HEIDELBERGER
JAHRBÜCHER
ONLINE
Band 6 (2021)

Gesellschaft der Freunde
Universität Heidelberg e.V.

Intelligenz: Theoretische Grundlagen und praktische Anwendungen

Rainer M. Holm-Hadulla, Joachim Funke & Michael Wink (Hrsg.)

HEIDELBERG
UNIVERSITY PUBLISHING

Künstliche Intelligenz und Algorithmen – Wahrer Fortschritt oder doch nur digitale Alchemie?

VINCENT HEUVELINE & VIOLA STIEFEL

IWR und Rechenzentrum, Universität Heidelberg

Zusammenfassung

Vom Schachcomputer über das selbstfahrende Auto bis hin zu den großen Science-Fiction-Erfolgen im medialen Bereich – KI ist heutzutage omnipräsent. Ausgehend von der Frage nach der Interaktion zwischen Mensch und Maschine beschäftigt sich dieser Beitrag zunächst mit der Unterscheidung zwischen starker KI, die primär in der cineastischen Welt beheimatet ist, und schwacher KI, unter die heute alle tatsächlichen KI-Systeme fallen. Der Aspekt des Lernens und die Rolle von Algorithmen hierbei sind von eminenter Bedeutung für die Erforschung und Weiterentwicklung der bis dato vorhandenen KI-Systeme. Auf Basis künstlicher neuronaler Netze lernen Computer beispielsweise, Katzen- von Hundebildern zu unterscheiden. Doch können der KI auch gewichtigere Entscheidungen übergeben werden? Und wie trifft der Algorithmus mathematisch betrachtet eine Entscheidung? Was passiert, wenn die Daten, mit denen der Computer lernt, fehlerbehaftet sind? Die Konsequenzen aus diesen Überlegungen eröffnen zweifelsohne eine nicht ausschließlich für die Forschung, sondern für die gesamte Gesellschaft relevante Bandbreite von neuen Themenstellungen, die mit dem zunehmenden Einsatz von KI immer zentraler werden.

https://doi.org/10.17885/heiup.hdjbo.2021.1.24390

1 Einleitung

Nicht selten wird die Frage aufgeworfen, wie es sein kann, dass ein Computer neue Erkenntnisse generiert, die der Programmierer gar nicht intendiert hat und die ihm völlig unbekannt sind. Man sollte eigentlich davon ausgehen können, dass der Computer sklavisch die Abfolge von binären Befehlen und Anweisungen ausführt, die der Programmierer mittels seiner Software-Entwicklung definiert hat. Der Computer macht eben keinen Fehler. Diese entstehen höchstens, wenn der Software-Entwickler ein paar von den berühmten Bugs in seiner Software übersehen hat und dadurch ein irrtümliches oder gar erratisches Verhalten der Maschine verursacht. In diesem Duktus sind die Verhältnisse zwischen Mensch und Maschine klar definiert: der Mensch diktiert die Aufgaben, die die Maschine umzusetzen hat. Der Computer, als ausführendes Instrument, wägt seine Reaktion dabei nicht kritisch ab und zeigt keine Gefühle wie etwa Langweile, auch wenn die auszuführenden Aufgaben aus repetitiven, drögen Schritten bestehen. Es wird erwartet, dass die Aufgaben deterministisch und reproduzierbar umgesetzt werden. Die Stromzufuhr sorgt dafür, dass die Bytes und Bits stets geordnet und nach Plan innerhalb der elektronischen Schaltungen fließen. In diesem Kontext gibt es wenig Platz in der Maschinen-Rolle für Unerwartetes oder gar Kreativität. Der Mensch meint somit, stets die vollständige Kontrolle über die Maschine zu besitzen. Diese weitverbreitete Sicht ist jedoch sehr trügerisch. Zum Beispiel kann ein Navigationssystem den kürzesten Weg zwischen der Stadt Heidelberg und dem wunderbaren, mittelalterlichen Ort Bad Wimpfen errechnen. Wir können jedoch nicht davon ausgehen, dass der Entwickler des Navigationssystems alle Geheimtipps der erkundeten Gegend – in unserem Fall Heidelberg – kennt und dafür Wege und Pläne dezidiert entwirft. Vielmehr wird der Programmierer ein Verfahren – allgemein im Sinne eines mathematischen Algorithmus – implementieren, das in der Lage ist, den kürzesten Weg zwischen zwei Punkten auf einer Karte zu berechnen. Die Frage, die sich hier stellt, ist, ob der Programmierer immer noch die Kontrolle über seine Software besitzt. Was bedeutet es, die Kontrolle über einen Algorithmus zu haben? Können unerwartete Ergebnisse entstehen, die zunächst nicht beabsichtigt wurden? Kann ein Algorithmus in der Kombination mit Daten ein Wissen generieren – hier der kürzeste bzw. schnellste Weg –, das der Programmierer in der Form überhaupt nicht kannte? In dieser Abhandlung werden wir uns Schritt für Schritt mit diesen fundamentalen Fragestellungen auseinandersetzen, die zum Verständnis und zur Bewertung von KI von essenzieller

