse Kontrolle über ihre Handlungen und ihren internen Zustand […]."
>
> „By *autonomy* researchers mean the ability of the systems to make their own decisions and execute tasks on the designer's behalf. […] It is precisely this autonomy that defines agents. Traditionally, software systems execute actions […] automatically."[816]
>
> „Unter Autonomie verstehen Forscher die Fähigkeit der Systeme, ihre eigenen Entscheidungen zu treffen und Aufgaben im Sinne des Entwicklers auszuführen. [...] Es ist genau diese Autonomie, die Agenten definiert. Traditionell führen Softwaresysteme Aktionen [...] automatisch aus."
>
> „To the extent that an agent relies on the prior knowledge of its designer rather than on its own precepts, we say that the agent lacks autonomy. A rational agent should be autonomous – it should learn what it can to compensate for partial or incorrect

815 *Wooldridge/Jennings* The Knowledge Engineering Review 10 (1995), 115 (116).
816 *Alonso* 2014, 235.

> prior knowledge. For example, a vacuum-cleaning agent that learns to foresee where and when additional dirt will appear will do better than one that does not. […] Hence, the incorporation of learning allows one to design a single rational agent that will succeed in a vast variety of environments."[817]
>
> „In dem Maße, in dem sich ein Agent auf das Vorwissen seines Entwicklers und nicht auf seine eigenen Regeln verlässt, sagen wir, dass es dem Agenten an Autonomie mangelt. Ein rationaler Agent sollte autonom sein – er sollte lernen, was er kann, um teilweises oder falsches Vorwissen zu kompensieren. Zum Beispiel wird ein Staubsauger, der lernt, vorauszusehen, wo und wann zusätzlicher Schmutz auftritt, besser agieren als ein Agent, der dies nicht tut. […] Die Einbeziehung des Lernens ermöglicht es also, einen einzigen rationalen Agenten zu entwickeln, der in einer Vielzahl von Umgebungen erfolgreich sein wird."

Die obigen Zitate vermitteln bereits einen Eindruck davon, dass der Begriff der Autonomie nicht einheitlich verwendet wird und dass sich insbes. Überschneidungen zur Lernfähigkeit ergeben. Um sich diesem Begriff dennoch anzunähern, bedarf es zwei grundlegender Schritte:

(1) Autonomie im Zusammenhang mit KI-Systemen ist grundsätzlich unabhängig zu sehen von einem „strengen" Autonomiekonzept wie es in Bezug auf den Menschen diskutiert wird.[818] Die Bezeichnung eines KI-Systems als autonom bedeutet also nicht, dass ihm von seinem Entwickler so etwas wie Freiheit im Kant`schen Sinne verliehen wurde.[819] Weniger irreführend wirkt daher der Begriff der „*technischen Autonomie*."[820]

(2) Autonomie im Zusammenhang mit einem KI-System beschreibt das Verhältnis zwischen Entwickler bzw. Nutzer einerseits und KI-System andererseits. Sie bezieht sich nicht auf eine bestimmte singuläre Eigenschaft des Systems, sondern auf das Ausmaß seiner Abhängigkeit bzw. Unabhängigkeit von Entwickler- und Nutzervorgaben. Darin stimmen die oben zitierten Ansätze grundsätzlich überein.[821]

Blickt man noch einmal auf die obigen Zitate, scheinen sich die Ansätze dann zu unterscheiden:

Einerseits soll ein Agent autonom sein, wenn er in der Lage ist, bestimmte *Aktionen* auszuführen, ohne dazu im Detail durch vom Entwickler vorgegebene

817 *Russell/Norvig* 2021, 60.
818 Vgl. in diesem Zusammenhang aber die Ausführungen zur schwachen und starken KI unten Kap. 2, A.IV.
819 *Yuan* RW 9 (2018), 477 (481); *Hubig/Harras* 2014, 52 ff.
820 *Kirn/Müller-Hengstenberg* KI 29 (2015), 59 (60); *Hubig/Harras* 2014, 48.
821 So schreiben *Russel* und *Norvig*: „To the extent that an agent relies on the prior knowledge of its designer rather than on its own precepts, we say that the agent lacks autonomy." Und *Wooldridge* und *Jennings* führen aus: „agents operate without the direct intervention of humans".

Regeln bzw. durch Eingaben des Nutzers angeleitet worden zu sein.[822] Der Agent agiere also nicht nur *automatisch*, indem er selbsttätig Handlungen nach vorgegebenen Regeln ausführt, sondern autonom, weil der Entwickler ihm nur ein Handlungsgerüst vorgegeben habe, innerhalb dessen er ohne weitere menschliche Vorgaben bestimmte Ausgaben produziert. Auf den Punkt gebracht: Ein KI-System sei autonom, weil es ohne konkrete menschliche Vorgaben selbst entscheide, bestimmte Handlungen vorzunehmen, und sie dann auch ausführe.[823] Darin unterscheide sich ein autonomes System von einem „*automatisierten*" System, welches zwar ohne menschliche Anleitung im Einzelfall eine Handlung ausführe,[824] diese Handlung aber auf im Detail vorgegebenen Regeln des Entwicklers beruhe und nicht auf einer selbsttätig getroffenen Entscheidung des Agenten.

Andererseits soll ein KI-System autonom sein, wenn es seine *Wissensbasis* auf Grund eigener Wahrnehmung über seine Umwelt erweitere bzw. modifiziere und damit unabhängig vom Wissen werde, das ihm der Entwickler vorgegeben habe.[825] Ein Agent sei autonom, weil er eigenes Umgebungswissen sammeln könne.

Auf den zweiten Blick jedoch unterscheiden sich die oben genannten Ansätze nur in ihrer jeweiligen Akzentuierung und stimmen in ihren Grundaussagen überein. Für alle Ansätze gilt: Ein KI-System ist autonom, weil es in einem gewissen Umfang unabhängig ist von Entwickler- und Nutzervorgaben. Diese Unabhängigkeit betrifft aber verschiedene Aspekte: Sie erfasst einerseits die *Wissensbasis* – ein KI-System erweitert bzw. modifiziert diese auf Grund eigener Wahrnehmungen über seine Umgebungen (Wissensautonomie) –, andererseits die Ebene der *Entscheidungsfindung* – ein KI-System entscheidet und handelt selbstständig und nicht nach im Einzelfall vorgegebenen Regeln (Entscheidungsautonomie).

Ein so verstandener Autonomiebegriff lässt noch einmal die Brücke schlagen zum Begriff der Lernfähigkeit. Ein lernfähiges KI-System ist in der Lage seinen Output anzupassen oder zu verbessern. Dies kann es, weil es „wissensautonom" ist, sich also eigenes Wissen über seine Umgebung aneignen kann, und eigene Entscheidungen trifft, ohne dass diese detailliert vom Entwickler oder im konkreten Fall vom Nutzer vorgegeben wären. Weil ein KI-System lernfähig ist, ist es autonom. Die Lernfähigkeit beschreibt eine Eigenschaft des KI-Systems, dessen Autonomie beschreibt ein gewisses Maß an Unabhängigkeit im Verhältnis zum

822 So führen *Wooldridge* und *Jennings* aus: „[…] agents operate without the direct intervention of humans […]" und nach *Alonso* ist Autonomie „the ability of the systems to make their own decisions".

823 Vgl. auch *Fateh-Moghadam* ZStW 131 (2019), 863 (875). Insbes. zur Abgrenzung von „nur" automatischen und autonomen Systemen *Sosnitza* CR 2016, 764 (765).

824 *Hubig* und *Harras* sprechen von automatischem Prozessieren im Sinne einer Unabhängigkeit von steuernden oder regelnden Benutzereingriffen, *Hubig/Harras* 2014, 48.

825 So *Russell/Norvig* 2021, 60.

Entwickler und Nutzer.[826] Zu unterscheiden ist zwischen *Wissensautonomie* – das KI-System kann sich eigenes Wissen über seine Umgebung aneignen und wird dadurch unabhängig von einer vorhandenen Wissensbasis – und *Entscheidungsautonomie* – das KI-System kann Entscheidungen treffen ohne konkrete Entwickler- oder Nutzervorgaben.

Der Begriff der Autonomie weist zuletzt auf die Stellung des Entwicklers bei der „Schaffung" von KI-Systemen hin. Der Entwickler entfernt sich von der Vorstellung des „klassischen" Programmierers, der ein Programm Zeile für Zeile in einer Programmiersprache erstellt, welches vom Computer direkt ausgeführt werden kann. In einer solchen „klassischen" Vorstellung kontrolliert der Entwickler jedes Detail der Maschine.[827] Bei der Entwicklung eines KI-System hingegen gibt der Entwickler nicht mehr jedes Detail vor, sondern er schafft einen Raum, innerhalb dessen das KI-System agieren kann.[828]

II. KI und maschinelles Lernen

Autonomie im Zusammenhang mit KI-Systemen bedeutet, das haben wir gerade gesehen, Wissens- und Entscheidungsautonomie. Im Folgenden geht es nun darum, durch welche Technik diese erreicht werden kann. Am Ende dieses Abschnitts wird dann ein weiterer Aspekt der Autonomie von KI-Systemen stehen, nämlich Autonomie bei der Regelbildung.

Menschliches Wissen speist sich aus zwei Grundquellen – Theorie und Erfahrung.[829] Übertragen auf die Schaffung von KI: Es gibt zwei grundsätzliche Herangehensweisen, einem System Wissen über seine Umgebung und das rationale Agieren in dieser zu vermitteln. Einerseits, indem die Entwicklerin „Theorie vermittelt" und Tatsachen- und Regelwissen vorgibt; andererseits, indem sie dem System die Möglichkeit gibt, aus *Erfahrung* (d.h. Daten)[830] eigenes Tatsachen- und Regelwissen zu generieren. Um dies an einem menschlichen Beispiel zu verdeutlichen: Menschen können eine Sprache erlernen, indem sie Vokabeln auswendig lernen und sich mit den grammatikalischen Strukturen der Sprache vertraut machen – Wie werden Vergangenheitsformen gebildet? Wie werden Wörter dekliniert? etc. Diese Vorgehensweise ist ein mühevolles Unterfangen und oftmals

826 Diese „Verwandtschaft" von lernendem und autonomen Agenten erklärt, weshalb Lernfähigkeit und Autonomie eines Agenten oft gleichgesetzt werden, vgl. bspw. *Zech* 2020, A 37. Ähnlich *Riehm/Meier* 2019, Rn. 6; *Gleß/Weigend* ZStW 126 (2014), 561 (563 f.); *Yuan* RW 9 (2018), 477 (481).

827 *Matthias* Ethics and Information Technology 6 (2004), 175 (181).

828 *Matthias* Ethics and Information Technology 6 (2004), 175 (181); vgl. auch *Zech* 2020, A 34; *Lenzen* 2018, 68. Zum dennoch erheblichen menschlichen Anteil bei der Entwicklung von KI siehe unten Kap. 2, A.II.9.

829 Vgl. Kant in der *Kritik der reinen Vernunft*, in der er von den »zwei Grundquellen« der Erkenntnis spricht, *Kant* 1787, B 74. Siehe zu dieser daraus abgeleiteten Unterscheidung von Wissen qua Erfahrung und Theorie *Kaminski/Resch/Küster* 2018, 253 f.; 259.

830 *Gründel* 2018.

fällt ein neidvoller Blick auf unsere Kinder. Kleinkinder haben noch nie etwas von Vokabeln oder grammatikalischen Strukturen gehört, dennoch erkennen sie früh Gegenstände, können sie bald darauf benennen und sind ab dem Alter von drei bis vier Jahren in der Lage, sich in meist grammatikalisch korrekten Sätzen auszudrücken. Sie lernen, wie ein Hund aussieht und dass dieses Objekt „Hund" genannt wird, weil die Eltern jedes Mal, wenn sie einen Hund sehen, darauf zeigen und „Hund" sagen. Kinder lernen grammatikalische Strukturen, weil ihre Eltern mit ihnen sprechen und sie aus dem Gehörten implizite grammatikalische Strukturen bilden. Kinder lernen Sprache aus Erfahrung, weil sie in der Lage sind zu generalisieren[831] bzw. Muster zu erkennen.[832]

Diesen Blick auf das Lernen der Kinder hat schon *Turing* in seinen Überlegungen zum „imitation game" geworfen:

> „Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain."[833]
>
> „Anstatt zu versuchen, ein Programm zu erzeugen, das den Verstand eines Erwachsenen simuliert, warum sollte man nicht lieber versuchen, ein Programm zu erzeugen, das den Verstand des Kindes simuliert? Wenn dies dann einer entsprechenden Ausbildung unterworfen würde, würde man das erwachsene Gehirn erhalten."

Turing stellte sich das kindliche Gehirn als unbeschriebenes Blatt vor, das entsprechend einfacher zu imitieren wäre.[834] Es müsse daher ein „child programme" – ein Kinderprogramm – geschrieben und dieses dann einem Lernprozess unterzogen werden. Diese Ausführungen überschrieb er mit dem Titel „Learning machines".[835]

In der Tat sind derzeit Ansätze prägend, die bei der „Erschaffung" intelligenter Agenten die Idee eines „child programme"[836] aufnehmen und auf machine learning bzw. maschinelles Lernen setzen.[837] Das Ziel maschinellen Lernens ist es, Wissen durch Erfahrung zu generieren, indem Lernalgorithmen aus Beispielsdaten ein komplexes Modell entwickeln.[838]

Der folgende Abschnitt gibt einen Überblick über einerseits die *Methoden*, die dem maschinellen Lernen zu Grunde liegen, wobei auf die künstlichen neuronalen Netze im Detail eingegangen wird, und andererseits über die *Lernformen*, die

831 *Ertel* 2021, 202.
832 *Lenzen* 2018, 59.
833 *Turing* Mind 59 (1950), 433 (456).
834 „Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. […] Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed." *Turing* Mind 59 (1950), 433 (456).
835 *Turing* Mind 59 (1950), 433 (454 ff.).
836 Siehe zur sog. Developmental Robotics, die zum Ziel hat, ein kontinuierliches Lernen in Robotern zu ermöglichen und Roboter auf diesem Weg eine „Kindheit" durchlaufen zu lassen, *Lenzen* 2018, 91 ff.
837 Zu den Parallelen des „child programme" und den heutigen Ansätzen zum maschinellen Lernen *Stephan/Walter* in: *Turing* 2021, 182.
838 *Döbel/Leis/Vogelsang u. a.* 2018, 9.

angewandt werden, um – bildlich gesprochen – aus einem „Kinderalgorithmus" einen „Erwachsenenalgorithmus" zu bilden. Dem vorangestellt ist ein Kurzüberblick über die sog. symbolbasierte KI, die jedenfalls vom klassischen Ansatz her auf das Generieren von Wissen durch Theorie setzt. Dieser Überblick verdeutlicht die Schwächen eines solchen Ansatzes und insofern das Bedürfnis nach dem Einsatz maschinellen Lernens. Ergänzend wird auf die Bedeutung von maschinellen Lernmethoden als statistische Verfahren, auf sog. offline- und online-Lernen, auf die Bedeutung von Big Data im Zusammenhang mit maschinellem Lernen und zuletzt auf die menschlichen Akteure beim maschinellen Lernen eingegangen.

1. Symbolbasierte KI – Wissen durch Theorie

In den Anfängen der KI-Forschung, d.h. von den 1950ern bis in die 1980er hinein, dominierten sog. symbolbasierte Ansätze.[839] Danach sollte Computern Wissen „wohlgeordnet" vorgegeben werden. „Wohlgeordnet" bedeutet in Form von Symbolen, die für Wissen und Fakten über die Umwelt stehen, und von Regeln, wie diese Symbole zu kombinieren und zu manipulieren sind (sog. Inferenzmaschine).[840] Symbolbasierte KI steht damit in der Tradition ursprünglicher, auf formaler Logik aufbauender Computertechnik.[841] Oder um die von *Turing* vorgenommene Differenzierung zwischen einem „adult mind" und einem „child mind" zu bemühen: Symbolbasierte KI hatte zum Ziel, ein erwachsenes Gehirn nachzuahmen.

Schlüsselkonzepte symbolbasierter Ansätze sind die heuristische Suche – die Suche mit Hilfe von Faustregeln – und das Planen – Problemlösungsanalyse durch hierarchische Strukturen von Zielen und Unterzielen.[842] Die Aufgabe eines Programms, das auf einem symbolbasierten Ansatz beruht, wird durch einen Suchbereich dargestellt, d.h. durch eine Menge von Möglichkeiten,[843] innerhalb derer die Lösung liegt und innerhalb derer sie auch gefunden werden muss.[844]

Aufbauend auf dem symbolbasierten Ansatz entstanden in den 1970ern und 1980ern sog. *Expertensysteme*. Dies sind Programme, die das (bspw. medizinische) Wissen menschlicher Experten als eine Reihe von Wenn-dann-Regeln darstellen und die dazu dienen können, Nicht-Experten auf dem betreffenden Gebiet

839 *Boden* 2014, 89.
840 Vgl. *Boden* 2014, 90; *Russell/Norvig* 2021, 1032 f.
841 Im Sinne eines von-Neumann-Computers oder einer Turing-Maschine, *Boden* 2014, 89.
842 Siehe dazu die grundlegenden Arbeiten insbes. von *Newell/Shaw/Simon* Psychologial Review 65 (1958), 151; *dies.* 1958. Zu einer zusammenfassenden Beschreibung *Boden* 2014, 90; *Franklin* 2014, 25.
843 Beispielsweise die Menge der erlaubten Spielzüge im Schach.
844 *Boden* 2014, 90.

zu beraten.[845] Ein Höhepunkt der symbolbasierten KI-Entwicklung ist der Sieg des IBM-Schachcomputers „Deep Blue" über den damaligen Schachweltmeister Gary Kasparov, der mit einer Kombination aus viel Rechenleistung und Suche nach dem besten (rationalen) Spielzug erreicht wurde.[846]

Wie die Begriffe heuristische Suche und Planen suggerieren, simulieren Programme, die auf symbolbasierten Ansätzen beruhen, *bewusste* menschliche Gedankenvorgänge. Daraus ergibt sich aber auch ihr Nachteil bzw. ihr eingeschränkter Anwendungsbereich. Die Entwicklung von Programmen, die auf symbolischer KI beruhen, setzt explizites Wissen und explizit bekannte Regeln voraus; beides ist aber nicht immer vorhanden bzw. zwar vorhanden, aber nur mit großem Aufwand in eine Symbolsprache (d.h. in eine Programmiersprache)[847] zu übersetzen:

> „The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order* it to perform."[848]
>
> „Die analytische Maschine erhebt keinen Anspruch, irgendetwas zu *erschaffen*. Sie kann *das* tun, *von dem wir wissen, wie wir* ihr *befehlen können* es zu tun."
>
> „[…] one has to provide the defining features of that "something" […]. Often this is not possible. For example, you may want to teach a program to recognize cats. But can you define "cat"? Can you even identify and list the relevant features seen in photographs of a dozen cats?"[849]
>
> „[...] man muss die bestimmenden Merkmale des "Etwas" angeben [...]. Oftmals ist dies nicht möglich. Vielleicht möchten Sie zum Beispiel einem Programm beibringen, wie man Katzen erkennt. Aber können Sie "Katze" definieren? Können Sie überhaupt die relevanten Merkmale ermitteln und aufzählen, die auf Fotos von einem Dutzend Katzen zu sehen sind?"

Symbolbasierte Ansätze stoßen also an ihre Grenzen, wenn es darum geht implizites oder umfangreiches Wissen und die dazugehörigen Regeln zu symbolisieren, d.h. in eine Programmiersprache zu integrieren. Dazu gehört bspw. die Bilderkennung – siehe dazu obiges Beispiel der Katze – sowie die Sprach- und Texterkennung. Was also sind die prägenden Merkmale einer Katze, was die prägenden Merkmale eines geschriebenen oder gesprochenen „A"? Ähnliches gilt bei der Entwicklung autonomer Fahrzeuge: Die Vielzahl an möglichen Verkehrssituationen kann schwerlich vorhergesehen werden. In vielen Bereichen sind die

845 Vgl. folgendes Beispiel nach *Boden* 2014, 91: „IF *these* symptoms are observed in the patient, AND the patient is an adult male, THEN infer that *this* bacterium is responsible, AND recommend *this* drug, given in *that* dosage for that length of time, as the treatment." – „WENN *diese* Symptome beim Patienten beobachtet werden UND der Patient ein erwachsener Mann ist, DANN schlussfolgere, dass *dieses* Bakterium dafür verantwortlich ist UND empfehle *dieses* Medikament in *dieser* Dosierung für diesen Zeitraum als Behandlung."