Natur sind. Dabei werden wir feststellen, dass das Verhältnis Mensch/Maschine in diesem Kontext doch subtiler ist als zunächst vermutet.

2 Starke versus schwache KI

Die Liste der Science-Fiction-Autoren und Filmemacher, die sich in ihren Werken der Thematik KI widmen, ist ausgesprochen lang und vielfältig. Eine beinahe Konstante in diesen Büchern und Filmen ist, dass die dargestellte KI in quasi allen Bereichen mindestens den menschlichen Fähigkeiten entspricht, wenn sie sie nicht gar übertrifft. Das überdurchschnittliche logische Denkvermögen von Commander Data aus der Science-Fiction-Reihe Star Trek macht ihn, in Verbindung mit seinem lückenlosen enzyklopädischen Wissen im Sinne von Big Data, zu einer faszinierenden Figur, die auf beinahe jedem Gebiet die menschliche Intelligenz übertrifft. Die Fähigkeit zur Kommunikation in allen natürlichen bekannten Sprachen, die der humanoiden Figur C3PO aus dem Film Star Wars zu eigen ist, ist nicht weniger beeindruckend. Dem Menschen mag diese kognitive Überlegenheit zumindest respekteinflößend, teilweise gar angsteinflößend vorkommen. Stephen Hawking beispielsweise hat immer vor den Gefahren gewarnt, die für die Menschheit von künstlicher Intelligenz ausgehen. Die allgegenwärtige mediale Darstellung von KI trägt nur bedingt zur Beruhigung und Entschärfung bei. James Cameron etwa setzt im zweiten Film seiner bekannten Filmreihe Terminator mit dem gleichnamigen Cyborg weitere Akzente im Zusammenhang mit der künstlichen Intelligenz: Terminator kombiniert die Superlative aller menschlichen Fähigkeiten zur Erreichung eines übergeordneten Ziels: die Menschheit retten. Dabei versteht und beherrscht Terminator sogar eine Fähigkeit, die als ausschließlich menschliches Attribut gilt – den Humor. All diese Figuren haben gemeinsam, dass sie die Ausprägung einer sogenannten starken künstlichen Intelligenz („strong artifical intelligence“) abbilden, die alle Seiten der menschlichen Intelligenz – auch und insbesondere in der Kombination der unterschiedlichen Fähigkeiten – umfasst [3,5].

Das Streben nach übermenschlichen Fähigkeiten, Superkräften und Hyperintelligenz fasziniert die Menschheit seit jeher und schlägt sich aus diesem Grund nicht zuletzt in sämtlichen medialen Formen nieder. Ein Beispiel aus der griechischen Mythologie wäre Ikarus, der schließlich daran scheiterte, zu hoch hinaus zu wollen. Inwieweit den heutigen Bestrebungen der Menschheit, sich auch ansonsten über eine starke KI gottähnlich machen zu wollen (man denke an Hararis Homo Deus),

das gleiche Schicksal droht, wird sich zeigen [4]. Derzeit kann man jedoch festhalten, dass die vorhandene Technologie weit davon entfernt ist, starke künstliche Intelligenz als Realität zu ermöglichen. Die heutigen KI-Systeme fallen unter die Kategorie der schwachen KI („weak artifical intelligence“) [1]: die menschliche Intelligenz bzw. die menschlichen kognitiven Fähigkeiten werden nur in abgegrenzten Teilbereichen erreicht und ggf. übertroffen. Bild- und Spracherkennung, automatisierte Übersetzung und selbstfahrende Autos sind nur ein paar Beispiele, für welche (schwache) künstliche Intelligenz heutzutage produktiv eingesetzt wird.