846 Vgl. *Franklin* 2014, 23; *Divisio* (Hrsg.) 2019b.

847 *Boden* 2014, 89.

848 Dieser Auszug stammt aus der 1843 veröffentlichten Analyse der britischen Mathematikerin *Ada Lovelace* zur analytischen Maschine von *Charles Babbage*, zitiert nach *Turing* Mind 59 (1950), 433 (450).

849 *Boden* 2014, 95 f.

Ermittlung und Symbolisierung von Wissen und Regel entweder nicht oder nur mit einem enormen Aufwand möglich.[850]

Angesichts der Erkenntnis, dass praktisch nie alle denkbaren Vorbedingungen explizit angegeben werden können, stagnierten in den 1980ern und 1990ern symbolische Ansätze und insbes. die beschriebenen Expertensysteme.[851] Symbolbasierte KI mit ihrem Anspruch, dem Computer Wissen „wohlgeordnet" vorzugeben, wird seitdem auch GOFAI genannt – good old fashioned artificial intelligence.[852]

2. *Von symbolbasierter KI zum maschinellen Lernen – Wissen durch Erfahrung*

Angesichts der beschriebenen Grenzen symbolbasierter KI ist derzeit das maschinelle Lernen die dominierende Methode zur Schaffung von KI-Systemen. Als prominentes Beispiel steht dafür das Programm AlphaGo der Google-Tochter DeepMind, welches in den Jahren 2016 und 2017 zwei der weltbesten Go-Spieler, den Südkoreaner Lee Sedol und den Chinesen Ke Jie, schlug.[853] Das Go-Spiel an Komplexität noch einmal übertreffend ist die Vorhersage von Proteinstrukturen auf der Grundlage der Aminosäuresequenz des Proteins, welche durch das System AlphaFold 2 (ebenfalls DeepMind) mit einer ähnlichen Genauigkeit getroffen werden kann wie bei experimentell bestimmten Strukturen.[854]

Um noch einmal den Sinn maschinellen Lernens zu verdeutlichen und die Grundfunktion zu erläutern, hierzu folgendes Beispiel zur Handschriftenerkennung:[855]

850 Vgl. auch *Russell/Norvig* 2021, 669.
851 *Döbel/Leis/Vogelsang u. a.* 2018, 15.
852 Vgl. nur *Boden* 2014.
853 Dazu bspw. *Döbel/Leis/Vogelsang u. a.* 2018, 28. Go gilt als viel komplexer und schwerer vorhersagbar als Schach und war nach dem Sieg von „Deep Blue" über Gary Kasparov (siehe oben Kap. 2, A.I.1.c.) demnach die neue Herausforderung. Diese konnte nicht über Methoden des GOFAI, sondern nur mit Hilfe von maschinellem Lernen erreicht werden.
854 *Callaway* Nature 588 (2020), 203 (203).
855 Beispiel und nachfolgende Darstellung nach *Döbel/Leis/Vogelsang u. a.* 2018, 12.

Ein symbolbasierter Ansatz müsste systematisch beschreiben, wie alle möglichen Varianten einer geschriebenen „7" der Ziffer „7" zugeordnet werden können („gehe so und so vor, um eine Ziffer 7 zu erkennen"). Beim maschinellen Lernen hingegen wird ein Lernalgorithmus geschaffen, der aus vielen Beispielen die entscheidenden Merkmale einer geschriebenen „7" herausfiltert und anschließend in der Lage ist, zu generalisieren, d.h. auch bei neuen Beispielen eine korrekte Zuordnung etwa zur Ziffer „7" vorzunehmen.[856]

> „Beim Maschinellen Lernen erzeugt der Lernalgorithmus ein Modell, das Beispieldaten generalisiert, so dass es anschließend auch auf neue Daten angewendet werden kann."[857]
> „[Maschinelles] Lernen ist ein Vorgang, Regeln zu bilden, die ein System in die Lage versetzen, bei zukünftigen Bearbeitungen derselben oder einer ähnlichen Aufgabe, sie besser zu erledigen."[858]
> Maschinelles Lernen ist Lernen aus Erfahrung in der Form von Daten.[859]

Um abschließend den Unterschied zwischen symbolischer KI und Ansätzen des maschinellen Lernens zu verdeutlichen: Symbolische KI operiert mithilfe eines *Top-Down-Ansatzes*, indem ein Problem und seine Teile mit Symbolen, Regeln und Repräsentationen deterministisch beschrieben werden – oder wiederum bildhaft: Symbolische KI hat den Anspruch, einen fertigen „Erwachsenenalgorithmus" zu schaffen.[860] KI hingegen, die auf maschinellem Lernen beruht, verfährt nach einem *Bottom-Up-Ansatz*; ein Problem wird zunächst auf niedrigerer Ebene beschrieben, indem erst ein „Kinderalgorithmus" geschaffen wird. Durch eine schrittweise Anpassung einer Vielzahl von Parametern kann das System schließlich die passenden Ergebnisse liefern.[861] Der angepasste Algorithmus, der „Erwachsenenalgorithmus" – in anderen Worten das ML-Modell – ist das Resultat dieses Lernprozesses.[862]

An dieser Stelle zeigt sich, dass die technische Autonomie von KI-Systemen um einen weiteren Aspekt erweitert werden kann: Zur Wissens- und Entscheidungsautonomie tritt eine Regelautonomie hinzu: Das System bildet die Regeln, nach denen es operiert, selbständig.

3. *Maschinelle Lernmethoden*

Turing hatte die Schwierigkeiten, die ein rein symbolbasierter Ansatz mit sich bringen würde, vorhergesehen und wie beschrieben eine Orientierung an einem „child mind" und die Entwicklung eines „child programme" vorgeschlagen.

856 *Döbel/Leis/Vogelsang u. a.* 2018, 12 f.
857 *Döbel/Leis/Vogelsang u. a.* 2018, 13.
858 *Kaminski/Glass* 2019, 130.
859 *Gründel* 2018.
860 *Divisio* (Hrsg.) 2019a.
861 *Divisio* (Hrsg.) 2019c; *ders.* (Hrsg.) 2019a.
862 *Kaminski/Glass* 2019, 130.

Zudem konstatierte er:

> „Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation […].“[863]
>
> „Intelligentes Verhalten besteht vermutlich in einer Abkehr von dem vollständig disziplinierten Verhalten, das mit dem maschinellen Berechnen[864] verbunden ist.“

So wie kindliche Gehirne eine andere Struktur als erwachsene Gehirne aufweisen und gerade diese ihnen das Lernen durch Erfahrung so einfach macht, müssen bei der Entwicklung lernender Maschinen andere Modelle und entsprechend andere Methoden zu Grunde gelegt werden.

Die derzeit bekannteste Methode maschinellen Lernens bilden die auf einem konnektionistischen Ansatz beruhenden *künstlichen neuronalen Netze*, teilweise wird gar maschinelles Lernen damit gleichgesetzt. Dies hat zur Folge, dass spezifische Eigenschaften künstlicher neuronaler Netze – insbes. zu beobachtende Black-Box-Effekte[865] – mit maschinellem Lernen insgesamt gleichgesetzt werden. Diesem Eindruck soll jedoch entgegengetreten werden, so dass entsprechend ein Kurzüberblick[866] über die verbreitetsten Lernmethoden gegeben wird.

Es lassen sich vier methodische Herangehensweisen unterscheiden, die jeweils von unterschiedlichen „Denkschulen“ geprägt sind.

a. Stochastik und Bayessche Verfahren

Grundlage des Bayesschen Verfahrens ist das sog. Bayes Theorem, das vom Mathematiker *Bayes* formuliert wurde. Es handelt sich dabei um eine mathematische Formel[867] für die Bestimmung einer bedingten Wahrscheinlichkeit, d.h. der Wahrscheinlichkeit, dass ein Ereignis (A) auftritt, unter der Bedingung, dass ein anderes Ereignis (B) eingetreten ist.[868] Damit kann bei der Bestimmung von Wahrscheinlichkeiten Vorwissen einbezogen werden.[869]

Im Kontext maschinellen Lernens eignet sich das Bayes Theorem dazu, das jeweils wahrscheinlichste Modell auf Basis der bestehenden Datenlage zu generieren. Schwerpunktmäßig geht es um das Schlussfolgern über zukünftige Ereignisse unter Unsicherheit durch Einbeziehung von Vorannahmen.[870]

863 *Turing* Mind 59 (1950), 433 (448).
864 Zu der schwierigen Übersetzung des Begriffs der Computation ins Deutsche *Lenzen* 2018, 36.
865 Siehe dazu unten Kap. 2, A.III.
866 In Anlehnung an *Döbel/Leis/Vogelsang u. a.* 2018, 16 ff.
867 Zur Formel *Ertel* 2021, 152.
868 *Döbel/Leis/Vogelsang u. a.* 2018, 17. Vgl. auch *Ertel* 2021, 145.
869 *Döbel/Leis/Vogelsang u. a.* 2018, 17.
870 *Döbel/Leis/Vogelsang u. a.* 2018, 17; *Danks* 2014, 156.

Beispiel: Aus dem Wissen darüber, ob ein Patient raucht oder nicht, ob das Wetter warm oder kalt ist, kann die wahrscheinlichste Diagnose hergeleitet werden, etwa Lungenerkrankung oder Erkältung.[871]

b. Analogismus

Analogistische Verfahren beruhen auf der Grundannahme, dass Objekte, die in Bezug auf bestimmte Merkmale große Ähnlichkeiten aufweisen, einer gemeinsamen Klasse angehören. Entsprechend werden Schlussfolgerungen auf Grund von Ähnlichkeiten gezogen.[872]

Im Zusammenhang mit maschinellem Lernen weisen analogistische Verfahren jedem Beispiel einen Wert zu, der in Abhängigkeit von ähnlichen Beispielen gewählt wird, für die bereits ein Wert vergeben wurde (= Regressionsanalyse).[873]

Beispiel: Sog. recommender-Systemen, etwa von Verkaufsplattformen, liegt der (analogistische) Gedanke zu Grunde, dass Kunden mit einer ähnlichen Bestellhistorie ähnliche Interessen haben. Entsprechend werden Vorschläge gemacht, die auf die Nutzer angepasst sind.[874]

c. Symbolismus

Symbolistische Verfahren im Zusammenhang mit maschinellem Lernen mögen auf den ersten Blick überraschen, wurde doch oben ausgeführt, dass es bei symbolischen Verfahren mehr um Wissensvermittlung durch Theorie geht, und dass die derzeitige KI-Forschung von diesem Ansatz weitgehend abgekommen ist.

Dennoch gibt es auch im Bereich des Symbolismus Verfahren, die beim maschinellen Lernen eingesetzt werden können, indem sie bestimmte Merkmale (Namen, Farben, Preisangaben, Temperaturwerte etc.) interpretieren. Sie sind jedoch nicht in der Lage zu berücksichtigen, dass dahinter Objekte mit Beziehungen oder gar Regeln zur Ableitung neuen Wissens stehen könnten.[875] Sehr verbreitet sind in diesem Zusammenhang sog. Entscheidungsbäume. Sie leiten eine Antwort her, indem sie sukzessive bestimmte Merkmale abfragen.[876]

Symbolische Lernverfahren kommen inzwischen auch in sog. hybriden Modellen zum Einsatz, die den Vorteil symbolischer Methoden mit den Vorteilen der sonstigen maschinellen Lernmethoden verknüpfen.[877]

871 *Döbel/Leis/Vogelsang u. a.* 2018, 17.
872 *Döbel/Leis/Vogelsang u. a.* 2018, 18.
873 *Döbel/Leis/Vogelsang u. a.* 2018, 18.
874 *Döbel/Leis/Vogelsang u. a.* 2018, 19.
875 *Döbel/Leis/Vogelsang u. a.* 2018, 22.
876 *Ertel* 2021, 231 f.
877 *Lenzen* 2018, 74; *Döbel/Leis/Vogelsang u. a.* 2018, 73; *Ertel* 2021, 11.

d. Konnektionismus – künstliche neuronale Netze

Der Konnektionismus orientiert sich bei der Entwicklung maschineller Lernverfahren an der Funktionsweise des menschlichen Gehirns. Lernen im menschlichen Gehirn findet durch eine Veränderung der Stärke der Verbindungen zwischen einzelnen Nervenzellen, der Neuronen, statt. Dieser Lernprozess wird in vereinfachter Form durch die Bildung sog. künstlicher neuronaler Netze nachgeahmt (siehe dazu sogleich unter 4.)

Das maschinelle Lernen mit Hilfe künstlicher neuronaler Netze ist derzeit das wohl bekannteste Lernverfahren und ist für viele aktuelle Durchbrüche in der KI-Forschung verantwortlich. Daher soll auf dieses Lernverfahren im folgenden Abschnitt vertieft eingegangen werden.

4. Insbesondere: Künstliche neuronale Netze und deep learning

Künstliche neuronale Netze[878] basieren auf der Annahme, dass Kognition durch die Interaktion einer großen Anzahl von Neuronen entsteht, die gemeinsam in der Lage sind, hochkomplexe Prozesse wie Wahrnehmung, Sprache, motorische Kontrolle etc. auszuführen.[879] Es ist daher das Ziel, neuronale Netze nachzubilden, die aus künstlichen Neuronen bestehen, also aus mathematischen Gleichungen von Neuronen und ihren Aktivitäten.[880]

Der theoretische Grundstein für künstliche neuronale Netze wurde bereits in den 1940ern gelegt,[881] etwa von *McCulluch* und *Pitts*, die 1943 ein einfaches Modell mit binären Neuronen entwickelten,[882] oder von *Hebb*, der 1949 eine Regel zum Zustandekommen des Lernens in natürlichen wie künstlichen neuronalen Netzen formulierte.[883] Erst aber als in den 1980er Jahren rein symbolische Ansätze nicht die erhofften Erfolge erzielten, leistungsfähigere Computer umfangreichere und tiefere künstliche neuronale Netze ermöglichten und zudem ausreichend Daten[884] vorhanden waren, um sie zu trainieren, erzielten künstliche neuronale Netze ihren Durchbruch.[885]

Ein künstliches neuronales Netz besteht aus Neuronen, die teilweise auch Units oder Knoten genannt werden.[886] Ein einfaches künstliches neuronales Netz

878 Teilweise ist auch von „connectionist models“ (vgl. *Sun* 2014) bzw. subsymbolischer KI (vgl. *Lenzen* 2018, 71) die Rede.
879 *Sun* 2014, 109.
880 *Yuan* RW 9 (2018), 477 (489). Es geht also nicht um das Nachbilden biologischer neuronaler Netze und Neuronen; dies ist Gegenstand der Computational Neuroscience, vgl. *Russell/Norvig* 2021, 839.
881 Siehe zur Entwicklung *Sun* 2014, 110.
882 *McCulloch/Pitts* Bulletin of Mathematical Biology 52 (1990 [Nachdruck von 1943]), 99.
883 *Hebb* 1949.
884 Siehe zur Bedeutung von Big Data bei maschinellem Lernen unten Kap. 2, A.II.8.
885 *Döbel/Leis/Vogelsang u. a.* 2018, 21.
886 *Yuan* RW 9 (2018), 477 (489); *Russell/Norvig* 2021, 802 f.

besteht aus zwei Schichten (sog. layer): Einer Eingabeschicht mit Input-Neuronen, die Informationen von der Umwelt aufnehmen, und einer Ausgabeschicht mit Output-Neuronen, die Informationen als Ergebnis an die Umwelt zurückgeben.[887] Es können sich dazwischen aber auch weitere Schichten mit sog. Hidden-Neuronen befinden.[888] Je mehr Neuronen und Neuronenschichten ein Netz enthält, desto komplexer und leistungsfähiger wird dies – man spricht in diesem Zusammenhang auch von deep neural networks und entsprechend von deep learning.[889] Mit zunehmender Komplexität künstlicher neuronaler Netze steigt die erforderliche Rechenleistung.[890]

Zwischen den Neuronen bestehen Verbindungen. Je nach Stärke und Bedeutung der Verbindung hat diese eine bestimmte Gewichtung („weight"). Je stärker die Gewichtung ist, desto größeren Einfluss nimmt ein Neuron über die Verbindung auf ein anderes Neuron; ist die Gewichtung null, übt ein Neuron über seine Verbindung hingegen keinen Einfluss auf das andere Neuron aus.[891]

Der Vorteil künstlicher neuronaler Netze besteht in einer „eingebauten" Lernfähigkeit; sie eignen sich daher für die Erfassung von implizitem Wissen, wenn also kein oder nur geringes systematisches Lösungswissen vorliegt und die Eingabeinformationen zum größten Teil unpräzise sind.[892] Künstliche neuronale Netze werden also z.B. eingesetzt bei der Spracherkennung, der Texterkennung, der Bild- und der Gesichtserkennung.[893] Zudem sind subsymbolische Verfahren fehlertolerant und haben die Fähigkeit zu generalisieren (d.h. sie können vorhandenes Wissen auf unterschiedliche Situationen anwenden).[894]

Künstliche neuronale Netze werden entwicklerseitig nicht programmiert, sondern trainiert; der Trainings- bzw. Lernprozess vollzieht sich in einer Veränderung der Gewichtungen der Verbindungen zwischen den künstlichen Neuronen. Um im obigen Beispiel des Erkennens einer Katze zu bleiben. Das künstliche neuronale Netz bekommt als Input-Information eine Vielzahl von verschiedenen Bildern von Katzen gezeigt und erarbeitet sich die entscheidenden Merkmale von Katzen über eine Veränderung der Gewichtungen der Verbindungen und zwar nicht nur durch die Veränderungen einzelner Gewichtungen, sondern aller Gewichtungen im Modell.[895] Bekommt das so trainierte künstliche neuronale Netz dann ein noch nicht bekanntes Bild gezeigt, gibt es anhand der erfassten

887 *Lenzen* 2018, 55; *Döbel/Leis/Vogelsang u. a.* 2018, 19.
888 *Yuan* RW 9 (2018), 477 (490); *Russell/Norvig* 2021, 801, 805.
889 *Lenzen* 2018, 55; *Yuan* RW 9 (2018), 477 (490).
890 *Lenzen* 2018, 55.
891 *Sun* 2014, 109; *Lenzen* 2018, 52 f.
892 *Sun* 2014, 109 f.
893 *Döbel/Leis/Vogelsang u. a.* 2018, 21.
894 *Sun* 2014, 109.
895 Vgl. *Lenzen* 2018, 57.

Inputinformationen eine Schätzung ab, ob auf dem Bild ebenfalls eine Katze dargestellt ist.

5. *Maschinelle Lernmethoden als statistische Verfahren*

Fast alle Modelle, die durch maschinelles Lernen erzeugt werden, sind *statistische Modelle*.[896] Diese Eigenschaft ist wichtig, um die Ausgabewerte, d.h. die Ergebnisse, die beim Einsatz intelligenter Agenten erzielt werden, bewerten zu können.

Dies soll noch einmal durch einen vergleichenden Rückgriff auf symbolische KI und KI, die auf maschinellem Lernen basiert, erläutert werden:

Nach über 30-jähriger Entwicklungszeit hat *Doug Lenat* 2016 seine Datenbank Cyc fertig gestellt. Diese Datenbank basiert ausschließlich auf symbolbasierter KI. Sie enthält 500.000 Begriffe, die über 17.000 verschiedene Arten von Beziehungen miteinander verbunden sind, und verfügt über eine Inferenzmaschine, die mehr als sieben Millionen Sätze enthält, wie die Begriffe miteinander zu verbinden sind. Gibt Cyc auf eine bestimmte Frage eine Antwort – reagiert sie auf einen bestimmten Input mit einem bestimmten Output – kann sie auch begründen, weshalb sie zu diesem Ergebnis gekommen ist. Denn Cyc beruht auf logischen Prinzipien, das dem System vorgegebene Wissen ist über definierte Regeln miteinander verbunden.[897]

2011 stellte IBM das von ihm entwickelte, auf künstlichen neuronalen Netzen beruhende System Watson vor, das das Ratespiel Jeopardy gegen zwei menschliche Konkurrenten gewann. Es handelt sich auch hier um eine große Datenbank, die u.a. große Mengen an Texten in natürlicher Sprache enthält (Wörterbücher, Enzyklopädien, Wikipedia) und innerhalb kurzer Zeit relevante Passagen und Fakten auffinden kann. Geht eine Frage ein, sucht eine Vielzahl von Algorithmen parallel nach der richtigen Ansicht. Je mehr Algorithmen unabhängig voneinander dieselbe Antwort erreichen, desto wahrscheinlicher ist es für das System, die richtige Antwort gefunden zu haben. Watson ermittelt also statistische Korrelationen und kann daher angeben, mit welcher Wahrscheinlichkeit er zu einem bestimmten Ergebnis kommt. Warum er einen bestimmten Eingabewert mit einem bestimmten Ausgabewert verknüpft, kann Watson dagegen – anders als Cyc – nicht angeben, da das vorhandene Wissen nicht über vorgegebene Regeln miteinander verknüpft ist.[898] Gleiches gilt für die Vorhersage von Proteinstrukturen durch das Programm AlphaFold 2: Die Vorhersage beinhaltet keine Angabe darüber, warum das Programm auf der Grundlage einer bestimmten Aminosäurensequenz zu einer bestimmten Proteinstruktur kommt.[899]

896 *Döbel/Leis/Vogelsang u. a.* 2018, 16; *Lenzen* 2018, 54, 72.
897 Siehe zu den Angaben in diesem Absatz *Lenzen* 2018, 71 f.
898 Siehe zu den Angaben in diesem Absatz *Franklin* 2014, 23 f.; *Lenzen* 2018, 72 f.
899 *Ball* 2020.