3 Schwache KI ist Mathematik

Die Dartmouth-Konferenz („Dartmouth Summer Research Project on Artifical Intelligence“), die im Jahr 1956 am Dartmouth College in Hanover, New Hampshire (USA) stattfand, gilt als Beginn der Beschäftigung mit künstlicher Intelligenz im Sinne derjenigen Konzepte und Ansätze, die wir heute verwenden.

Der Name „Artifical Intelligence/künstliche Intelligenz“ stammt von dem Initiator der Konferenz, John McCarthy. Im Rahmen dieser Konferenz sind auch Marvin Minsky, Claude Shannon, John von Neumann und Ray Solomonoff zu nennen, die die weiteren Entwicklungen der KI sehr stark geprägt haben. Die genaue Betrachtung sowohl der behandelten Themen als auch der vertretenen Expertise macht sehr deutlich, dass die zugrundeliegenden KI-Konzepte auf mathematischen Abstraktionen fußen [6,7]. Im Kontext dieser Aufbruchstimmung erscheint das Schlusswort der Konferenz einerseits vielversprechend, auf der anderen Seite jedoch auch noch vorsichtig unverbindlich: „[. . .] 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“ [32]. Eine für die letzten Dekaden wichtige Erkenntnis aus dieser Schlussfolgerung und aus der gesamten Konferenz ist, dass ein Computer mehr als nur die Kernaufgabe des wissenschaftlichen Rechnens – die klassische numerische Simulation – ausführen kann. Ein Computer ist tatsächlich in der Lage, mithilfe entsprechender Algorithmen zu lernen. Die Erfahrung zeigt jedoch, dass dies allgemein nur gelingen kann, wenn man über erhebliche Rechenleistung und Trainingsdaten verfügt [14]. Die letzten Jahrzehnte haben darüber hinaus gezeigt, dass die Entwicklungen der Computer-Technologie als Katalysator für die KI-Entwicklung fungiert haben.

Das Mooresche Gesetz (Moore's Law), das eigentlich eher eine Beobachtung ist, besagt, dass alle 18 Monate eine Verdoppelung der vorhandenen Rechenleistung stattfindet. Diese Faustregel, die bis heute nicht widerlegt worden ist, entspricht einem exponentiellen Wachstum. Derartig gewaltige, technologische Entwicklungen sind ohne Zweifel ausschlaggebend für die längst gefeierten Erfolge von KI beim Schach- und Go-Spiel. Supercomputer und Hochleistungsrechner, die für KI sehr stark eingesetzt werden, sind inzwischen Symbole der größtmöglichen verfügbaren Rechenleistung geworden. Die Anzahl an Gleitkommaoperationen pro Sekunde – kurz FLOPS (Floating Point Operations Per Second) – ist ein Maß für die Leistungsfähigkeit eines Rechners. Heutzutage sind die schnellsten Supercomputer in der Lage, ca. 500 PetaFLOPS (Peta=10^{15}) zu berechnen. Dies entspricht in etwa

500 000 000 000 000 000 Operationen pro Sekunde.

Auch für Experten sind solche Größenordnungen herausfordernd und schwer zu fassen. Diese geballte Rechenleistung ist zwar eine notwendige, jedoch nicht hinreichende Bedingung zum erfolgreichen Einsatz von KI. Die in KI notwendigen Lernverfahren basieren auf Trainingsdaten, auf Grundlage derer der Rechner in der Lage ist, Sachverhalte zu lernen. Allgemein gilt: Je präziser und umfangreicher diese Daten vorliegen, desto besser kann der Rechner lernen. Ein in Lehrbüchern klassisches Beispiel dafür ist das Unterscheiden zwischen Katzen und Hunden auf Basis von Bildern. Der Rechner wird dahingehend trainiert, dass er mithilfe einer großen Anzahl an Bildern von Katzen und Hunden „eingelernt" wird. Beim Präsentieren eines für den Rechner komplett neuen Bildes ist er dann in der Lage zu unterscheiden, ob es sich um eine Katze oder eben einen Hund handelt. Der Clou dabei: Es ist irrelevant, ob er genau die vorgestellte Rasse des Hundes oder der Katze kennt. Wie ein Mensch hat der Computer mithilfe der Trainingsdaten Merkmale gelernt, anhand derer er die Tiere – fast immer richtig – unterscheiden kann. Dabei ist es wesentlich, dass die Unterscheidungsmerkmale nicht explizit vorab programmiert worden sind, sondern automatisch über die entsprechenden KI-Algorithmen durch den Rechner definiert werden.

Die Erfahrung zeigt, dass die Qualität und die Menge der Trainingsdaten dabei ausschlaggebend sind. Hier spielen die technologischen IT-Entwicklungen eine entscheidende Rolle. In einer Ära der allgegenwärtigen digitalen Kommunikation, aber auch der Vernetzung z. B. über das Internet of Things (IoT), werden digitale Daten in etlichen Konstellationen produziert. Der Begriff „Big Data" ist

diesbezüglich selbstredend: in etlichen Anwendungsbereichen werden erhebliche Mengen an strukturierten und unstrukturierten Daten produziert. Dabei sind nicht die Daten als solche wichtig, sondern die Möglichkeit, aus diesen Daten ein Wissen und Erkenntnisse zu generieren. Viele KI-Systeme leben von dem Vorhandensein solcher Daten und sind somit in der Lage, aus diesen wichtige und neue Erkenntnisse zu gewinnen [8]. In diesem Zusammenhang ist es bedeutsam, dass die Datenmenge teilweise so umfangreich geworden ist, dass ein einzelner Mensch nicht in der Lage wäre, diese Daten ohne Rechner-Unterstützung - teilweise mit KI-basierten Ansätzen - zu analysieren. Verliert der Mensch somit die Kontrolle über seine Entscheidung? Wie zuverlässig sind die KI-Systeme tatsächlich? Eine Antwort auf diese Fragen erfordert einen genaueren Blick in die Funktionsweise von KI-Algorithmen.

4 Algorithmen für KI

Der Einsatz von KI setzt voraus, dass der verwendete Computer die Fähigkeit des Lernens besitzt. Ein erster - jedoch irrtümlicher - Gedanke wäre, die Verantwortung des expliziten Lernens an den Programmierer des Computers zu übertragen. Der Programmierer als „Chef-Pädagoge“ der Maschine kann über die entwickelte Software Regeln und Handlungsanweisungen vorgeben, nach deren Muster die Maschine in vorgegebenen Situationen zu reagieren hat. Ein wesentlicher Nachteil an einer solchen Vorgehensweise wäre, dass der Programmierer selbst über das benötigte Wissen verfügen müsste, um dieses in eine entsprechende Software übertragen zu können. Der Programmierer des Go-Spiels müsste demzufolge alle Tipps und Tricks des Go-Spiels kennen und dabei explizit implementieren. Konzeptionell wäre dadurch das vorhandene Wissen der Maschine stets an die Fähigkeiten des Programmierer-Teams gebunden und dadurch de facto erheblich beschränkt. Eine solche Maschine wäre kaum in der Lage, die weltbesten Go-Spieler zu besiegen. Die Erfolge der Software Alpha-Go gegen die Go-Champions dieser Welt haben jedoch einen Gegenbeweis dafür geliefert. Die Maschine konnte mit deutlicher Überlegenheit das alte chinesische Spiel Go stets gewinnen. Wie ist das überhaupt möglich?

Der eigentliche Trick besteht darin, der Maschine beizubringen, wie sie selbst lernen kann, quasi nach dem Prinzip Hilfe zur Selbsthilfe. Der Programmierer bleibt zwar weiterhin „Chef-Pädagoge“ der Maschine. Dabei versucht er jedoch

nicht, die Inhalte, die er allgemein weder kennt noch beherrscht, der Maschine beizubringen. Vielmehr übermittelt er eine Methodik, wie die Maschine selbst lernen kann. Solche Ansätze fallen unter den Begriff „Maschinelles Lernen“: Aus digitalen Daten soll künstlich ein Wissen generiert werden, mit dem die Maschine in der Lage ist, eigenständig Entscheidungen treffen zu können [10]. Die praktische Umsetzung solcher Ansätze erfolgt mithilfe von eigens dafür entwickelten Algorithmen. Diese Algorithmen lassen sich grob in zwei Kategorien einteilen: überwachtes Lernen (engl. supervised learning) und unüberwachtes Lernen (engl. unsupervised learning) [11,13]. Die genaue Beschreibung dieser unterschiedlichen Gruppen würde den Rahmen dieser Abhandlung sprengen. Wir werden uns entsprechend im Folgenden ausschließlich auf die Algorithmen der Kategorie des überwachten Lernens konzentrieren.