Maschinelle Lernverfahren als statistische Modelle produzieren also Ausgabewerte, die mit einer probabilistischen Einschätzung[900] verbunden sind, sie entdecken statistische Korrelationen in bestimmten Daten, sie garantieren aber keine Kausalitäten.[901]

6. *Maschinelle Lernformen*

Die unter den Ziff. 3. und 4. vorgestellten maschinellen *Lernmethoden* betreffen das hinter einem Lernalgorithmus stehende Modell.[902] Daneben werden für das maschinelle Lernen verschiedene *Lernformen* eingesetzt, die sich nach der Art der Trainingsdaten sowie dem Verhältnis zwischen Algorithmus und Trainer unterscheiden.[903] Je nach Lernform können unterschiedliche Lernmethoden zum Einsatz kommen.[904]

Eine erste Form maschinellen Lernens ist das sog. *überwachte Lernen* (supervised learning).[905] Ziel des überwachten Lernens ist es, eine Regel (eine Funktion f) zu erlernen, die das Verhältnis von Input und Output beschreibt (f(x) = y).[906] Sowohl der Input x als auch der Output y sind während des Lernvorgangs bekannt, d.h. zu jedem Trainingsbeispiel (bspw. Bild eines Hundes oder einer Katze) liegt die richtige Antwort vor (bspw.: Dieses Bild zeigt einen Hund, dieses eine Katze), die Trainingsdaten sind mit einem „Label" versehen.[907] Ob also der Lernalgorithmus während des Lernvorgangs die richtige Regel gebildet hat (also Hund und Katze unterscheiden kann), lässt sich anhand des vorhandenen Labels sofort zurückspiegeln.[908] Am Ende des Lernvorgangs soll eine Regel gebildet worden sein, die auch bei nicht mit Labeln versehenen Daten korrekte Ausgabewerte erzielt; um im obigen Beispiel zu bleiben: Der Algorithmus soll auch bei nicht gelabelten Bildern erkennen können, ob darauf ein Hund oder eine Katze zu sehen ist. Dies wird im Anschluss an die Trainingsphase in einer Testphase überprüft.[909]

Überwachtes Lernen wird nicht nur zur Bilderkennung eingesetzt, sondern auch zur Spracherkennung oder zum Filtern von Spam-E-Mails.[910]

Auch beim *unüberwachten Lernen* (unsupervised learning)[911] wird mit Daten trainiert mit dem Ziel, eine bestimmte Regel (eine Funktion (f)) zu bilden. Aller-

900 *Kaminski* 2020, 163. Vgl. auch *Cremers/Engländer/Gabriel u. a.* 2019, 11.
901 *Barredo Arrieta/Díaz-Rodríguez/Del Ser u. a.* IF 58 (2020), 82 (8). Vgl. auch *Danks* 2014, 158 f.
902 *Döbel/Leis/Vogelsang u. a.* 2018, 29.
903 *Döbel/Leis/Vogelsang u. a.* 2018, 29.
904 Vgl. überblicksartig *Ertel* 2021, 274; *Döbel/Leis/Vogelsang u. a.* 2018, 29 ff.
905 Dazu *Danks* 2014, 154; *Russell/Norvig* 2021, 671 ff.
906 *Kaminski/Glass* 2019, 130; *Döbel/Leis/Vogelsang u. a.* 2018, 25.
907 *Kaminski/Glass* 2019, 130; *Döbel/Leis/Vogelsang u. a.* 2018, 25.
908 *Döbel/Leis/Vogelsang u. a.* 2018, 25.
909 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 4.
910 *Döbel/Leis/Vogelsang u. a.* 2018, 26.
911 Dazu *Danks* 2014, 154; *Russell/Norvig* 2021, 671.

dings stehen beim unüberwachten Lernen lediglich Eingabewerte und keine Ausgabewerte fest, d.h. die Trainingsdaten sind nicht mit einem Label versehen.[912] Dennoch sollen in den Daten Strukturen oder Muster gefunden werden.[913] Ein Beispiel unüberwachten Lernens ist die sog. Clusteranalyse, bei der überprüft wird, ob in den Trainingsdaten lokale Häufungen, sog. Cluster, zu finden sind.[914] Die Clusteranalyse wird bspw. eingesetzt zur Segmentierung von Kundendaten, um bestimmte Zielgruppen zu identifizieren, oder zu einer ersten „Datenexploration", auf die dann weitergehende Analysen folgen.[915]

Lernen durch Verstärkung (reinforcement-learning)[916] bedeutet Lernen durch Belohnung und Tadel bzw. Versuch und Irrtum/Erfolg. Der Algorithmus soll durch Versuch und Irrtum/Erfolg herausfinden, welche Aktionen in einer bestimmten Situation gut sind und welche nicht. Dafür wird ein Feedback in Form eines mathematischen Äquivalents zur „Belohnung" bzw. zum „Tadel" erteilt, wenn das Ziel erreicht bzw. verfehlt wird.[917] Die Maschine soll ihren Algorithmus mit Hilfe des Feedbacks so anpassen, dass sie die Aktionen auswählt, welche die Nutzenfunktion maximieren.[918] Verstärkendes Lernen kommt bspw. in der Robotik zum Einsatz – Roboter erlernen das Gehen durch Versuch und Irrtum[919] – oder in Empfehlungssystemen[920] im Internet, die Kaufvorschläge generieren.

7. *Offline- und online-learning*

Eine letzte wichtige Differenzierung bei der Betrachtung des maschinellen Lernens ist die Differenzierung zwischen offline-learning und online-learning,[921] die Rede ist auch von lernenden und im Betrieb weiterlernenden Modellen.[922]

Zum *Offline-Lernen*: Es wird zunächst ein untrainiertes Modell geschaffen, das dann mit Hilfe von Trainingsdaten seinen Lernalgorithmus nach und nach anpasst (Trainingsphase). In einer daran anschließenden Phase wird das erlernte Modell mit Hilfe von weiteren Trainingsdaten auf seine Funktionsfähigkeit getestet (Testphase).[923] Bevor nun das KI-Modell aus seiner Trainingsumgebung in eine reale Umgebung entlassen wird, wird es „eingefroren", es „geht offline":[924]

912 *Döbel/Leis/Vogelsang u. a.* 2018, 6; *Kaminski/Glass* 2019, 131.
913 *Kaminski/Glass* 2019, 131.
914 *Kaminski/Glass* 2019, 131.
915 *Döbel/Leis/Vogelsang u. a.* 2018, 26.
916 *Russell/Norvig* 2021, 671, 840 ff.
917 *Döbel/Leis/Vogelsang u. a.* 2018, 28.
918 *Döbel/Leis/Vogelsang u. a.* 2018, 28.
919 *Döbel/Leis/Vogelsang u. a.* 2018, 28; *Franklin* 2014, 26.
920 *Unabhängige hochrangige Expertengruppe für Künstliche Intelligenz* 2018, 4.
921 *The Royal Society* 2017, 20.
922 Vgl. bspw. *Döbel/Leis/Vogelsang u. a.* 2018, 203. *Wigger* spricht ähnl. von „gelernten" und „selbstlernenden" Systemen", *Wigger* 2020, 55.
923 *Lenzen* 2018, 57.
924 *Lenzen* 2018, 58; *Zech* 2020, A 37; *Wigger* 2020, 55 f.; *Mayrhofer* 2023, 27.

Der sich in der Trainings- und Testphase schrittweise optimierende Lernalgorithmus kann sich nicht mehr weiter verändern. Während sich also in der Trainingsphase die Funktion f zwischen Input und Output kontinuierlich verändert hat, ist sie in der Anwendungsphase konstant. Die Regelautonomie des Systems endet also mit der Einsatzphase, das System kann die Transformationsregeln, auf Grund derer es operiert, nicht mehr selbständig anpassen. Jedes nachfolgende Training des Lernalgorithmus erfolgt in einer Trainingsumgebung und wird über Softwareupdates ausgeführt.[925] Auf diese Weise besteht die Möglichkeit, die Funktionsfähigkeit des Lernalgorithmus zu prüfen, bevor das KI-Modell mit einem Nutzer interagiert.[926]

Online-Lernsysteme hingegen durchlaufen zwar auch eine Trainings- und Testphase. Jedoch wird der Lernalgorithmus nicht „eingefroren", bevor er in einer realen Umgebung zum Einsatz kommt.[927] Im Einsatz wird die Funktion f zwischen Input und Output kontinuierlich angepasst, das KI-Modell lernt im Einsatz weiter.[928] Die Regelautonomie bleibt also auch in der Einsatzphase erhalten. Eine Überprüfung der Anpassungen des Lernalgorithmus ist nicht möglich.[929]

Offline-Lernen ist derzeit noch der übliche Ansatz bei maschinellem Lernen.[930] D.h., im Einsatz befindliche Lernalgorithmen verändern sich nicht mehr. Diese Systeme mögen zwar auch während des Einsatzes wissens- und entscheidungsautonom sein, weil sie sich eigenes Wissen über ihre Umgebung aneignen können und weil sie Entscheidungen ohne konkrete Entwickler- oder Nutzervorgaben treffen können. Sie sind aber nicht regelautonom in dem Sinne, dass sie während des Einsatzes ihre Transformationsregeln verändern könnten.

Sofern Online-Lernsysteme eingesetzt werden, sind sie häufig mit der Lernform des reinforcement learning, dem verstärkenden Lernen, verbunden.[931] Lernalgorithmen werden also in einer realen Umgebung eingesetzt und passen sich auf Grund von positivem (Lob) und negativem (Tadel) Feedback kontinuierlich an, so bspw. recommender-Systeme. Daneben sind inzwischen so bedeutende Anwendungen wie ChatGPT als online-Lernsysteme konzipiert, so dass die Bedeutung dieser Systeme absehbar zunehmen wird.

925 *Gründel* 2018.
926 Siehe dazu und zum gesamten Absatz *The Royal Society* 2017, 20.
927 *Lenzen* 2018, 58; *Mayrhofer* 2023, 28.
928 Vgl. auch *Cremers/Engländer/Gabriel u. a.* 2019, 11; *Wigger* 2020, 56; *Lohmann* 2021, 89 f.
929 Siehe dazu und zum gesamten Absatz *The Royal Society* 2017, 20.
930 *Gründel* 2018. Im Bereich des autonomen Fahrens kommen Online-Lernsysteme bislang bspw. nur für nicht sicherheitsrelevante Funktionen zum Einsatz, *Wigger* 2020, 56.
931 Vgl. auch *Matthias* Ethics and Information Technology 6 (2004), 175 (179).

8. Big Data und maschinelles Lernen

Die Ablösung symbolbasierter KI durch maschinelle Lernverfahren lag in der oben beschriebenen Begrenztheit symbolbasierter KI und daneben in einer zunehmenden computertechnischen Leistungskraft begründet. Ein weiterer Erfolgsfaktor für den „Siegeszug" maschineller Lernverfahren war aber auch die zunehmende Verfügbarkeit großer Datenmengen als Trainingsmaterial (Big Data).[932] Denn maschinelles Lernen bedeutet Lernen aus Erfahrung in Form von Daten und gerade bei überwachten und unüberwachten Lernformen bedarf es eines großen Datenbestands. Maschinelles Lernen ist insofern auch eine „datengestützte Technologie".[933]

Einerseits gilt, dass ein Lernalgorithmus sein Lernziel der Regelbildung umso besser erreichen und Fehlerquoten verringern kann, je mehr Trainingsdaten ihm zur Verfügung gestellt werden.[934] Entscheidend sind andererseits aber nicht nur die Quantität der Trainingsdaten, sondern auch deren Qualität. Denn die Qualität der Daten bestimmt darüber, was eine KI am Ende zu leisten in der Lage ist – je besser die Daten, umso besser das Modell.[935]

Problematisch ist es, wenn Daten mit einem sog. Bias verbunden sind. Bias bedeutet in diesem Kontext, dass der Datensatz nicht die Wirklichkeit der Umgebung widerspiegelt, in der ein trainiertes Modell später eingesetzt werden soll, er also nicht repräsentativ ist.[936] Geht ein solcher Bias in den Lernprozess ein, wird sich dieser auch in der Einsatzphase zeigen – „bias in – bias out".[937] Wenig repräsentative Trainingsdaten können also zu einer Regelbildung führen, die zwar mit den Trainingsdaten korrekte Ausgabewerte produziert, nicht aber bei der Anwendung auf ungelabelten Daten. Werden zur Erkennung von Flugzeugen bspw. nur Bilder von Flugzeugen verwendet, die sich in der Luft befinden, kann es sein, dass der Algorithmus den blauen Hintergrund als relevantes Merkmal heranzieht; entsprechend wird das trainierte Modell später bei Bildern von Flugzeugen, die sich in einem Hangar befinden, versagen, obwohl es zuvor in der Trainingsphase „zuverlässig" Flugzeuge erkannt hat.[938] Weitere Beispiele für einen Daten-Bias in Form eines racial oder gender bias: Ein Objekterkennungssystem von Google, welches dunkelhäutige Personen als Gorillas kennzeichnet, weil das System nicht mit ausreichend Bildern dunkelhäutiger Personen trainiert wurde,[939] ein Fieberthermometer wurde in der Hand eines weißen Menschen als

932 *Russell/Norvig* 2021, 44.
933 *Döbel/Leis/Vogelsang u. a.* 2018, 47. Zudem *Yuan* RW 9 (2018), 477 (482).
934 *Döbel/Leis/Vogelsang u. a.* 2018, 47; *Lenzen* 2018, 60.
935 *Gründel* 2018; *Lenzen* 2018, 60; *Cremers/Engländer/Gabriel u. a.* 2019, 11.
936 *Lionbridge AI* (Hrsg.) 2020; *Döbel/Leis/Vogelsang u. a.* 2018, 47.
937 Die Rede ist auch von „garbage in – garbage out", *Döbel/Leis/Vogelsang u. a.* 2018, 47.
938 *Döbel/Leis/Vogelsang u. a.* 2018, 47.
939 *Sommerer* 2020b, 102; *Kasperkevic* 2015.

Fernrohr, in der Hand eines dunkelhäutigen Menschen als Waffe identifiziert;[940] Spracherkennungsprogramme, die Frauenstimmen schlechter erkennen, weil sie v.a. mit Männerstimmen trainiert wurden.[941]

Die Regelbildung kann zudem verzerrt werden, wenn es sich um „weiterlernende" Modelle handelt, die also jenseits ihrer Trainings- und Testphase ihren Algorithmus kontinuierlich verändern. Der Chatbot Tay von Microsoft etwa musste abgeschaltet werden, weil er durch entsprechende Nutzereingaben im Internet rassistische Ausdrucksweisen erlernte und einsetzte.[942]

Ein weiterer Aspekt der Datenqualität ist der der Robustheit. Ein Algorithmus ist wenig robust, wenn kleine Änderungen in den Eingabewerten zu großen Änderungen in den Ausgabewerten führen. Dies ist bspw. der Fall, wenn Straßenschilder teilweise von Aufklebern oder Graffitis verdeckt sind, so dass autonome Fahrzeuge sie falsch interpretieren, etwa ein Stoppschild als Geschwindigkeitsbegrenzung.[943]

9. Menschliche Akteure beim maschinellen Lernen

Die Begriffe Künstliche Intelligenz, Autonomie, Lernfähigkeit und maschinelles Lernen sind wirkmächtig – ihnen wohnt die Tendenz inne, beim technischen Laien eine überhöhte Vorstellung von der Leistungsfähigkeit der Technik hervorzurufen. KI-Systemen werden Eigenschaften zugeschrieben, die sie so (noch) nicht haben. Und es werden auf diese Technik Vorstellungen von künstlichen „Wesen" projiziert, die eher filmischen und literarischen Vorbildern entsprechen.

Um dieser Tendenz entgegenzutreten, wurde oben etwa gezeigt, dass KI vom Ansatz her eine Imitation menschlicher Verhaltensweisen beschreibt, was u.a. durch das Modell des intelligenten Agenten und die damit verbundene Agentenarchitektur *sense-plan-act* verwirklicht werden soll. Ebenso wurde der Begriff der Autonomie von KI-Systemen aus seinem Bedeutungszusammenhang mit menschlicher Autonomie herausgelöst und als Relationsbegriff eingeordnet, der eine gewisse Unabhängigkeit des Systems von Entwickler- und Nutzervorgaben beschreibt. Auch in Bezug auf das maschinelle Lernen erscheint eine solche Klarstellung nötig, und zwar einerseits in Bezug auf die beteiligten „menschlichen" Akteure und andererseits in Bezug auf die Bedeutung des Lernens. Letzteres erfolgt unter dem Punkt IV.3., Ersteres sogleich:

Wie *Yuan* zu Recht hervorhebt, entspricht die Vorstellung, dass beim maschinellen Lernen, „‚irgendwie' Daten bereitgestellt werden, und der Algorithmus ‚ei-

940 *Döbel/Leis/Vogelsang u. a.* 2018, 48.
941 *Tatman* 2017, 53.
942 *Döbel/Leis/Vogelsang u. a.* 2018, 48.
943 *Döbel/Leis/Vogelsang u. a.* 2018, 52.

genverantwortlich' die Arbeit erledigt, […] nicht der Realität."[944] Um die Akteure und den Umfang ihrer Beteiligung im Rahmen des maschinellen Lernens herauszuarbeiten, bietet sich eine Betrachtung des Gesamtprozesses der Entwicklung von Modellen des maschinellen Lernens an. Dieser Gesamtprozess wird regelmäßig in fünf Phasen unterteilt:[945]

Die *Phase eins* beschäftigt sich mit dem Design eines Algorithmus und dessen Implementierung, indem das Design in eine Programmiersprache transferiert wird. Beteiligt sind in dieser Phase vornehmlich Wissenschaftler und Informatiker. Letztere werden meist im Auftrag von Unternehmen tätig, die mit der Implementierung kommerzielle Interessen verfolgen, teilweise wird diese Leistung aber auch unentgeltlich durch eine Gemeinschaft von Programmierern geleistet.[946]

Die *Phase zwei* betrifft die Methodenauswahl und dient v.a. der Datensammlung und -auswahl durch sog. data scientists.[947] Im obigen Punkt „Big data und maschinelles Lernen" wurde die Bedeutung nicht nur der Quantität, sondern v.a. auch der Qualität der verwendeten Daten bereits herausgearbeitet. Die Qualität der in der Trainingsphase erlernten Regelbildung liegt jedenfalls wesentlich in der Hand des data scientists, da er über die Qualität der Daten entscheidet: Repräsentative und robuste Datenmengen führen zu korrekter Regelbildung, nicht repräsentative und nicht robuste Datenmengen verzerren die Regelbildung und führen zu teilweise falschen Klassifizierungen. Die erlernte Funktion kann nur so gut sein wie die Datenmengen, die der Entwickler dem Modell zur Verfügung gestellt hat.[948]

In *Phase drei* wird eine maschinelle Lernmethode mit den Trainingsdaten zusammengebracht und es findet das eigentliche Training des Systems statt.[949] Der data scientist hat auf den Ausgang des Trainings Einfluss, indem er bestimmte Parameter[950] festlegen und bei Bedarf, d.h. wenn das Modell noch nicht gut genug ist, verändern kann.[951]

Phase vier beschreibt die Einbettung eines Systems in einen Anwendungskontext. Dieser ist besonders relevant, wenn es sich um maschinelle Entscheidungssysteme handelt, deren „Entscheidungen" von menschlichen Anwendern inter-

944 *Yuan* RW 9 (2018), 477 (492).
945 So z.B. durch *Borges/Grabmair/Krupka u. a.* 2018, 45 ff.; *Zweig* 2018, 17 ff. Die jeweils beschriebenen Phasen sind prozessual nicht identisch, stimmen jedoch inhaltlich weitgehend überein. Eine konzise Zusammenfassung findet sich bei *Yuan* RW 9 (2018), 477 (491 f.).
946 *Zweig* 2018, 17 f.
947 Zu dieser Phase *Zweig* 2018, 18 f.; *Borges/Grabmair/Krupka u. a.* 2018, 45 ff. (die jedoch diese Phase detailliert noch einmal unterteilen in „Data acquisition", „Preprocessing", „Transformation" und „Post-processing").
948 *Döbel/Leis/Vogelsang u. a.* 2018, 48.
949 Siehe dazu oben unter Punkt 3. und 6.
950 Siehe zu diesen Parametern *Borges/Grabmair/Krupka u. a.* 2018, 54.
951 *Zweig* 2018, 19; *Borges/Grabmair/Krupka u. a.* 2018, 46 f. (unter dem Punkt „Training and application of AI model").

pretiert und weiterverwendet werden sollen. In einer letzten *Phase fünf* kann ein Modell auf Grundlage der Anwendungserfahrung re-evaluiert werden.[952]

Die dargestellten Phasen zeigen, dass für die Erstellung eines ML-Modells in großem Umfang die Beteiligung menschlicher Akteure notwendig ist. Zwar bildet ein lernender Algorithmus ein Modell selbstständig ohne bzw. mit wenig Anleitung durch den Entwickler,[953] bzw. lernt sogar nach der Testphase weiter. Dennoch haben menschliche Akteure sowohl während als auch nach der Testphase einen Einfluss durch die Möglichkeit, Parameter anzupassen. Erst recht nehmen sie ganz erheblichen Einfluss auf das ML-Modell durch die Auswahl und Aufbereitung der Daten.

Die Betrachtung des Gesamtprozesses der Erstellung eines Modells zeigt auch, dass daran eine Vielzahl von Personen beteiligt sind, *Zweig* etwa spricht davon, dass „leicht Hunderte Personen" involviert sein können.[954]

III. Maschinelles Lernen und Black-Box-Effekte

> „Wenn ein Löwe sprechen könnte, wir könnten ihn nicht verstehen."[955]

Dieses Zitat *Ludwig Wittgensteins* illustriert die grundsätzliche Unmöglichkeit einer Verständigung zwischen sehr verschiedenen Lebensformen, die keine gemeinsame Lebenspraxis als Grundlage von Sprachverständnis haben.

Wittgensteins Bild vom sprechenden, aber unverständlichen Löwen leitet über zum nächsten Punkt – maschinelles Lernen und Black-Box-Effekte. Wir haben gesehen, dass technische Autonomie und Lernfähigkeit zwei prägende Eigenschaften von KI-Systemen sind, die mit Hilfe maschinellen Lernens erreicht werden können. Übertragen wir *Wittgensteins* Bild auf ein KI-System, das man auf Grund dieser Eigenschaften zwar nicht als Lebewesen, aber doch als etwas betrachten kann, was als nicht mehr nur vom Menschen beherrschtes Werkzeug erscheint: Wenn die KI sprechen könnte, könnten wir sie dann verstehen?

Turing und *Wiener*, die sich beide in frühen Grundlagenarbeiten mit KI beschäftigt haben, sahen dies nicht so; sie glaubten, dass mit Autonomie und Lernfähigkeit eine weitere Eigenschaft künstlich intelligenter Systeme hinzutreten werde, die ein „Verstehen" hindern werde:

> „An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behaviour. […] This is in clear contrast with normal procedure when using a machine to do computations: one's object is then to have a clear mental picture of the state of the machine at each moment in the computation […] Most of the

952 *Zweig* 2018, 19 f.
953 *Kaminski* 2020, 155.
954 *Zweig* 2018, 20. Siehe auch den Hinweis bei *Yuan* RW 9 (2018), 477 (491).
955 *Wittgenstein* 1984, 568.

programmes which we can put into the machine will result in its doing something that we cannot make sense of at all, or which we regard as completely random behaviour."[956]

> „Ein wichtiges Merkmal einer lernenden Maschine ist, dass ihr Lehrer oft sehr weitgehend unwissend ist, was in ihrem Inneren vor sich geht, auch wenn er das Verhalten seines Schülers bis zu einem gewissen Grad noch vorhersagen kann. [...] Dies steht in klarem Gegensatz zum normalen Vorgehen bei der Durchführung von Berechnungen mit einer Maschine: Das Ziel besteht dann darin, ein klares geistiges Bild vom Zustand der Maschine zu jedem Zeitpunkt der Berechnung zu haben [...] Die meisten Programme, die wir in die Maschine eingeben werden können, werden dazu führen, dass sie etwas tut, das wir überhaupt nicht verstehen können oder das wir als völlig zufälliges Verhalten betrachten."

„It may be seen that the result of a programming technique of automatization is to remove from the mind of the designer and operator an effective understanding of many of the stages by which the machine comes to its conclusions and of what the real tactical intentions of many of its operation may be."[957]

> „Das Ergebnis einer Programmiertechnik der Automatisierung mag sein, dass der Programmierer und der Betreiber einer solchen Maschine kein tatsächliches Verständnis mehr haben werden von den vielen Schritten, auf Grund derer die Maschine zu ihren Schlussfolgerungen kommt, und von den tatsächlichen taktischen Zielen vieler ihrer Operationen."

„It may well be that in principle we cannot make any machine the elements of whose behavior we cannot comprehend sooner or later. This does not mean in any way that we shall be able to comprehend these elements in substantially less time than the time required for operation of the machine, or even within any given number of years or generations."[958]

> „Es mag sein, dass wir keine Maschine erschaffen können, deren Verhaltenselemente wir nicht früher oder später verstehen. Dies bedeutet jedenfalls nicht, dass wir diese Elemente in erheblich kürzerer Zeit verstehen als die Zeit, die erforderlich für den Betrieb der Maschine ist, oder, dass wir sie innerhalb einer bestimmten Anzahl an Jahren oder Generationen verstehen."

Turing legt dar, dass für ein erfolgreiches Bestehen des von ihm konzipierten „imitation game" nicht nur die weitere gerätetechnische Entwicklung, sondern v.a. die richtige Programmierung und die Schaffung von „learning machines" entscheidend sei. Als Konsequenz der Entwicklung von „learning machines" sah er voraus, dass der Lehrer, d.h. der Entwickler, nicht mehr wissen werde, was im Inneren der Maschine vorgehe und deren Ergebnisse nicht verstehen könne.

Nach *Wiener* führen Programmiertechniken der Automatisierung[959] dazu, dass Entwickler und Betreiber die Entscheidungsbildung und die dazugehörigen einzelnen Schritte nicht mehr verstehen können. Er brachte zudem das Zeitlimit bei

956 *Turing* Mind 59 (1950), 433 (458 f.).
957 *Wiener* Science 131 (1960), 1355 (1357).
958 *Wiener* Science 131 (1960), 1355.
959 Insofern *Wiener* von Automatisierungstechniken spricht, meint er autonome und lernfähige Maschinen, vgl. *Wiener* Science 131 (1960), 1355.

der Entwicklung eines solchen Verständnisses ins Spiel: Was nicht in angemessener Zeit verstanden werden kann, könne eben gar nicht verstanden werden.[960]

Wie ist nun der Stand 60 bis 70 Jahre später, also zu einem Zeitpunkt, zu dem „learning machines" tatsächlich umgesetzt sind? Die Europäische Kommission beschreibt KI-Systeme wie folgt:

> „Opacity: The more complex emerging digital technologies become, the less those taking advantage of their functions or being exposed to them can comprehend the processes that may have caused harm to themselves or to others. Algorithms often no longer come as more or less easily readable code, but as a black-box that has evolved through self-learning and which we may be able to test as to its effects, but not so much to understand."[961]
>
> „Opazität: Je komplexer aufkommende digitale Technologien werden, desto weniger können diejenigen, die ihre Funktionen ausnutzen oder ihnen ausgesetzt sind, die Prozesse nachvollziehen, die ihnen selbst oder anderen Schaden zugefügt haben könnten. Algorithmen zeigen sich oft nicht mehr als mehr oder weniger leicht lesbarer Code, sondern als Black-Box, die sich durch Selbstlernen entwickelt hat und deren Auswirkungen wir vielleicht testen können, aber nicht so sehr verstehen."

Auch die Europäische Kommission spricht also bei der Beschreibung von KI-Systemen davon, dass die damit verbundenen Prozesse nicht verstehbar und nachvollziehbar seien.

Fehlende Transparenz, die Opazität von KI bzw. deren Black-Box-Effekt, werden derzeit in der Tat als prägendes Element im Verhältnis KI und Entwickler bzw. Nutzer wahrgenommen;[962] es hat sich sogar ein Wissenschaftszweig der eXplainable Artificial Intelligence (XAI) entwickelt, der sich im Detail mit der Opazität von KI und deren Ursachen beschäftigt und Wege sucht, diese besser nachvollziehbar zu machen.

Der folgende Abschnitt beleuchtet diese Thematik näher. In einem ersten Schritt (1. Epistemische Opazität) geht es um den Begriff der Opazität, dessen Herkunft und Bedeutung im Zusammenhang mit computertechnischen Systemen im Allgemeinen und mit KI im Konkreten. In einem zweiten Schritt (2. Dimensionen epistemischer Transparenz) wird dagegen der Begriff der Transparenz beleuchtet, also die Anforderungen, damit ein KI-System gerade nicht opak ist. Zuletzt geht es speziell um die Ursachen von Opazität bei ML-Modellen (3. Epistemische Opazität bei KI-Systemen). In einem letzten Schritt (4. Exkurs: Maschinelles Lernen und „statistische Transparenz") wird zuletzt der epistemischen Opazität die „statistische Transparenz" von KI-Systemen gegenübergestellt.

960 So die Zusammenfassung der Aussagen von *Wiener* durch *Lenhard* 2015, 104.

961 *Expert group on Liability and New Technologies* 2019, 33.

962 Vgl. nur *Expert group on Liability and New Technologies* 2019, 33; *Yuan* RW 9 (2018), 477 (490); *Zech* 2020, A 41; *Datenethikkommission* 2019, 169; *Fateh-Moghadam* ZStW 131 (2019), 863 (885).

1. Epistemische Opazität

Der Begriff der Opazität im Zusammenhang mit Computersystemen im Allgemeinen wurde – soweit ersichtlich – erstmals vom US-amerikanischen Philosophen *Humphreys* verwendet. *Humphreys* definiert Opazität, genauer gesagt epistemische Opazität (epistemic opacity), wie folgt:

> „[…] a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process."[963]
>
> „[…] ein Prozess ist für einen kognitiven Agenten X zum Zeitpunkt t nur dann epistemisch opak, wenn X zum Zeitpunkt t nicht alle epistemisch relevanten Elemente des Prozesses kennt. Ein Prozess ist für X prinzipiell epistemisch opak dann und nur dann, wenn es angesichts der Natur von X für X unmöglich ist, alle epistemisch relevanten Elemente des Prozesses zu kennen."

Epistemische Opazität ist also mit *Humphreys* zunächst ein Reflexionsbegriff, da er das *Verhältnis* eines Subjekts zu einem Prozess beschreibt.[964] Das Subjekt kann jedermann sein, aber auch und v.a. der Experte.[965] Für dieses Subjekt sind nun bestimmte Elemente des Prozesses nicht oder nicht zu einem bestimmten Zeitpunkt epistemisch einsichtig. Die Uneinsichtigkeit bezieht sich nicht auf irgendeinen Prozess, sondern auf den Prozess des wissenschaftlichen Vorgehens; epistemische Opazität meint also methodische Opazität.[966] Wie *Wiener* operiert dabei auch *Humphreys* mit einem Zeitlimit, indem er relative Opazität als Uneinsichtigkeit für ein Subjekt zu einem bestimmten Zeitpunkt und absolute Opazität als Uneinsichtigkeit unabhängig vom Zeitfaktor unterscheidet.

Epistemische Opazität beschreibt zusammengefasst die absolute bzw. zeitlich relative Uneinsichtigkeit der wissenschaftlichen Methode eines Prozesses selbst für Experten. (Epistemische) Transparenz liegt hingegen vor, wenn die wissenschaftliche Methode einsichtig ist; oben wurde dies in Bezug auf „klassische" Technik auch als epistemische Verstehbarkeit beschrieben.[967]

Humphreys wendete den Begriff der epistemischen Opazität auf Computersimulationen etwa in Form von Klimasimulationen[968] an. Er stellte dabei fest:

> „The computations involved in most simulations are so fast and so complex that no human or group of humans can in practice reproduce or understand the processes."[969]

963 *Humphreys* Synthese 169 (2009), 615 (618).
964 *Kaminski* 2018, 321 mit Fn. 8.
965 Vgl. *Kaminski* 2020, 161.
966 *Kaminski/Resch/Küster* 2018, 258.
967 Siehe oben Kap. 1, B.II.3.
968 Vgl. dazu erläuternd *Kaminski/Resch/Küster* 2018, 254 f ff.
969 *Humphreys* Synthese 169 (2009), 615 (619).

> „Die Berechnungen in den meisten Simulationen sind so schnell und so komplex, dass kein Mensch oder eine Gruppe von Menschen die Prozesse in der Praxis reproduzieren oder verstehen kann."

Der Begriff und das Konzept der (epistemischen) Opazität wurden – in Anlehnung an *Humphreys* – auch auf KI-Systeme übertragen: Die technisch autonom erfolgenden Regelbildungen seien so komplex und abstrakt, dass sie für Menschen und gerade auch für Experten zum derzeitigen Stand der Wissenschaft nicht reproduzierbar oder verstehbar seien.[970] Sie seien damit epistemisch opak bzw. nicht transparent. Oder in anderen Worten: Eingabe- und Ausgabewerte, Input und Output von KI-Systemen sind sichtbar, aber die Transformationsfunktion, die die beiden Werte miteinander verbindet, verbirgt sich in der Black-Box, die selbst für Experten nicht mehr zu öffnen ist.[971]

2. *Dimensionen epistemischer Transparenz*

Der Wissenschaftszweig der eXplainable Artificial Intelligence (XAI) beschäftigt sich – wie erwähnt – mit der Opazität von KI-Systemen und sucht Wege, diese besser nachvollziehbar zu machen. Dafür definiert die XAI zuallererst die Anforderungen, die an transparente KI-Systeme zu stellen sind. Dabei zeigt sich, dass es verschiedene Dimensionen epistemischer Transparenz geben kann:

Transparenz kann Simulierbarkeit bedeuten, welche gegeben ist, wenn ein Mensch auf Grundlage der Eingabewerte und der Parameter alle Rechenschritte eines ML-Modells in angemessener Zeit auswerten kann.[972] Ein System ist also simulierbar und damit transparent, wenn ein Mensch nachrechnen kann, was das Modell errechnet hat. Simulierbarkeit ist im Gegenteil nicht gegeben, wenn ein solches Nachrechnen in angemessener Zeit nicht möglich ist; es geht damit um die von *Humphreys* angesprochene – zeitlich gesehene – relative epistemische Opazität, also die Uneinsichtigkeit der wissenschaftlichen Methode zu einem bestimmten Zeitraum.

Verständlichkeit / Unterteilbarkeit bedeutet, dass alle Komponenten eines ML-Modells, Eingangswerte, Parameter und Rechenschritte, intuitiv verständlich sein müssen.[973] Was darunter genau zu verstehen ist, beschreibt der Physiker *Feynman* (nicht nur im Hinblick auf KI-Systeme) wie folgt:

> „I understand what an equation means if I have a way of figuring out the characteristics of its solution without actually solving it. So if we have a way of knowing what should

970 *Kaminski* 2020, 161. Vgl. auch *Expert group on Liability and New Technologies* 2019, 33.
971 *Kaminski* 2020, 158, 162.
972 *Lipton* 2017; *Schaaf* 2020.
973 *Schaaf* 2020; *Lipton* 2017.

> happen in given circumstances without actually solving the equations, then we ‚understand' the equations, as applied to these circumstances."[974]
>
> „Ich verstehe, was eine Gleichung bedeutet, wenn ich einen Weg habe, die Eigenschaften ihrer Lösung herauszufinden, ohne sie tatsächlich zu lösen. Wenn wir also eine Möglichkeit haben, zu wissen, was unter bestimmten Umständen passieren sollte, ohne die Gleichungen tatsächlich zu lösen, dann ‚verstehen' wir, wie die Gleichungen auf diese Umstände angewendet werden."

Verständlichkeit in diesem Sinne ist also zu unterscheiden von der Möglichkeit nachzurechnen. Es geht darum, zu erkennen, wie die Elemente eines Modells zueinander im Verhältnis stehen, um so die methodische Grundkonzeption eines Modells zu verstehen. Der Output kann dann vorhergesehen werden, ohne das Modell nachzurechnen.

Algorithmische Transparenz betrifft den Lernalgorithmus selbst. Diese Form der Transparenz ist also gegeben, wenn die Regel, die Funktion, die aus einem bestimmten Input einen bestimmten Output produziert, und entsprechend jegliche Reaktionen des Modells verstanden werden können.[975]

Diese Formen der Transparenz sind *systemimmanent* – ein Modell ist entweder transparent in einem oben beschriebenen Sinne oder eben opak.[976] Sog. post-hoc-Erklärbarkeit (post-hoc-explainability) von KI-Systemen hingegen beruht auf Bemühungen, trotz einer systemimmanenten Opazität Ergebnisse des KI-Systems durch zusätzliche Module oder Systeme erklärbar zu machen – etwa durch schriftliche, visuelle oder beispielhafte Erklärungen.[977]

3. *Epistemische Opazität bei KI-Systemen*

Epistemische Opazität von KI-Systemen kann sich aus zwei Quellen speisen: Aus ihrer *Komplexität* (a.) und ihrer tatsächlichen *analytischen Unverständlichkeit* (b.):

a. (relative) epistemische Opazität auf Grund Komplexität

Die Komplexität betrifft eine grundsätzliche Problematik, die bereits *Wiener* angesprochen hatte, nämlich die, dass „Mensch und Maschine" auf unterschiedlichen Zeitskalen (time scales) operieren:

974 *Feynman/Leighton/Sands* 2010, 2-1.
975 *Barredo Arrieta/Díaz-Rodríguez/Del Ser u. a.* IF 58 (2020), 82 (88); *Schaaf* 2020; *Lipton* 2017; *Lipton* 2017.
976 Vgl. *Sheh/Monteath* KI 32 (2018), 261 (264).
977 *Barredo Arrieta/Díaz-Rodríguez/Del Ser u. a.* IF 58 (2020), 82 (89); *Lipton* 2017; *Sheh/Monteath* KI 32 (2018), 261 (263).

> „[…] man and machine operate on two distinct time scales, so that the machine is much faster than man […]“[978]
>
> „[…] Mensch und Maschine arbeiten auf zwei unterschiedlichen Zeitskalen, so dass die Maschine viel schneller ist als der Mensch […].“

Der Zeitaufwand, den ein Mensch benötigt, um Berechnungen eines KI-Systems *nachzurechnen*, ist um ein Vielfaches höher, als der Zeitaufwand für die ursprünglichen Berechnungen durch das System. Ergebnisse eines KI-Systems können also schon deswegen – relativ – epistemisch opak sein, weil ein Mensch die einzelnen Rechenschritte unter Zugrundelegung der Eingabewerte und der Modellparameter nicht in angemessener Zeit nachrechnen kann. Denn KI-Systeme weisen oft eine enorme Größe auf, erfordern eine Vielzahl einzelner Rechenschritte[979] und sind somit angesichts der eingeschränkten Rechenkapazität des Menschen komplex.

Vgl. dazu folgendes Beispiel eines künstlichen neuronalen Netzes zur Handschriftenerkennung nach *Kaminski*:[980]

> „Die erste Schicht enthält z.B. 784 Neuronen, welche einem Gitter mit der Auflösung von 28 × 28 Pixeln entspricht. Dieses Gitter wird unter die handschriftlich notierte Ziffer gelegt, sodass für jede Box der jeweilige Licht- und Schattenwert bestimmt werden kann. Jedes der 784 Neuronen weist daher einen Wert zwischen 0 und 1 auf; dieser Wert entspricht dem Farbwert (von 0 für weiß bis 1 für schwarz). Die letzte Schicht weist 10 Neuronen auf, denen die zu erkennenden Zahlenwerte (0–9) entsprechen. Diese letzte Schicht kann ebenfalls wieder Werte zwischen 0 und 1 aufweisen; dem entspricht die Wahrscheinlichkeit, mit der der Zahlenwert der handschriftlichen Ziffer (von 0–9) korrekt erkannt wurde. Dazwischen finden sich, in diesem simplen Beispiel, zwei so genannte „hidden layers“ mit jeweils 16 Neuronen. Jedes Neuron der einen Schicht ist mit allen Neuronen der nächsten Schicht verbunden. Zudem werden so genannte Gewichte und Biaswerte eingeführt. In diesem Netz ergeben sich so 13.002 Parameter.“

Allein die enorme Anzahl der Parameter in diesem simplen(!) künstlichen neuronalen Netz zu erfassen, übersteigt die menschliche Rechenkapazität und zeigt, dass künstliche neuronale Netze zu komplex sind, als dass sie menschlich nachgerechnet werden könnten und damit simulierbar wären.[981]

978 *Wiener* Science 131 (1960), 1355 (1358).
979 Vgl. *Lipton* 2017.
980 *Kaminski* 2020, 160.
981 *Schaaf* 2020.

b. (absolute) epistemische Opazität auf Grund analytischer Unverständlichkeit

Zum Aspekt der absoluten epistemischen Opazität auf Grund analytischer Unverständlichkeit sei noch einmal das wittgenstein`sche Eingangszitat zu diesem Abschnitt wiederholt:

> „Wenn ein Löwe sprechen könnte, wir könnten ihn nicht verstehen."[982]

Selbstverständlich „spricht" die KI nicht zu uns, aber ein KI-System verfügt über eine Memory-Einheit, in der Eingangswerte, Parameter und Ausgangswerte gespeichert sind, und diese können auch eingesehen werden.[983] Also könnte man annehmen, dass doch jedenfalls eine Nachvollziehbarkeit des KI-Systems bzw. eine gewisse algorithmische Transparenz möglich sein könnte?

Einerseits wäre ein solches Unterfangen zeitintensiv – Stichwort Komplexität –, andererseits wäre es auch nicht zielführend.[984] Denn die gespeicherten Eingangswerte, Parameter und Ausgangswerte sind nicht explizit dargestellt, also in einer Form gespeichert, die für Menschen verständlich ist.[985] Die in einem künstlichen neuronalen Netz enthaltenen Gewichtungen zwischen den einzelnen Neuronen werden bspw. in Werten zwischen 0 und 1 angegeben, dies lässt aber nur eine sehr abstrakte Interpretation zu,[986] da nicht klar ist, welche symbolische Wertung dahinter steht.

Das KI-System „spricht" zwar über Zahlen und Gleichungen zu uns, diese „Sprache" ist aber für Menschen nicht verständlich, weil damit keine quasi-linguistische Bedeutung verbunden ist:

> „[…] it is possible to simply print out all of the neural network weights and activations. For all but the most trivial problems, these „Execution" versions of introspective explanations are neither understandable nor deep enough to be useful." [987]
>
> „[…] es ist möglich, einfach alle Gewichte und Aktivierungen des neuronalen Netzes auszudrucken. Für alle außer den trivialsten Problemen sind diese "Ausführungs"-Versionen der introspektiven Erklärungen weder verständlich noch tief genug, um nützlich zu sein."
>
> „[…] there [is] nowhere […] a list or catalog of all learned information, as there is in symbolic programs. […] Connectionist systems lack an explicit representation and the contained information can only be deduced from their behavior. […] We cannot have a look at the information that is stored inside the network."[988]
>
> „[...] es [gibt] nirgendwo [...] eine Liste oder einen Katalog aller gelernten Informationen, wie es sie in symbolischen Programmen gibt. [...] Konnektionistischen Systemen fehlt eine explizite Repräsentation und die enthaltenen Informationen können

982 *Wittgenstein* 1984, 569.
983 *Sheh/Monteath* KI 32 (2018), 261 (263).
984 *Sheh/Monteath* KI 32 (2018), 261 (263).
985 *Matthias* Ethics and Information Technology 6 (2004), 175 (178).
986 *Kaminski* 2020, 160.
987 *Sheh/Monteath* KI 32 (2018), 261 (263).
988 *Matthias* Ethics and Information Technology 6 (2004), 175 (178).

> nur aus ihrem Verhalten abgeleitet werden. [...] Wir können keinen Blick auf die Informationen werfen, die im Netzwerk gespeichert sind."

Insbes. künstliche neuronale Netze sind auch deswegen epistemisch opak, weil sie analytisch unverständlich sind. Das Modell, seine Parameter und der Lernalgorithmus sind in einer Weise dargestellt, dass deren Bedeutung für Menschen nicht verständlich und damit nicht analysierbar ist.

Die epistemische Opazität von KI-Systemen wurde anhand von künstlichen neuronalen Netzen als Lernmethode exemplifiziert. Es bleibt jedoch zu betonen, dass nicht jedes KI-System in gleichem Maße epistemisch opak ist wie ein künstliches neuronales Netz. Denn es existieren Lernmethoden, die einerseits weniger komplex sein können – und damit Transparenz im Sinne einer Simulierbarkeit ermöglichen – andererseits analytisch verständlich sind, da die Modell-Informationen in symbolischer, expliziter und damit für Menschen nachvollziehbarer Weise dargestellt sind.[989] Dies gilt insbes. für sog. Entscheidungsbäume, die auf symbolischen Lernmethoden beruhen; auch diese können zwar komplex werden, sie bleiben aber in der Regel analytisch verständlich.[990]

4. Exkurs: Maschinelles Lernen und „statistische Transparenz"

Epistemische Opazität beschreibt die Intransparenz des technischen Transformationsprozesses mit der Folge, dass dieser Prozess nicht erklärt werden kann und somit das technische Können weitergeht als das technische Verstehen. Dass ein KI-System nicht in diesem Sinne nachvollziehbar gemacht und sein „Agieren" nicht auf diese Weise begründet werden kann, heißt aber nicht, dass eine Begründung bzw. Bewertung von KI-Systemen nicht möglich wäre bzw. nicht erfolgen würde. KI-Systeme werden vielmehr stets daraufhin evaluiert, wie gut ihre Vorhersagekraft ist,[991] in anderen Worten, wie gut die von ihnen gebildeten Korrelationen mit tatsächlichen Kausalitäten übereinstimmen.

Yuan etwa beschreibt, dass bspw. im Falle von Klassifikationsaufgaben die Anzahl der wahr positiven, falsch positiven, wahr negativen und falsch negativen Vorhersagen in bestimmte Verhältnisse gesetzt und Zahlenwerte gebildet werden, die zwischen 0 und 1 liegen. Je mehr sich diese Zahlenwerte 1 annähern, desto besser ist die Vorhersage.[992]

Wenn KI-Systeme also auch nicht epistemisch transparent sein mögen, so sind sie jedenfalls „statistisch" transparent, weil – vereinfacht gesprochen – bekannt

989 Vgl. zu einer Übersicht *Barredo Arrieta/Díaz-Rodríguez/Del Ser u. a.* IF 58 (2020), 82 (94 ff.) Daneben *Sheh/Monteath* KI 32 (2018), 261 (265).
990 Vgl. nur *Ertel* 2021, 231 f.; *Borges/Grabmair/Krupka u. a.* 2018, 52 f. (mit Hinweis auch auf die logistische Regression). Siehe aber *Sheh/Monteath* KI 32 (2018), 261 (265).
991 *Yuan* RW 9 (2018), 477 (491).
992 *Yuan* RW 9 (2018), 477 (491).

ist, mit welcher Trefferwahrscheinlichkeit ein System operiert. Es ist ein „probabilistisches Urteil" möglich, das eine Aussage über die Verlässlichkeit eines ML-Modells trifft.[993]

IV. Can machines think? – schwache vs. starke Künstliche Intelligenz

„Can machines think?" – mit dieser Frage leitete *Turing* seinen Beitrag zum „imitation game" ein. Er ersetzte diese Frage zunächst mit einer „more accurate form of this question"[994], indem er danach fragte, ob eine „machine" – ein Computer – sich im oben beschriebenen „imitation game" bewähren könne. Er sah jedoch, dass seine ursprüngliche Frage nach der denkenden Maschine weiterhin relevant sein werde, und zwar jedenfalls, sobald eine Maschine sich im „imitation game" bewähre – *Turing* ging davon aus, dass dies am Ende des 20. Jahrhunderts der Fall sein werde. Dann werde sich der allgemeine Sprachgebrauch und die allgemeine Meinung so geändert haben, dass man von denkenden Maschinen sprechen könne, ohne damit rechnen zu müssen, auf Widerspruch zu stoßen.[995]

Wir haben gesehen, dass KI-Systeme noch nicht so weit entwickelt sind, dass sie sich im „imitation game" bewähren können, und dass *Turing* mit seiner Einschätzung, wie schnell die Entwicklung solcher Systeme vonstattengehen würde, falsch lag. Dennoch weisen KI-Systeme mit ihrer technischen Autonomie und Lernfähigkeit im oben beschriebenen Sinne Eigenschaften auf, die im wissenschaftlichen Diskurs – jenseits einer rein anwendungsbezogenen Erforschung der KI – die Frage aufkommen lassen, ob sie dadurch zu „denkenden Einheiten" werden. Und so liegt auch dieser Arbeit die Hypothese zu Grunde, dass mit KI ein autonomer und in seiner Funktionsweise opaker *Akteur* dem Menschen als „zweite Natur" gegenübertritt. Daran schließen insbes. in der Moralphilosophie sowie in der (Straf-)Rechtswissenschaft Verantwortungsdiskussionen an, die sich auch damit beschäftigen, ob KI-Systeme als weitere Verantwortungssubjekte anzuerkennen sind.[996]

An dieser Stelle soll zunächst die Frage des „Can machines think?" aufgenommen werden und ein Überblick über den hierzu geführten Diskurs gegeben werden. Die Meinungspole dazu werden mit dem Begriffspaar der *schwachen* und *starken* KI bezeichnet, so dass diese Begrifflichkeiten zunächst einmal erläutert werden (1.). In einem zweiten Schritt (2.) wird der Diskussionsstand zur Frage nach der (a.) Möglichkeit und der (b.) technischen Umsetzung starker KI wieder-

993 *Kaminski/Resch/Küster* 2018, 258, 266, 271 im Hinblick auf Computersimulationen, die aber wie Verfahren des maschinellen Lernens epistemische Opazität aufweisen können.
994 Mit einer „genaueren Form dieser Frage".
995 *Turing* Mind 59 (1950), 433 (442); zu Turings „imitation game" sowie zu seinem Verständnis einer „machine" siehe oben Kap. 2, A.I.1.a. bzw. Kap. 1, B.II.1.a.
996 Siehe dazu unten Kap. 3, C.

gegeben. In einem dritten Schritt (3.) wird schließlich KI auf der Grundlage ihrer derzeitigen technischen Umsetzung in der Form des maschinellen Lernens der schwachen KI zugeordnet. Gerade der letzte Punkt wird noch einmal die Bedeutung und die (begrenzte) Leistungsfähigkeit derzeitiger KI-Systeme aufzeigen.

1. Schwache und starke KI

Die Begriffe der schwachen und starken KI führte der US-amerikanische Philosoph *Searle* für die Bedeutung von Computersimulationen bei der Gehirnforschung ein:[997]

> „ […] I find it useful to distinguish what I will call "strong" AI from "weak" or "cautious" AI […]. According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. […] But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states.“
>
> „[…] ich finde es nützlich, das, was ich "starke" KI nennen werde, von "schwacher" oder "vorsichtiger" KI […] zu unterscheiden. Gemäß der schwachen KI besteht der hauptsächliche Wert des Computers für die Erforschung des Gehirns darin, dass er uns ein sehr mächtiges Werkzeug an die Hand gibt. Aber gemäß der starken KI ist der Computer nicht nur ein Hilfsmittel bei der Erforschung des Geistes, sondern der richtig programmierte Computer verfügt tatsächlich über ein Gehirn, in dem Sinne, dass Computer, die die richtigen Programme erhalten, buchstäblich verstehen können und sonstige kognitive Zustände aufweisen.“

Nach *Searle* ist schwache KI ein Mittel zum Zweck, um neue Erkenntnisse in der Hirnforschung zu erlangen. Starke KI hingegen beschreibe Computer, die tatsächlich verstehen und sonstige kognitive Zustände aufweisen könnten.

Die Bedeutung von schwacher und starker KI allgemein fassen *Russel* und *Norvig* wie folgt zusammen:

> „[…] the assertion that machines could act as if they were intelligent is called the weak AI hypothesis […], and the assertion that machines that do so are actually thinking (not just simulating thinking) is called the strong AI hypothesis.“[998]
>
> „[...] die Behauptung, dass Maschinen so tun könnten, als wären sie intelligent, wird als schwache KI-Hypothese bezeichnet [...], und die Behauptung, dass Maschinen, die dies tun, tatsächlich denken (und nicht nur das Denken simulieren), wird als starke KI-Hypothese bezeichnet."

Nimmt man *Searle* sowie *Russel und Norvig* zusammen, bedeutet schwache KI also, dass intelligente Systeme ein Mittel zum Zweck sind, indem sie menschliche

997 *Searle* BBS 1980, 417 (417).
998 *Russell/Norvig* 2010, 1020.

Intelligenz simulieren, starke KI dagegen, dass diese Systeme tatsächlich Intelligenz aufweisen und denken wie Menschen.[999]

2. *Möglichkeit und technische Umsetzung starker KI?*

Bei der Diskussion um schwache vs. starke KI sind zwei Fragen auseinanderzuhalten: (a.) Ist eine starke KI prinzipiell möglich oder kann KI immer nur „schwach" sein, weil menschliche Intelligenz allenfalls simuliert werden kann? (b.) Sofern man von der prinzipiellen Möglichkeit starker KI ausgeht, ist diese derzeit bereits technisch umgesetzt?

a. *Möglichkeit starker KI*

Ein Hauptvertreter der Möglichkeit einer starken KI ist *Haugeland*. Er beschreibt das Ziel von KI im Sinne einer starken KI wie folgt:

> „The fundamental goal of this research is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: *machines with minds,* in the full and literal sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, *computers ourselves*. That idea – the idea that thinking and computing are radically the same – is the topic of this book."[1000]
>
> „Das grundlegende Ziel dieser Forschung ist es nicht, Intelligenz zu imitieren oder eine clevere Fälschung zu produzieren. Ganz und gar nicht. KI will nur das Echte: Maschinen mit Verstand, im vollen und wörtlichen Sinne. Das ist keine Science-Fiction, sondern echte Wissenschaft, die auf einer ebenso tiefgründigen wie gewagten theoretischen Vorstellung beruht: Nämlich, dass wir im Grunde genommen selbst Computer sind. Diese Idee – die Idee, dass Denken und Rechnen radikal das Gleiche sind – ist das Thema dieses Buches."

Haugeland bejaht also die Möglichkeit einer starken KI, weil er der Überzeugung ist, dass Denken und Rechnen das Gleiche ist; eine Maschine könne denken, weil das menschliche Gehirn selbst nichts anderes als eine denkende Maschine sei. Dieser Ansatz, der als „computational theory of the mind" bekannt ist, ist die wesentliche Grundlage derzeitiger Forschung zur starken KI.[1001] Daneben werden eine Vielzahl von Gründen für und auch wider die Möglichkeit einer starken KI angebracht und dabei grundlegende – und in vielerlei Hinsicht ungelöste – Fragen aufgeworfen: Was ist menschliche Intelligenz? Was ist Denken oder was bedeutet Bewusstsein?[1002] Es zeigen sich Ähnlichkeiten zu dem Determinismus/In-

999 So auch das Verständnis von *Arkoudas/Bringsjord* 2014, 35 f.
1000 *Haugeland* 1985, 2.
1001 *Arkoudas/Bringsjord* 2014, 36; *Rescorla* 2020.
1002 Zu einer Übersicht klassisch *Turing* Mind 59 (1950), 433 (442 ff.) Daneben *Arkoudas/Bringsjord* 2014; *Rescorla* 2020; *Russell/Norvig* 2021, 1035 ff. Anschaulich *Nida-Rümelin/Weidenfeld* 2018, 32 ff.

determinismus-Streit über die Existenz menschlicher Willensfreiheit, der in der Strafrechtswissenschaft im Bereich der Schuld geführt wird.[1003]

Eines der prominentesten Argumente gegen die Annahme, dass Computer denken können wie Menschen, ist ein Gedankenexperiment[1004] von *Searle*, das als „Chinese Room Argument" bekannt ist. *Searle* stellt sich dafür folgende Konstellation vor:

> „Imagine a native English speaker […] who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese."
>
> „Stellen Sie sich vor, dass ein englischer Muttersprachler […], der kein Chinesisch kann, in einen Raum voller Kisten mit chinesischen Symbolen (eine Datenbank) und einem Buch mit Anweisungen zur Verarbeitung der Symbole (das Programm) gesperrt wird. Stellen Sie sich vor, dass Leute außerhalb des Raums andere chinesische Symbole einsenden, die, ohne dass die Person im Raum es weiß, Fragen auf Chinesisch sind (die Eingabe). Und stellen Sie sich vor, dass der Mensch im Raum durch Befolgen der Anweisungen im Programm in der Lage ist, chinesische Symbole auszugeben, die richtige Antworten auf die Fragen sind (die Ausgabe). Das Programm ermöglicht der Person im Raum, den Turing-Test für das Verstehen von Chinesisch zu bestehen, aber sie versteht kein Wort Chinesisch."

Aus seinem Gedankenexperiment zieht *Searle* folgende Schlussfolgerung:

> „The point of the argument is this: if the man in the room does not understand Chinese on the basis of implementing the appropriate program for understanding Chinese then neither does any other digital computer solely on that basis because no computer, qua computer, has anything the man does not have."[1005]
>
> „Die Kernaussage des Arguments ist folgende: Wenn der Mann im Raum kein Chinesisch versteht, indem er das entsprechende Programm zum Verstehen von Chinesisch umsetzt, dann tut es auch kein anderer digitaler Computer allein auf dieser Basis, weil kein Computer, qua Computer, etwas hat, was der Mensch nicht hat."

Searle argumentiert also, dass zwar ein Computer programmiert werden könne, der Sprache *scheinbar* versteht, weil korrekte Antworten (als Output) gegeben würden, dass aber nie ein wirkliches Verständnis erzeugt werden könne. Denn Computer würden lediglich syntaktische Regeln verwenden, aber kein Verständnis von Bedeutung oder Semantik haben. *Searle* setzt sich damit in Widerspruch zu *Haugelands* These, dass Denken und Rechnen dasselbe seien: Denn, wenn der Computer Sprache nicht verstehen kann, kann er auch nicht denken. Denken be-

1003 Siehe dazu unten Kap. 3, C.I.2.
1004 Eingehend hierzu *Cole* 2020; daneben *Danks* 2014, 160; *Stephan/Walter* in: *Turing* 2021, 187 f.
1005 Für dieses und das vorangegangene Zitat: *Searle* 1999, 115.

schränkt sich demnach nicht auf die Fähigkeit syntaktische Regeln anzuwenden, sondern erfasst auch die Fähigkeit eines semantischen Verstehens.

Mit seinem Gedankenexperiment will *Searle* zudem zeigen, dass ein Computer, der nicht verstehen kann, auch kein Bewusstsein oder finales Handeln aufweisen kann.

> „I demonstrated years ago with the so-called Chinese Room Argument that the implementation of the computer program is *not* by itself *sufficient for consciousness or intentionality* [...]. Computation is defined purely formally or syntactically, whereas minds have actual mental or semantic contents, and we cannot get from syntactical to the semantic just by having the syntactical operations and nothing else."[1006]
>
> „Ich habe vor Jahren mit dem sogenannten Chinese Room Argument gezeigt, dass die Ausführung des Computerprogramms an sich *nicht für Bewusstsein oder Intentionalität ausreicht* [...]. Computation ist rein formal oder syntaktisch definiert, wohingegen Verstand tatsächliche mentale oder semantische Inhalte hat, und wir können nicht vom Syntaktischen zum Semantischen kommen, indem wir nur die syntaktischen Operationen haben und sonst nichts."
>
> [Hervorhebungen durch die Verf.]

Bewusstsein und finales Handeln setzen nach *Searle* damit die Fähigkeit voraus, nicht nur Regeln anwenden zu können, sondern auch die damit verbundene Bedeutung zu verstehen.

Die Frage, ob starke KI möglich ist, ist also umstritten und nicht geklärt. Die Möglichkeit einer starken KI bejahen v.a. diejenigen, die keinen kategorialen Unterschied zwischen den Rechenprozessen eines Computers und den kognitiven Vorgängen im Gehirn eines Menschen erkennen. Es bestehe die prinzipielle Möglichkeit einer denkenden Maschine, weil das menschliche Gehirn nichts Anderes sei. Der Unterschied zwischen einem „einfachen" Computer, einem KI-System und einem denkenden Menschen ist danach kein kategorialer, sondern ein relativer – der Mensch in seiner „Rechenkapazität" ist derzeit noch überlegen. Diejenigen, die die Möglichkeit einer starken KI ablehnen, sehen dagegen einen kategorialen Unterschied zwischen den kognitiven Vorgängen im Gehirn und den Rechenvorgängen von KI-Systemen. So *Searle*: Computer würden lediglich syntaktische Operationen vornehmen, ohne dass dadurch eine Vorstellung von Semantik, von Bedeutung und Verständnis, verbunden sei.

b. Technische Umsetzung starker KI

Die Frage nach der prinzipiellen Möglichkeit einer starken KI ist die eine, die nach deren technischer Umsetzung die andere.

Auch Vertreter einer starken KI-These gehen beim aktuellen Stand der KI-Entwicklung noch nicht von der Existenz denkender Maschinen aus. Sie halten

1006 *Searle* 2010, 17.

starke KI zwar für prinzipiell möglich, technisch aber noch nicht umsetzbar bzw. umgesetzt, etwa weil bislang kein Computer in der Lage war, den Turing-Test oder den „Total Turing Test" zu bestehen.[1007] KI-Systeme, die derzeit von einer anwendungsbezogenen Herangehensweise an KI geschaffen werden können, haben also noch nicht die Fähigkeiten entwickelt, die nach den Vertretern einer starken KI-These erforderlich wären, um von einer starken KI zu sprechen. Dies sei etwa gegeben im Falle einer „Artificial General Intelligence"[1008] bzw. einer „technologischen Singularität", wenn also ein KI-System jede ihm gestellte Aufgabe erlernen könnte. Derzeitige KI-Systeme sind in ihrer Anwendung beschränkt – der Algorithmus, der ein autonomes Fahrzeug „steuert", kann kein Go spielen oder Proteinstrukturen vorhersagen[1009] – man spricht in diesem Zusammenhang daher auch von „narrow ai".[1010] Die Möglichkeiten für eine technologische Singularität werden zwar erforscht, auf dem Weg dorthin ergeben sich jedoch viele Hürden; KI-Systeme müssten etwa mit weniger Daten und Beispielen trainiert werden können – um bspw. wie ein Kind nach wenigen Beispielen zu wissen, wie eine Katze aussieht –, ihre Labels selber lernen, auf neue Kontexte umlernen können (ohne die anderen Kontexte zu „vergessen" – sog. „catastrophic forgetting") etc.[1011] Erste Ansätze jedenfalls zu einem Transferlernen weisen sog. Foundation Models auf; diese ermöglichen eine breite Palette von Anwendungsmöglichkeiten in verschiedenen Bereichen, ohne dass sie wesentlich modifiziert oder angepasst werden müssen,[1012] auch diese sind jedoch noch von einer technischen Singularität entfernt. Ob eine solche je erreichbar ist, lässt sich auf dem jetzigen technischen Stand nicht seriös beurteilen.[1013]

Diese Feststellung wirft die Frage auf, auf welchen konkreten Untersuchungsgegenstand sich die weiteren Ausführungen, insbes. zum innovativen Charakter von KI und den daraus abzuleitenden veränderten bzw. zu verändernden

1007 Vgl. *Russell/Norvig* 2021, 1035 ff. Es ist wichtig festzuhalten, dass *Turing* sich selbst nicht eindeutig festgelegt hat, ob eine Maschine, die das „imitation game" für sich entscheidet, tatsächlich eine denkende Maschine sein kann. Er setzte sich zwar mit den Argumenten für und wider die Möglichkeit denkender Maschinen auseinander, blieb aber eine eindeutige Antwort schuldig. Insofern wird dem Turing-Test u.a. eine bloß sprachnormative Bedeutung zugewiesen; seine Funktion soll also nicht darin liegen, Existenzaussagen aufzustellen, sondern lediglich darin, den Sprachgebrauch zu regeln, *Heintz* 1993, 273. Denn *Turing* führt aus „I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." – „Ich glaube, dass sich am Ende des Jahrhunderts der Wortgebrauch und die allgemeine akademische Meinung so sehr verändert haben werden, dass man von denkenden Maschinen sprechen kann, ohne Widerspruch erwarten zu müssen", *Turing* Mind 59 (1950), 433 (442).

1008 Überblicksartig dazu *Bostrom/Yudkowsky* 2014, 318 ff.

1009 Siehe zu AlphaGo bzw. AlphaFold 2 oben Kap. 2, A.II.2.

1010 *Kurzweil* 2005, 92, 264.

1011 Einen guten Überblick über die anstehenden Herausforderungen des machine learning geben *Döbel/Leis/Vogelsang u. a.* 2018, 64 ff.

1012 *Bommasani/Hudson/Adeli u. a.* 2021, 3 mit umfangreichen wN.

1013 Vgl. *Grace/Salvatier/Dafoe u. a.* Journal of Artificial Intelligence Research 62 (2018), 729; *Boucher* 2020, 13 ff.; *Lenzen* 2018, 33.

Verantwortungsstrukturen, konzentrieren werden. Es kann um derzeit bzw. in absehbarer Zeit technisch umsetzbare Systeme, v.a. auf der Grundlage des oben beschriebenen maschinellen Lernens, gehen oder um Systeme, die erst hypothetisch gedacht sind, sich im Innovationsprozess noch nicht annähernd über die Phase der Invention hinausbewegt haben, und deren Realisierung noch nicht sicher bzw. jedenfalls zeitlich gesehen nicht absehbar ist. Gerade in der Rechtswissenschaft[1014] wird oftmals letzterer Ansatz gewählt und entsprechend nehmen Diskussionen zur KI als Verantwortungssubjekt einen größeren Raum ein.[1015] Die vorliegende Arbeit wählt jedoch den entgegengesetzten Weg und nimmt nur derzeit bzw. in absehbarer Zeit technisch umsetzbare Systeme in den Blick. Es ist zwar Aufgabe der Rechtswissenschaft, über den Ist-Zustand hinaus zu denken und auch das in den Blick zu nehmen, zu analysieren und aufzuarbeiten,[1016] was sein könnte; dies gerade, wenn man auch Rechtswissenschaft in der Innovationslogik von Beschleunigung und systematischer Durchsetzung des Neuen denkt.[1017] Bereits jetzt zu erörtern, was eine „denkende Maschine" für Konsequenzen für das (Straf)Recht haben würde, würde jedoch bedeuten, den zweiten Schritt vor dem ersten machen zu wollen. Es sollte also zunächst aufgearbeitet werden, was eine technisch autonome, lernfähige und in Teilen opake Technik für bestehende Verantwortungsstrukturen bedeutet.

3. *Maschinelles Lernen als schwache KI*

Die Ausführungen zur schwachen vs. starken KI haben den Untersuchungsgegenstand dieser Arbeit noch einmal konkretisieren lassen. Zudem kann auf dieser Grundlage das maschinelle Lernen im Folgenden als eindeutig schwache KI eingeordnet und noch einmal aufgezeigt werden, was dieses zum aktuellen Stand in der Lage ist zu leisten bzw. was gerade nicht.

Searle argumentiert, dass die Rechenoperationen eines Computers rein formal oder syntaktisch ablaufen – „Computation is defined purely formally or syntactically" –, es gehe um Regeln zur Bildung von Zeichenfolgen. Aus rein syntaktischen Rechenoperationen folgten aber keine semantischen Zustände – „we cannot get from syntactical to the semantic just by having the syntactical operations".

Seine Argumentation und das zugrundeliegende Beispiel des Chinese Room Argument beziehen sich auf eine – in der Theorie gedachte – KI, die auf einem symbolbasierten Ansatz beruht:[1018] Denn die Anleitung („book of instructions"),

1014 So z.B. bei *Gaede* 2019, 24 ff. und dann passim; *Christoph* 2022, 70; siehe im Übrigen zu den Beiträgen, die sich mit der Frage v.a. einer Schuldfähigkeit von (starken) KI-Systemen beschäftigen die Nachweise unten Kap. 3, C.I.2., 3.

1015 Ebenfalls krit. zu dieser Entwicklung *Beck* 2020a, § 7 Rn. 79.

1016 So auch *Simmler/Markwalder* ZStW 129 (2017), 20 (22); *Beck* JR 2009, 225 (230).

1017 Siehe dazu oben Kap. 1, A.I.

1018 *Cole* 2020.

wie chinesische Schriftzeichen anzuordnen sind, um eine richtige Antwort zu geben, d.h. das Programm, ist vorgegeben.

Überträgt man *Searles* Gedankenexperiment des Chinese Room Argument auf das maschinelle Lernen, ergibt sich insofern ein Unterschied. Denn das KI-System bildet die syntaktischen Sprachregeln aus einer Vielzahl entsprechender Beispielsdaten selbst. Die Regelbildung erfolgt auf der Grundlage statistischer Methoden, indem das KI-System Muster oder statistische Korrelationen in Daten erkennt.[1019] Würde man das Chinese Room Argument auf maschinelle Lernmethoden anpassen, würde also der Computer über keine – menschengemachte – Anleitung verfügen, wie die chinesischen Schriftzeichen anzuordnen sind. Er würde sich diese Regeln vielmehr selbst auf Grundlage einer Datenbasis, der „boxes of Chinese symbols" im Raum, selbst erschließen.

Der Übergang von einer „deterministischen Maschine" im Sinne einer symbolbasierten KI zu einer „probabilistischen Maschine", die auf statistischen Lernmethoden beruht, ändert jedoch an der Schlussfolgerung aus dem Chinese Room Argument nichts.[1020]

Ein KI-System, das auf maschinellen Lernverfahren beruht, mag bspw. Bilder von Hunden, Katzen oder Flugzeugen ähnlich zuverlässig wie ein Mensch erkennen. Diese Fähigkeit basiert darauf, dass das KI-System in einem großen Datenpool bestimmte Muster, d.h. statistische Korrelationen erkennen kann. Diese müssen aber nicht zwingend auf tatsächlichen Kausalitäten beruhen. Aus Korrelationen werden auf der Grundlage derzeitiger KI-Systeme keine Kausalitäten. Computer erkennen nicht, *was* ein Hund, eine Katze oder ein Flugzeug bedeutet. Oder in den Worten *Searles*: „We cannot get from syntactical to the semantic just by having the syntactical operations and nothing else".[1021]

Danks führt in diesem Zusammenhang zum maschinellen Lernen aus: [1022]

> „The successes of machine learning result from structural inference; these methods use patterns or statistical regularities in the data, and are (relatively speaking) indifferent to the semantics of the input variables."[1023]
>
> „Die Erfolge des maschinellen Lernens resultieren aus struktureller Schlussfolgerung; diese Methoden nutzen Muster oder statistische Regelmäßigkeiten in den Daten und sind (relativ gesehen) gleichgültig gegenüber der Semantik der Eingabevariablen."

Der Begriff des maschinellen Lernens könnte daher auch mit dem Begriff der statistischen Mustererkennung ersetzt werden und man sollte sich jedenfalls davor hüten, das maschinelle Lernen ungefiltert mit dem menschlichen Lernen gleichzusetzen.

1019 S. zum maschinellen Lernen oben Kap. 2, A.II.2.
1020 Vgl. auch *Nida-Rümelin/Weidenfeld* 2018, 47.
1021 Nachweis siehe oben Fn. 1006.
1022 *Danks* 2014, 159.
1023 *Danks* 2014, 160.

Auch KI-Systeme, die auf maschinellem Lernen basieren, unterfallen demnach jedenfalls einem Verständnis von schwacher KI, wie es *Searle* formuliert. Und nach den Vertretern einer starken KI sind diese Systeme wie gesehen noch nicht leistungsfähig genug und damit in diesem Sinne keine starke, sondern eine schwache KI.

Es bleiben natürlich an dieser Stelle die Fragen offen, ob mit *Searle* ein kategorialer Unterschied zwischen den Rechenoperationen eines KI-Systems und dem eines Menschen besteht oder mit den Vertretern einer starken KI lediglich ein relativer und, wenn man ersterer Ansicht folgt, worin dieser Unterschied bestehen soll.[1024] Wie sich im weiteren Verlauf dieser Arbeit zeigen wird, kann eine definitive Antwort auf diese (empirisch noch nicht zu beantwortende) Frage ausbleiben, da es jedenfalls aus der Perspektive des Strafrechts darauf nicht ankommt.[1025]

V. Rückblick und Ausblick

Der vorangegangene Abschnitt sollte dazu dienen, den Untersuchungsgegenstand dieser Arbeit zu präzisieren.

1. Dafür wurden zunächst Ansätze zum Begriff der Künstlichen Intelligenz vorgestellt: Ein v.a. auf *Turing* und sein „imitation game" zurückgehender Ansatz stellt darauf ab, ob ein Computer menschliches Verhalten imitieren kann. Damit zeigt dieser Ansatz gleichzeitig auf, welche Fähigkeiten zu dieser Imitationsleistung erforderlich sind, etwa die Fähigkeit zur Sprach- und Bildverarbeitung, Lernfähigkeit oder die Fähigkeit zur Robotik. Der anwendungsorientierte Ansatz des intelligenten Agenten beschreibt KI hingegen v.a. im Hinblick auf die erforderliche Architektur künstlich intelligenter Systeme (sense, plan, act) und gibt als Maßstab für Intelligenz ein rationales, d.h. v. a. ein zielorientiertes Verhalten vor. Ein solches Agieren erfordert v.a. (technische) Autonomie und Lernfähigkeit des intelligenten Agenten.

1. Lernfähigkeit bedeutet, dass ein System in der Lage ist, nicht nur auf der Grundlage von entwicklerseitig vorgegebenem Tatsachenwissen zu operieren, sondern auch auf der Grundlage „eigener Wahrnehmung"; das System kann insofern seinen Output im Hinblick auf das gesammelte Wissen anpassen und verbessern. Während Lernfähigkeit v.a. eine Eigenschaft eines künstlich intelligenten Systems beschreibt, beschreibt der Begriff der (technischen) Autonomie das Verhältnis des Systems zum Entwickler bzw. dem Nutzer. Weil das System selbstständig Tatsachenwissen sammeln kann und weil es bei der Entscheidung, wie es agiert, entwicklerseitig einen Spielraum eingeräumt

1024 Siehe dazu unten Kap. 3, C.II.2.a.
1025 Siehe unten Kap. 3, C.II.2.b.

bekommen hat, besitzt es eine gewisse Unabhängigkeit von Entwickler- und Nutzervorgaben. Es ist wissens- und entscheidungsautonom.

2. Zu diesen beiden Aspekten technischer Autonomie tritt im Falle maschinellen Lernens Regelautonomie hinzu: Das System bildet die Regeln, auf Grund derer es operiert, selbstständig. Grundlage dafür sind Daten und darin abgebildete statistische Korrelationen. Für das maschinelle Lernen stehen verschiedene Lernmethoden zur Verfügung – allen voran die künstlichen neuronalen Netze – sowie verschiedene Lernformen (das überwachte/unüberwachte Lernen, das verstärkende Lernen). Die Regelautonomie, die Möglichkeit des Systems seine Regeln, auf Grund derer es operiert, selbstständig zu verändern, endet jedoch in der Regel mit der sog. Trainingsphase. Gängige KI-Systeme sind sog. offline-Systeme, die im Betrieb nicht weiterlernen. Die oft verbreitete Auffassung, dass KI-Systeme ihre Regeln *stets* selbstständig ändern können, ist so jedenfalls nicht richtig.
3. Zur Lernfähigkeit und technischen Autonomie künstlich intelligenter Systeme tritt beim maschinellen Lernen, v.a. in Form künstlicher neuronaler Netze, ein weiteres Charakteristikum hinzu, nämlich das der Opazität oder des Black-Box-Effekts. Opazität bedeutet, dass die von KI-Systemen gebildeten Regeln – ihre Transformationsfunktionen – für Menschen und gerade auch für Experten zum derzeitigen Stand der Wissenschaft nicht verstehbar sind. Dies liegt begründet in der Komplexität von KI-Systemen – die Rechenprozesse sind zu umfangreich, als dass ein Mensch diese in angemessener Zeit nachrechnen könnte (relative Opazität) – sowie in der analytischen Unverständlichkeit von ML-Modellen (absolute Opazität). Die KI „spricht" auf eine für Menschen nicht verständliche Weise.
4. Technisch autonome und lernfähige KI-Systeme sind trotz ihrer Leistungsfähigkeit der sog. schwachen KI zuzuordnen. Auf der Grundlage maschinellen Lernens sind sie in der Lage, erfolgreich Muster, d.h. statistische Korrelationen, in Datensätzen zu erkennen. Aus Korrelationen werden jedoch keine Kausalitäten. Ob darüber hinaus zukünftige KI-Systeme so leistungsfähig gemacht werden können, dass sie von Vertretern einer starker KI als eine solche qualifiziert würden, ist umstritten. Jedenfalls ist eine solche KI derzeit und auch in absehbarer Zeit technisch nicht umsetzbar. Die vorliegende Arbeit blendet diese Form von KI daher aus.

B. Künstliche Intelligenz als technische Innovation

Im Folgenden wird nun analysiert, ob und weshalb KI eine technische Innovation ist, weil sie nicht dem klassischen Verständnis von Technik entspricht und das technische Risiko verändert. Diese Ausführungen bauen auf dem im Grund-

lagenteil erarbeiteten Verständnis von Technik und Risiko auf. Entsprechend der Ausdifferenzierung des Technikbegriffs werden KI-Technik als Form (I.), KI als Experiment (II.) und KI als Erwartung (III.) untersucht; dies ermöglicht die Feststellung, dass KI eine „transklassische" Technik ist (IV.). Zuletzt wird die Verknüpfung von KI-Technik und Risiko untersucht, um Veränderungen des technischen Risikos herauszuarbeiten (V.).

I. KI als Form

Klassische Technik als *Form* beschreibt das Regelhafte, das nach einer festen Transformationsregel einen bestimmten Input in den immer gleichen Output verwandelt. Gleichzeitig bestimmt sich Technik damit durch eine Trennung von Form und Kontext.[1026] KI-Systeme entsprechen an sich dieser klassischen Grundform; auch sie verwandeln Input in Output. Auf Grund ihrer Regel-, Wissens- und Entscheidungsautonomie verändern sich jedoch bestimmte Elemente von Technik als Form.

Solange sich eine KI im „Lernprozess" befindet, ist die Transformationsregel nicht statisch: Auf einen bestimmten Input x folgt nicht immer der Output y, sondern auf den Input m folgt einmal der Output p, das nächste Mal aber q oder r. Während des Lernprozesses bestehen keine *fixed rules* (nach *Turing*), keine statische Determiniertheit (nach *von Foerster*) und keine rigide Kopplung (nach *Luhmann*); *von Foerster* würde anstatt von einer trivialen von einer nicht-trivialen Maschine sprechen.[1027] Sobald und sofern ein KI-System nach dem Training „eingefroren" wird, endet aber die Varianz der Transformationsregel und sie operiert wieder „trivial" und auf klassische Weise. Lediglich bei „weiterlernenden" Systemen (online-learning)[1028] hält diese Eigenschaft an und diese Form der KI-Technik erfüllt dauerhaft nicht die klassische Vorstellung von Technik, die einen Input nach einer festen Transformationsregel in den immer gleichen Output verwandelt.

Die Regelautonomie (solange sie anhält) sowie die Wissens- und Entscheidungsautonomie stellen zudem einen weiteren Aspekt von Technik als Form in Frage, nämlich das Vorhandensein einer artifiziellen Grenze zwischen dem Inneren der Maschine und dem komplexen Außen der Umgebung – der Trennung von Form und Kontext. Ist ein KI-System wissens- und entscheidungsautonom, weil es auf der Grundlage von Informationen aus seiner Umgebung seine vom Entwickler vorgegebene Wissensbasis erweitern und innerhalb des vom Entwickler vorgegebenen Entscheidungsspielraums seinen Output anpassen kann,

1026 S.o. Kap. 1, B.II.1.c.
1027 *Kaminski* 2014b, 74 f.; *ders.* 2014a, 13 f.; *Kaminski/Glass* 2019, 132; *Kaminski* 2020, 157.
1028 Zur Unterscheidung zwischen offline- und online-learning bzw. lernenden und weiterlernenden Systemen siehe oben Kap. 2, A.II.7.

bestimmt die Umgebung mit, nach welchen Regeln Technik operiert. Form und Kontext verschwimmen.[1029]

II. KI als Experiment

Unter Technik als Experiment wurde das technische Handeln untersucht, das Steuern sowie das Regeln. Zunächst zum *Steuern* im Sinne eines Einsatzes von Technik als Mittel zum Zweck:

Steuern kann menschliches Handeln verstärken oder entlasten. Ein Fahrrad setzt menschliche Muskelkraft effizienter um als bei der natürlichen Art der Fortbewegung, ein Kfz ersetzt menschliche Muskelkraft gänzlich, muss aber noch durch den menschlichen Nutzer, den Fahrer, durch den Straßenverkehr navigiert werden. Steuern in einem klassischen Sinne bedeutet den Einsatz von Technik als Werkzeug, der menschliche Nutzer bleibt die „Zentralgestalt" des Geschehens.[1030]

Bei KI-Systemen rückt der „Anteil" menschlichen Handelns am Output in den Hintergrund: In der „Idealvorstellung"[1031] eines autonomen Fahrzeugs wird der Fahrer zum „Passagier", der zu Beginn der Fahrt einsteigt und nach Nennung des Fahrziels passiv bleiben kann, alles andere erledigt die KI – die Berechnung der Route, das Navigieren durch den Straßenverkehr. Beim Steuern von KI ist zwar immer noch menschliches Handeln erforderlich, dieses beschränkt sich jedoch weitestgehend auf das Ingangsetzen der Technik – der Anteil der technischen Prozesse am „Gesamtgeschehen" hingegen nimmt zu. Jedenfalls aber setzt die Nutzung einer KI noch immer voraus, dass ein Mensch diese als Mittel zum Zweck einsetzen will; dies gilt auch bei weiterlernenden Systemen, die im Betrieb ihre Transformationsregel ändern können.

Auch bei der *Entwicklung* von KI-Systemen nimmt der menschliche Anteil ab, weil das System die Regeln, auf Grund derer es operiert, selbst lernt und der Entwickler nicht jedes Detail festlegt, sondern einen Raum schafft, innerhalb dessen das KI-System agieren kann. Dennoch ist der menschliche Anteil bei der Entwicklung solcher Systeme, insbes. beim maschinellen Lernen, weiterhin hoch und dient dazu, dass ein KI-System den vorgegebenen Zweck erzielt und im späteren Einsatz die Technik regelbasiert agiert.[1032]

Technische Autonomie bewirkt, dass ein KI-System nicht *äußerlich* kontrollierbar ist – weil es im Einzelfall nicht steuerbar ist bzw. die Entwickler die Funktionsweise nicht im Detail vorgeben (was konzeptionell ja gerade so sein soll); es ist aber sehr wohl *innerlich* kontrollierbar, weil alle Zwecke und Ver-

1029 *Kaminski* 2014a, 13 ff.; *Resch/Kaminski* MaM 29 (2019), 9 (12).
1030 Zum technischen Handeln in der Form des Steuerns oben Kap. 1, B.II.2.
1031 Siehe zu den Autonomiegraden beim autonomen Fahren unten Fn. 1119.
1032 Siehe zum menschlichen Anteil beim maschinellen Lernen oben Kap. 2, A.II.9.

haltensmöglichkeiten durch die Entwickler gesetzt wurden und die Nutzer über ihren Einsatz entscheiden.

KI-Technik hat auch Auswirkungen auf den „experimentellen" Anteil technischen Handelns, auf das *Regeln* als Gewährleistung des Steuerungserfolgs: KI-Systeme sind konzeptionell nicht darauf angelegt, das System vollständig gegen Umwelteinflüsse von außen abzuschirmen. Sie sind offen gegenüber ihrer Umgebung, ihrem Kontext, um ihr Tatsachen- und Regelwissen zu erweitern. Dies verändert das „Regeln" im Sinne der Fehlersuche und -behebung, sofern KI „kaputt" ist oder „fehlerhaft" operiert.[1033] Regeln in diesem Sinne setzt zunächst einmal die Feststellung voraus, dass die Technik fehlerhaft operiert; im Falle klassischer Technik ist diese Feststellung einfach, weil der erwartete Output nicht eintritt. Ein Fahrzeug springt nicht an, bei Betätigen des Lichtschalters geht das Licht nicht an etc. Wenn aber Form und Kontext verschwimmen, ist es nicht einfach, festzustellen, ob die enttäuschte Erwartung auf einen Output x auf einem bereits im System vorhandenen Fehler beruht oder auf einer neuen Regelbildung bzw. auf neuem Tatsachenwissen.[1034] Sofern die Feststellung „fehlerhaft" und „kaputt" getroffen ist, kann auf Grund der Offenheit des Systems gegenüber seinem Kontext wiederum eine Vielzahl von Umwelteinflüssen ursächlich sein, die den veränderten Output erklären kann. Ist eine erklärende Hypothese schließlich gefunden, kann diese zwar induktiv erhärtet werden, sofern der Fehler in einer Vielzahl von Fällen auftritt. Nimmt man aber hinzu, dass das KI-System epistemisch opak ist, ist eine deduzierende Bestätigung nicht möglich. Denn die allgemeinen Regeln, auf Grund derer die KI operiert, sind (noch) nicht bekannt.[1035]

III. KI als Erwartung

Zuletzt zu den menschlichen Erwartungen an KI-Technik:

Die *praktische* Verstehbarkeit – also zu wissen, wie Input und Output zusammenhängen – kann bei KI-Systemen auf Grund ihrer technischen Autonomie herabgesetzt sein. Einerseits kann der Zusammenhang zwischen Input und Output nicht erkannt werden, solange der Lernprozess anhält und die Transformationsregel nicht statisch ist. Aber auch bei offline-Systemen (die in der Anwendungsphase nicht weiter lernen) kann die praktische Verstehbarkeit eingeschränkt sein, wenn das KI-System auf Grund seiner Wahrnehmungsfähigkeit das Tatsachenwissen erweitert, also auf einen neuen Input trifft, und auf Grund dessen seinen Output anpasst. Das System wird dadurch komplexer und so kann es schwierig

1033 Zu diesem Aspekt des Regelns bzw. zur Leitunterscheidung „fehlerfrei/heil" und „fehlerhaft/kaputt" oben Kap. 1, B.II.2.

1034 *Kaminski* 2014a, 15; *Kaminski/Glass* 2019, 132; *Resch/Kaminski* MaM 29 (2019), 9 (12). Vgl. auch *Nordmann* 2008b, 177 f.

1035 Dazu *Hubig/Harras* 2014, 44 f., 48. Vertiefend *Hubig* 2006, 208 ff.

sein, Systemzusammenhänge zu verstehen.[1036] In dem Maße, wie die praktische Verstehbarkeit eingeschränkt sein kann, können die Erwartungen der Vorhersehbarkeit und Wiederholbarkeit eines bestimmten Outputs bei KI-Technik herabgesetzt sein: Auf Grund der Wissens- und Regelautonomie kann sich der Output verändern und ist als solcher nicht sicher vorhersehbar. Gleichzeitig ist ein einmal eingetretener Output dann auch nicht beliebig wiederholbar.[1037]

Erheblich herabgesetzt bzw. sogar teilweise ausgeschlossen ist zuletzt die *epistemische* Verstehbarkeit, also zu wissen, *warum* Input und Output zusammenhängen. KI-Systeme sind oft hochgradig komplex, so dass ein Nachrechnen der einzelnen Rechenschritte kaum möglich ist, zudem werden in vielen Fällen Lernmethoden eingesetzt, die menschlich nicht verstehbare Regeln ausbilden. Sogar Experten können nicht mehr in der Lage sein, KI-Systeme zu verstehen.[1038]

IV. Zwischenstand – KI als transklassische Technik

KI ist Technik, aber sie entspricht nicht der klassischen Vorstellung von Technik: Sie ist das Gegenteil von Regelhaftigkeit, solange der Lernprozess anhält; Form und Kontext, Technik und Umwelt verschwimmen. KI verdrängt den Menschen als Zentralgestalt des technischen Prozesses, weil der Anteil des menschlichen Steuerns abnimmt und dieser Prozess weniger verstehbar wird. Zudem verändert sich das technische Handeln, das zur Aufrechterhaltung der Regelhaftigkeit aufgewendet wird: Es geht nicht darum die Form der Technik gegen den Kontext, gegen Umwelteinflüsse, abzuschirmen, vielmehr bezieht KI die Umwelt gerade mit ein. Dadurch wird im Falle kaputter und fehlerhaft agierender KI die Fehlersuche und -behebung erschwert, d.h. die damit verbundenen Schritte von Abduktion, Induktion und Deduktion.

KI-Technik ist eine innovative, weil transklassische Technik.[1039]

V. KI-Technik und Risiko

Auf der Grundlage der „klassischen" Verknüpfung von Technik und Risiko zuletzt zur Frage, ob und wie der transklassische Charakter von KI auch das technische Risiko verändert:

1036 Vgl. *Kaminski/Glass* 2019, 132.
1037 *Kaminski* 2020, 157.
1038 *Kaminski/Glass* 2019, 132; *Kaminski* 2020, 158, 161.
1039 Der Begriff der transklassischen Technik geht zurück auf *Hubig/Harras* 2014, 45: Es seien technische Systemzustände nicht bekannt und für die Mensch-Technik-Beziehung könne ein Verlust an Transparenz (das technische Subjekt sei epistemisch eingeschränkt) und an Gestaltbarkeit (das technische Subjekt sei als Aktionssubjekt zurückgedrängt von autonom gewordener Technik) diagnostiziert werden. Den Begriff der transklassischen Technik verwendet ebenso *Kaminski* 2020, 154 ff.

1. KI-Technik und Risikoprognose

a. KI und Risikowissen

Der Risikobegriff enthält immer ein prognostisches Element – eine Risikoentscheidung erfordert eine Aussage darüber, welche der möglichen Folgen eintreten kann. Dafür sind Risikowissen über bestimmte gesetzmäßige Zusammenhänge und eine Verknüpfung dieses Wissens mit Wahrscheinlichkeitsangaben erforderlich.[1040]

Wie präzise eine Risikoprognose ist bzw. ob eine solche überhaupt getätigt werden kann, hängt von der Qualität des Wissens über gesetzmäßige Zusammenhänge ab. Das Risikowissen über technische Zusammenhänge im Allgemeinen hat regelmäßig einen dynamischen Charakter. Befindet sich eine Technik noch in der Phase der Innovation, ist das Wissen über gesetzmäßige Zusammenhänge eingeschränkt oder noch nicht vorhanden, es kann an deduktiven Erklärungsmodellen (einem epistemischen Verstehen) fehlen und für induktive Erklärungsmodelle fehlt es an Erfahrungswissen. In der Phase der Diffusion einer Technik wächst regelmäßig das Erfahrungswissen, da mit dem „Experiment Technik" Wissen über mögliche Störfaktoren gewonnen wird und eine Risikoprognose präziser werden lässt.[1041]

Bezogen auf KI-Systeme: Die Risikoprognose bei KI-Systemen ist allein deswegen schon weniger präzise, weil sich diese derzeit in der Innovationsphase einer erstmaligen Entwicklung und Nutzung und manche gar erst in der Phase der Invention befinden. Das Erfahrungswissen über gesetzmäßige Zusammenhänge und über diejenigen Fälle, in denen der Steuerungserfolg ausbleibt, beruht lediglich auf den Daten aus der Trainingsphase bzw. ersten Daten aus der Einsatzphase. Beginnt jedoch das „Experiment Technik" mit einer breiten Diffusion von KI-Systemen, wird das Erfahrungswissen über gesetzmäßige Zusammenhänge und mögliche Störfaktoren zunehmen und entsprechend die Risikoprognose präziser werden.

Die technische Autonomie und epistemische Opazität von KI-Systemen führen zudem dazu, dass die Risikoprognose gegenüber klassischer Technik dauerhaft erschwert ist. Anzuknüpfen ist hierfür an die obige Beschreibung von „KI-Technik als Experiment" und „KI-Technik und Erwartung"; übertragen auf die Risikoprognose bedeutet dies: Handelt es sich um (wegen ihrer Komplexität relativ oder absolut) epistemisch opake Systeme, beruht die Risikoprognose auf statistischen Daten aus der Trainings- und schließlich der Einsatzphase. Ein Entwickler kann vorhersagen, wie genau und präzise das System unter welchen Umständen

1040 Siehe oben Kap. 1, B.III.1.b.aa.
1041 S.o. Kap. 1, B.II.2. sowie B.III.1.b.aa.

agieren wird, weil das KI-System statistisch transparent ist.[1042] Der Nutzen dieses statistischen Wissens trägt aber nur insoweit, als die konkrete Anwendung eines KI-Systems bereits bekannt ist. Handelt es sich um Foundation Models, die auf Grund ihrer Fähigkeit zum Transferlernen (Wissen kann von einer Aufgabe auf die nächste übertragen werden) verschiedene Anwendungsmöglichkeiten haben, ist der Nutzen statistischer Werte für die Risikoprognose eingeschränkt.[1043] Dies gilt ebenso für weiterlernende Systeme, da sich das statistische Wissen auf die Trainingsphase bezieht.[1044] Fehlt es daneben an einem deduktiven Erklärungsmodell, weshalb die Technik so funktioniert, wie sie funktioniert, wird es umso schwieriger, die potentiellen Störfaktoren auf das System bzw. die systemimmanenten Fehler zu prognostizieren bzw. im Falle von Fehlfunktionen vermutete Störfaktoren auf die Technik zurück zu beziehen. Verdeutlicht am simplen Beispiel einer Bilderkennungs-KI, welche in der Trainingsphase zuverlässig Flugzeuge zu erkennen vermochte, aber in der Einsatzphase versagte: Weil die Transformationsregeln unbekannt waren, konnte zunächst nicht vorhergesehen werden, dass die Bilderkennungs-KI Flugzeuge nur deshalb „zuverlässig" erkennen hatte können, weil es den blauen Hintergrund, der bei allen Bildern mit Flugzeugen zu sehen war, als relevantes Merkmal herangezogen hatte.[1045] Zusammengefasst in den Worten *Teubners*: „Nur noch experimentell kann das Neue ausprobiert, nicht mehr vorherberechnet, sondern nur noch nachträglich auf seine Folgen überprüft werden."[1046]

Die Risikoprognose ist also bei KI-Systemen dauerhaft erschwert, Risiken sind aber nicht völlig unvorhersehbar, da jeweils statistisches Wissen vorhanden ist, welches mit dem Praxiseinsatz von KI anwachsen wird.[1047]

1042 S.o. Kap. 2, A.III.4.

1043 Zu Foundation Models bereits oben Kap. 2, A.IV.2.b., zudem unten Kap. 2, B.V.4.

1044 Vgl. *Zech* 2020, A37.

1045 Zu diesem Beispiel *Döbel/Leis/Vogelsang u. a.* 2018, 47.

1046 *Teubner* AcP 218 (2018), 155 (176).

1047 A.A. hingegen scheinbar *Teubner* AcP 218 (2018), 155 (164): „Besonders einschneidend ist das Autonomierisiko, das vom *prinzipiell unvorhersehbaren* Verhalten selbstlernender Algorithmen erzeugt wird." – Hervorhebungen durch die Verf.; zudem *Zech* 2016, 175: „Ein besonderes und neuartiges Risiko ergibt sich aus dem *prinzipiell unvorhersehbaren Verhalten* selbstlernender Algorithmen." – Hervorhebungen durch die Verf. Entschärfen lassen sich diese scheinbaren Unterschiede mit einer von *Bostrom* und *Yudkowsky* in diesem Kontext vorgenommenen Unterscheidung zwischen lokalem und nicht-lokalem Wissen über ein KI-System (*Bostrom/Yudkowsky* 2014, 319 f.): Vorhersehbar bleibt das generelle Agieren eines KI-Systems, weil diesbezüglich insbes. auf Grund statistischen Wissens eine Risikoprognose möglich ist (nicht-lokales Wissen); nicht vorhersehbar mag dagegen das Agieren eines KI-Systems im Einzelfall sein (lokales Wissen). Eine solche Vorhersehbarkeit ist für klassische Technik aber ebenfalls nicht gegeben, siehe oben Kap. 1, B.II.3. Siehe dazu in Ansätzen nun *Zech* 2020, A42.

b. Erhöhte Schadenswahrscheinlichkeit und Schadenshöhe?

> „Die Unberechenbarkeit des Systemverhaltens erhöht dabei u.a. das Risiko für die mit der KI konfrontierten Rechtsgüter."[1048] „Aufgrund der Unberechenbarkeit deren Verhaltens besteht eine hohe Wahrscheinlichkeit, dass die Schadenshöhe über dem Level liegen wird, das bei herkömmlichen Produkten nach einer Fehlfunktion üblich ist."[1049]
> „Die von diesen Systemen ausgehende Unberechenbarkeit hat zur Folge, dass eine exakte, hundertprozentige Vorhersage, wie das System in einer Situation agiert, nicht möglich ist. Auf der einen Seite bedeutet das, dass je unvorhersehbarer und somit risikoreicher ein solches Produkt ist, desto wahrscheinlicher sind Schäden […]."[1050]

Aus obigen Zitaten lässt sich die Aussage ableiten, dass aus einer Unvorhersehbarkeit KI-technischen Agierens (in der obigen Diktion: aus eingeschränktem Risikowissen) eine per se erhöhte Schadenswahrscheinlichkeit und Schadenshöhe beim Einsatz von KI-Systemen gegenüber klassischer Technik folge.[1051] *Lohmann* etwa argumentiert, dass ein autonom agierendes System, dessen Systemverhalten unvorhersehbar sei, stets die Schadens*wahrscheinlichkeit* erhöhe („je unvorhersehbarer und somit risikoreicher"). Nach *Haagen* folgt daraus auch ein per se erhöhtes Schadens*ausmaß*, da eine Fehlfunktion für den nicht mit der Bedienung beschäftigten Nutzer einen Überraschungseffekt bedeute, der zu einer verlängerten Reaktionszeit bei der Schadensabwehr führe; damit könne ein Schaden weniger effektiv abgewehrt werden und sich ein Schaden intensivieren. Der Nutzer könne – so *Haagen* – nicht trainieren, wie er das autonome System bei Fehlfunktionen beherrschen und Schaden abwenden könne; anders bei klassischer Technik:

> „Herkömmliche Produkte kennen nur die gleichen Abläufe, sodass der Nutzer diese stets beherrschen kann. Er weiß [sic!] was auf ihn zukommt und kann sich darauf einstellen, wie er zu reagieren hat."[1052]

Diese Annahme, dass sich aus fehlendem Wissen über das Agieren eines KI-Systems ein per se erhöhtes Risiko ergebe, vermag indes nicht zu überzeugen. Fehlendes Risikowissen und ein erhöhtes Risiko stehen nicht objektiv, sondern lediglich subjektiv in einem Zusammenhang: Dass ein Nutzer nicht vorhersehen kann, wie ein autonomes Fahrzeug ihn zur Arbeit bringt, auf welcher Route, mit welchen Verkehrsvorgängen etc. bedeutet nicht zwangsläufig, dass sich hieraus mit einer erhöhten Wahrscheinlichkeit ein Schaden entwickeln wird. Es ist möglich, dass das System so verlässlich ist wie ein Fahrzeug klassischer Art und

1048 *Haagen* 2021, 220.
1049 *Haagen* 2021, 222.
1050 *Lohmann* 2021, 158 f.
1051 Ähnlich sprechen *Gless* und *Weigend* von „Risiken [, die] besonders hoch und schwer zu kontrollieren" seien, *Gleß/Weigend* ZStW 126 (2014), 561 (583).
1052 *Haagen* 2021, 223; zu diesem Gedanken auch *Coeckelbergh* SEE 26 (2020), 2051 (2055).

ein Fahrer, der dies steuert.[1053] Die Risikoforschung hat hingegen gezeigt, dass in Fällen eingeschränkten Risikowissens und eingeschränkter Risikokontrolle das subjektiv empfundene Risiko erhöht ist.[1054] Dies hat aber keinen Einfluss auf das objektive Risiko.

Fehlendes Risikowissen und eine verminderte Steuerung KI-technischer Systeme bewirken zwar, dass menschliche Abwehrmaßnahmen gegen einen drohenden Schaden, die den Eintritt eines Schadens bzw. jedenfalls die Schadensintensität verringern könnten, schwerer und später ergriffen werden können. Aber auch daraus lässt sich nicht pauschal eine erhöhte Schadenswahrscheinlichkeit und Schadenshöhe ableiten, denn dies würde voraussetzen, dass klassische Technik und klassisches technisches Handeln eines Menschen ein geringeres Risiko bergen. Angesichts der Tatsache, dass auch klassische Technik bzw. der Mensch bei seinem technischen Handeln versagen kann, überzeugt dieser Schluss nicht.

KI-Systeme mögen im Einzelfall und je nach Art ihrer Anwendung ein erhöhtes Risiko gegenüber klassischen Systemen haben, allein auf Grund ihrer Unterschiede zur klassischen Technik lässt sich ein solcher Schluss jedoch nicht ziehen.[1055]

2. *KI-Technik: Risiko oder Gefahr?*

> „But once we think of these as technical systems in their own right, naturalized technologies cease to be objects of science and of experience, they take on a life of their own such that we no Ionger appear to perceive, comprehend, or control them, such that we no longer think of them as mechanisms or something 'devised by human Wit', but something instead that has receded into the fabric of uncomprehended nature with its occult qualities."[1056]
>
> „Sobald wir sie jedoch als eigenständige technische Systeme betrachten, hören naturalisierte Technologien auf, Objekte der Wissenschaft und der Erfahrung zu sein, sie entwickeln ein Eigenleben, so dass wir sie scheinbar nicht mehr wahrnehmen, verstehen oder kontrollieren, so dass wir sie nicht mehr als Mechanismen oder als etwas ‚von menschlichem Geist Erfundenes' betrachten, sondern als etwas, das in das

1053 Erste Studien deuten darauf hin, dass autonome Fahrzeuge insgesamt weniger Unfälle verursachen als Fahrzeuge, die von Menschen gesteuert werden. Eine Studie spricht von 1/3 weniger Verkehrsunfällen, die durch eine präzisere Wahrnehmungsfähigkeit und durch den Wegfall menschlicher Fahruntüchtigkeit vermieden werden könnten. Um die Unfallrate weiter zu senken, müssten für autonome Fahrzeuge erhöhte Sicherheitsmaßnahmen, etwa eine reduzierte Geschwindigkeit, vorgesehen werden, *Mueller/Cicchino/Zuby* Journal of safety research 75 (2020), 310. Eine Studie von McKinsey kommt zudem zum Ergebnis, dass 90 % aller *tödlichen* Unfälle durch autonome Fahrzeuge verhindert werden könnten, *Bertoncello/Wee* 2015.

1054 Siehe dazu oben Kap. 1, B.III.1.b.aa.

1055 Ähnlich *Beck* 2020b, 453 Rn. 17, die betont, dass eine statistische Vorhersage über das Risiko durch die Herstellung und Nutzung autonomer Fahrzeuge mangels ausreichender Erfahrungswerte noch nicht möglich sei.

1056 *Nordmann* 2008b, 176. Erläuternd hierzu in Bezug auf KI-Systeme *Kaminski* 2014b, 61, der dafür den Begriff der „lernenden Maschinen" verwendet.

> Gefüge der unbegreiflichen Natur mit ihren verborgenen Eigenschaften zurückgetreten ist."

Nach *Nordmann* weist KI-Technik Eigenschaften auf, die wir mit der Natur verbinden – indisponibel und menschlicher Einsicht oft verschlossen – er spricht von „the fabric of uncomprehended nature with its occult qualities". Knüpft man an die *Luhmann*`sche Unterscheidung zwischen Risiko und Gefahr an, könnten auf der Grundlage dieser Aussage Schadenspotentiale einer KI schon gar keine Risiken mehr sein, sondern Gefahren; denn was per se unvorhersehbar und unbeherrschbar ist, kann auch nicht auf eine menschliche Entscheidung zurückbezogen werden.[1057] Und diese Aussage könnte die eingangs formulierte These bestätigen, dass sich mit KI eine Technik zu entwickeln scheint, die sich aus der (vermeintlichen) Steuerung und Beherrschbarkeit durch den Menschen herauslöst und als autonomer und in seiner Funktionsweise opaker Akteur dem Menschen als „zweite Natur" gegenübertritt:

In der Tat – die technische Autonomie künstlich intelligenter Systeme führt dazu, dass der Anteil menschlichen Steuerns bei der Nutzung und bei der Entwicklung von KI-Systemen abnimmt. Zudem bewirkt die Opazität von KI, dass ihr Agieren und ihre Funktionsweise nicht vorhersehbar sein können. Dies bedeutet aber nicht, dass KI-Systeme dem Menschen als „zweite" Natur gegenübertreten würden. KI-Systeme sind weiterhin innerlich steuerbar, weil der menschliche Einfluss auf den technischen Output bei der Entwicklung hoch ist und der Nutzer über den Einsatz eines KI-Systems entscheidet. Die Präzision einer Risikoprognose ist zwar eingeschränkt, aber nicht vollständig ausgeschlossen. KI ist eine transklassische Technik, sie ist aber nicht auch eine „naturalisierte" Technik,[1058] Schadenspotentiale einer KI sind Risiken und keine Gefahren gleich dem Wirken der Natur.

3. KI-Technik: Sicherheitsrisiko und Verantwortungsrisiko

Im Grundlagenteil „Risiko und Technik" wurden drei Verknüpfungen von Technik und Risiko genannt: Das der Technik *immanente* Risiko des Versagens des Steuerungserfolgs; das mit dem Einsatz von Technik *generell* einhergehende Risiko, dass Umweltschäden eintreten; und zuletzt die Risikozunahme durch von Technik und Wissenschaft generiertes Wissen über Schadenspotentiale und wie diese im Zusammenhang mit menschlichen Entscheidungen stehen. Für den vorliegenden Untersuchungsgegenstand wurde die erste Verknüpfung von Technik und Risiko als relevant identifiziert – das der Technik immanente Risiko des Versagens des Steuerungserfolgs und der damit einhergehende Konflikt zwischen

1057 Zum Risikobegriff und zur Abgrenzung zur Gefahr siehe oben Kap. 1, B.III.1.a.
1058 Von einer „naturalisierten" Technik spricht aber *Nordmann* 2008b, passim.

Interessen der technischen Innovatoren, Nutzungsinteressen der Allgemeinheit und den Interessen der Betroffenen am Erhalt ihrer körperlichen Unversehrtheit.

Im Kontext KI-Technik und Risiko werden in Ergänzung zu dieser Risikoverknüpfung (im Folgenden wird diese als *Sicherheitsrisiko* bezeichnet) weitere Verknüpfungen diskutiert: Ausgehend davon, dass KI auch zunehmend eingesetzt wird (oder werden soll), um menschliche Entscheidungen zu ersetzen (etwa automatisierte Gesichtserkennung bei der Suche nach Tatverdächtigen, KI-basierte Lügendetektoren in Strafverfahren)[1059], soll es in dem aus dem Versagen des Steuerungserfolgs resultierenden Konflikt nicht nur um Interessen der Betroffenen im Hinblick auf Leben und körperliche Unversehrtheit gehen; es werden eine Reihe weiterer Rechtsgüter in Bezug genommen. Dazu stellvertretend das Weißbuch der Europäischen Kommission zur Künstlichen Intelligenz:

> „Infolge der Nutzung von KI können die Werte, auf denen die EU gründet, beeinträchtigt und Grundrechte verletzt werden. Dies gilt auch für das Recht auf freie Meinungsäußerung, die Versammlungsfreiheit, die Achtung der Menschenwürde, die Nichtdiskriminierung, den Schutz personenbezogener Daten und des Privatlebens, das Recht auf einen wirksamen gerichtlichen Rechtsbehelf und ein faires Verfahren sowie den Verbraucherschutz."[1060]

Diese Risiken für die genannten weiteren Rechtsgüter können nicht nur aus Fehlern in der Gestaltung von KI-Systemen, sondern auch aus Fehlern bei der Verwendung von Daten folgen:[1061] Wenn ein Objekterkennungssystem Bilder dunkelhäutiger Personen als Gorillas klassifiziert oder bei dunkelhäutigen Menschen mit einer viel geringeren Präzision als bei weißen Menschen operiert, wenn ein Spracherkennungsprogramm weibliche Stimmen schlechter erkennt als männliche Stimmen, ist dies oft darauf zurückzuführen, dass das Modell mit zu wenig Datenmaterial der jeweils diskriminierten Gruppe trainiert wurde.[1062]

Zudem wird eine weitere Verknüpfung von KI-Technik und Risiko diskutiert, die zwar ihren Ausgangspunkt im Versagen des Steuerungserfolgs und dadurch eintretenden Schäden hat. Es geht aber nicht um das unmittelbare technische Risiko, sondern um das mittelbare Risiko, dass für eingetretene Schäden retrospektiv niemand in Verantwortung genommen werden kann, weil v.a. Nachweismöglichkeiten fehlen. Dieses Risiko wird im Nachfolgenden als *Verantwortungsrisiko* bezeichnet.

1059 Siehe dazu *Ibold* ZStW 134 (2022), 504.

1060 *Europäische Kommission* 2020b, 12 f.

1061 *Europäische Kommission* 2020b, 13.

1062 Zu diesen Beispielen *Sommerer* 2020b, 102; *Kasperkevic* 2015; *Tatman* 2017, 53. Zu erwähnen ist auch das Programm COMPAS, das in den USA teilweise eingesetzt wird, um die Rückfallwahrscheinlichkeit von Straftätern zu prognostizieren; für schwarze Straftäter wurde generell eine zweimal höhere Rückfallwahrscheinlichkeit prognostiziert, obgleich diese sodann nicht rückfällig wurden (*Angwin/Larson/Mattu u. a.* 2016; krit. dazu *Flores/Bechtel/Lowenkamp* FPJ 80 (2016), 38).

Stellvertretend hierzu:

> „Die besonderen Merkmale vieler KI-Technologien wie Opazität („Blackbox-Effekt"), Komplexität, Unvorhersehbarkeit und teilautonomes Verhalten können die […] wirksame Durchsetzung von EU-Rechtsvorschriften zum Schutz der Grundrechte erschweren. […] Treten Sicherheitsrisiken tatsächlich auf, ist es […] schwierig, potenziell problematische Entscheidungen, die unter Einbeziehung von KI-Systemen getroffen wurden, zurückzuverfolgen. […] Im Falle KI-gestützter Systeme […] kann es […] schwierig sein, einen Produktfehler, den entstandenen Schaden und den Kausalzusammenhang zwischen diesen beiden nachzuweisen."[1063]
>
> „Hinzu kommt, dass selbst technisch einfache algorithmische Systeme oftmals in komplexe sozioinformatische Ökosysteme eingebunden sind, d.h. informations- und arbeitsteilige Prozesse, in denen eine Vielzahl von Herstellern und Betreibern mitwirkt. […].[1064]

Die *KI-spezifischen Eigenschaften* (technische Autonomie und epistemische Opazität) sollen also nicht nur Einfluss auf das technische Risiko haben, sondern auch auf die Möglichkeit, für die Realisierung dieses Risikos jemandem – im hiesigen Kontext – strafrechtliche Produktverantwortung zuzurechnen. Dieses Verantwortungsrisiko soll zudem von dem Fertigungsprozess beeinflusst sein, innerhalb dessen KI-Systeme entwickelt und hergestellt werden, einem „komplexen sozioinformatorischen Ökosystem" mit einer Vielzahl verschiedener arbeitsteiliger Prozesse und beteiligten Personen. Um dieses Verantwortungsrisiko bewerten zu können, insbes. ob es nicht auch schon im Kontext der strafrechtlichen Produktverantwortung für „herkömmliche" Produkte relevant wird, bedarf es im Rahmen des Realbereichs der strafrechtlichen Produktverantwortung einer Erörterung des Fertigungsprozesses von KI-Systemen. Eine abschließende Bewertung ist schließlich nur nach Analyse der Strafbarkeitsvoraussetzungen der strafrechtlichen Produktverantwortung möglich.

Für den Untersuchungsgegenstand der strafrechtlichen Produktverantwortung sind nur das Sicherheitsrisiko und das Verantwortungsrisiko relevant; KI-Risiken für andere Rechtsgüter als Leben und körperliche Unversehrtheit scheiden aus, da die §§ 222, 229 StGB nur diese beiden Rechtsgüter schützen und entsprechend nur Handlungen unter Strafe stellen, die einen solchen Rechtsgüterbezug aufweisen.[1065] Das Verantwortungsrisiko findet ergänzend zum Sicherheitsrisiko Berücksichtigung, weil dieses jedenfalls einen mittelbaren Bezug zu diesen Rechtsgütern aufweist und Aussagen über die Wirksamkeit der strafrechtlichen Produktverantwortung ermöglicht.

1063 *Europäische Kommission* 2020b, 14 f.
1064 *Datenethikkommission* 2019, 169, 171.
1065 Näher dazu unten Kap. 3, E.II.3.b.cc.

4. KI-Technik und Chance

Risikobehaftete Entscheidungen werden nicht um ihrer selbst willen, sondern wegen der damit verbundenen Chancen getroffen.[1066] Chancen sind das Spiegelbild des Risikos – sie beschreiben mögliche Folgen in der Zukunft, die nicht als negativ, sondern als positiv bewertet werden. Dies gilt auch für technisches Handeln und insbes. für KI-technisches Handeln. Wenngleich technisches Handeln stets mit dem Risiko verbunden ist, dass der Steuerungserfolg versagen kann, und wenngleich bei KI-technischem Handeln die Risikoprognose gegenüber klassischer Technik erschwert ist, werden diese Risiken um der damit verbundenen Chancen willen eingegangen:

Technisches Handeln – die Entwicklung und Herstellung sowie die Nutzung technischer Produkte führt ganz allgemein zu einer Expansion menschlicher Handlungsmacht; es verstärkt, entlastet oder ersetzt den natürlichen Mitteleinsatz.[1067] Gleiches gilt für KI-Technik: Sie ermöglicht einen Technikeinsatz in Bereichen, die bisher ausschließlich auf natürlichen Mitteleinsatz angewiesen waren, bzw. führt dazu, dass in bereits technisierten Bereichen der Mensch verstärkt entlastet wird. Dazu folgende Beispiele:

Bei allen technischen Anwendungen in Bezug auf Sprache und Bilder hat erst der Einsatz des maschinellen Lernens zu wirklichen Erfolgen geführt. Konventionelle Programmiertechniken, die ein explizites Wissen und explizite Regeln über (handgeschriebene und gesprochene) Sprache voraussetzen, sind zuvor an der Komplexität von Sprache gescheitert. Inzwischen ist aber mit Hilfe von KI die automatische Erkennung von geschriebener wie gesprochener Sprache weit fortgeschritten; die Regeln, wie Sprache ausgesprochen und mit der Hand geschrieben wird, lernt der Computer auf der Grundlage von Daten selbst.[1068] Zudem ermöglichen inzwischen Anwendungen wie GPT-4, ChatGPT bzw. Dall-E (sog. generative KI) sogar das automatische Generieren von Texten bzw. Bildern, deren Inhalt von den Nutzern durch „prompts“ nur grob vorgegeben wird.[1069]

Im Bereich Mobilität kann durch KI eine bereits bestehende Entlastung des Menschen verstärkt werden: Der Einsatz von KI in Fahrzeugen (autonome Fahrzeuge) ermöglicht eine Fortbewegung, ohne dass der Nutzer jederzeit mit der Steuerung des Fortbewegungsmittels beschäftigt ist, und er sich vom Verkehrsgeschehen abwenden kann; die Eliminierung menschlichen Versagens soll sogar

1066 Siehe dazu oben Kap. 1, B.III.1.c.
1067 Dazu oben Kap. 1, B.III.2.b.
1068 Siehe oben Kap. 2, A.II.
1069 Zu den Problemen konventioneller Programmiertechniken und speziell zur Sprach- und Bilderkennung siehe oben Kap. 2, A.II.2.; zu GPT-4, ChatGPT und DALL-E siehe die Anbieterseite: https://openai.com/product.

die Unfallwahrscheinlichkeit reduzieren.[1070] Die Vielzahl an möglichen Verkehrssituationen, die ein autonom fahrendes Fahrzeug auf welche Weise bewältigen muss, kann schwerlich vorhergesehen und in explizite Regeln übersetzt werden; also kommt auch dort maschinelles Lernen zum Einsatz.

In der Forschung kann KI den Prozess der wissenschaftlichen Erkenntnis unterstützen, indem KI schneller als der Mensch bestimmte Muster aufdecken kann. Dazu gehört bspw. die „drug discovery", also die Suche nach neuen Medikamenten. Diese Suche erfolgt regelmäßig auf der Grundlage von Laborexperten; KI kann diese Suche nun beschleunigen, indem auf der Grundlage von Daten über bekannte Wirkstoffe Vorhersagen getroffen werden, welche Moleküle in welcher Kombination positive Auswirkungen auf den menschlichen Körper haben könnten.[1071] Ein weiteres Beispiel ist die Anwendung AlphaFold, welche Proteinstrukturen auf der Grundlage der Aminosäuresequenz des Proteins vorhersagen kann; davon verspricht sich die Forschung ein erleichtertes Verständnis über körperliche Prozesse und Krankheiten sowie die beschleunigte Entwicklung von Medikamenten.[1072] In der angewandten Medizin soll KI zuletzt in der radiologischen Diagnostik zum Einsatz kommen und schneller und präziser als der Mensch – so jedenfalls das Ziel – Krankheiten erkennen bzw. ausschließen können.

Zuletzt ist bei der Betrachtung der Chancen von KI auch noch ihr Charakter als „enabling technology" zu beachten: Sie ist eine Form von Software-Technik, die für verschiedene Arten von Produkten zum Einsatz kommen kann und dort verbesserte oder bisher nicht mögliche Anwendungs- und Nutzungsmöglichkeiten eröffnet.[1073] Besondere Chancen ergeben sich auch aus Foundation Models, die sich durch ihre Fähigkeit zum Transferlernen (Wissen kann von einer Aufgabe auf die nächste übertragen werden) und ihren enormen Umfang (viel Speicher, viel Daten) auszeichnen. Auf der Grundlage von Foundation Models können auch kleinere Unternehmen ohne entsprechendes Spezialwissen KI-Anwendungen entwickeln; der Vorteil von Foundation Models liegt also v.a. darin, dass der potentielle wirtschaftliche oder wissenschaftliche Nutzen die Kosten für Entwicklung im Vergleich zu regulären Anwendungen bei Weitem übersteigt.[1074]

1070 Siehe zur prognostizierten Unfallwahrscheinlichkeit bei autonomen Fahrzeugen oben Fn. 1053. Zu diesem Aspekt auch *Wigger* 2020, 64 f.
1071 *Bommasani/Hudson/Adeli u. a.* 2021, 56.
1072 Zu AlphaFold oben Kap. 2, A.II.2. Zudem *Grolle* SPIEGEL 40 (4.10.2022).
1073 *Brühl* SZ 5.5.2022.
1074 *Bommasani/Hudson/Adeli u. a.* 2021, 149 ff.

VI. Rückblick

KI beschreibt eine transklassische, aber nicht auch eine „naturalisierte" Technik, die dem Menschen als „zweite Natur" gegenübertreten würde. Denn KI-Systeme sind weiterhin innerlich steuerbar, weil der menschliche Einfluss auf den technischen Output bei der Entwicklung weiterhin hoch ist und der Nutzer über den Einsatz eines KI-Systems entscheidet. Schadenspotentiale einer KI sind Risiken und keine Gefahren gleich dem Wirken der Natur. Der Charakter des technischen Risikos verändert sich nicht grundlegend: KI schafft kein zusätzliches *Sicherheitsrisiko* und auch die Schadenswahrscheinlichkeit und Schadenshöhe kann bei KI-Systemen nicht als pauschal erhöht angesehen werden. Veränderungen ergeben sich allerdings bei der Risikoprognose; diese ist bei KI-Systemen gegenüber klassischer Technik nicht nur vorübergehend in der Innovationsphase, sondern wegen der Opazität von KI-Systemen dauerhaft erschwert.

Im Zusammenhang mit KI wird vermehrt ein hier als *Verantwortungsrisiko* bezeichnetes Risiko betont; dies umbeschreibt den Umstand, dass bei der Verwirklichung des technischen Risikos retrospektiv möglicherweise keiner dafür in Verantwortung genommen werden kann. Dieses Risiko bedarf noch näherer Analyse i.R.d. Realbereichs der strafrechtlichen Produktverantwortung.

Dem KI-technischen Risiko stehen zuletzt KI-spezifische Chancen gegenüber: KI ermöglicht einen Technikeinsatz in Bereichen, die bisher ausschließlich auf natürlichen Mitteleinsatz angewiesen waren, bzw. führt dazu, dass in bereits technisierten Bereichen der Mensch verstärkt entlastet wird.