Im überwachten Lernen lernt der Computer aus gegebenen Paaren von Ein- und Ausgaben. Beispielsweise wird ein Bild mit einer Katze (Eingabe) mit dem Wert 0 (Ausgabe) verknüpft und ein Bild mit einem Hund (Eingabe) mit dem Wert 1 (Ausgabe) versehen. Die Kunst besteht darin, eine Abbildung bzw. eine Funktion zwischen den Eingaben und Ausgaben so zu definieren, dass sogar unbekannte Bilder richtig nach dem genannten Prinzip klassifiziert werden. Für die Definition von solchen Abbildungen hat sich die Verwendung von sog. neuronalen Netzen sehr bewährt [9,12].

In der Biologie beziehen sich neuronale Netze auf Strukturen des Gehirns von Tieren und Menschen. Dabei bilden die Neuronen ein äußerst komplexes Netzwerk – in der menschlichen Großhirnrinde arbeiten 10 Milliarden Neuronen in feinster Abstimmung. Jedes Neuron ist mit ca. 2000 anderen Neuronen verbunden. Die Fähigkeit des Lernens erfolgt über eine Veränderung der Verbindungsstärken der vorhandenen Neuronen. Somit sind die Informationen nicht in einzelnen Neuronen gespeichert, sondern werden durch den gesamten Zustand des neuronalen Netzes mit allen Verbindungsstärken repräsentiert. Im Bereich der KI verwendet man künstliche neuronale Netze, die stark von ihrem biologischen Pendant inspiriert sind [2]. Bei künstlichen neuronalen Netzen geht es allerdings darum, eine Abstraktion im Sinne einer Modellbildung zu erlangen, anhand derer die Abbildung zwischen Eingaben und Ausgaben bestmöglich definiert werden kann. Der Lernprozess auf Basis von solchen künstlichen neuronalen Netzen besteht darin, Gewichte entlang der Verbindungen (Kanten) eines Graphens zu bestimmen, für welche die Neuronen als Knoten fungieren. Die Paare von Eingaben/Ausgaben

als Trainingsdaten werden dafür verwendet, diese Gewichte festzulegen. Mathematisch betrachtet handelt es sich hierbei um eine Modellkalibrierung im Sinne einer Parameteridentifikation. Dabei sind die Gewichte entlang der Kanten des neuronalen Netzes die Parameter, die identifiziert werden sollen: der Lernprozess als Problem der Parameteridentifikation.

Es soll nicht unerwähnt bleiben, dass aus Sicht der Mathematik noch etliche Fragen rund um die Eigenschaften solcher neuronaler Netze offen sind. Zum Beispiel ist die Festlegung der Dimensionierung eines solchen neuronalen Netzwerkes für eine vorgegebene Anwendung nach wie vor eine Herausforderung, die allgemein empirisch über zahlreiche Tests definiert werden muss. Eine gewisse digitale Alchemie ist hier immer wieder notwendig. Für einfache neuronale Netze kann man beweisen, dass die zugrundeliegende Methodik bekannten Verfahren aus dem Bereich der numerischen Optimierung entspricht. Somit hat man für solche Verfahren die ausgesprochen wichtige Stütze der Mathematik, die ein Fundament sowohl für das Verständnis als auch für Konvergenz-Aussagen der Verfahren liefert. Für zahlreiche Verfahren, die sich in der Praxis bewährt haben, gibt es leider kaum mathematische Erkenntnisse, warum diese Verfahren funktionieren und ob das tatsächlich immer der Fall ist. Das erklärt, warum diese Technologie immer wieder als Blackbox-Modell bezeichnet wird. Für die Verwendung von KI in kritischen Bereichen kann eine solche Lage manche Gefahren bergen. Wie empfindlich reagiert das neuronale Netzwerk z. B. auf fehlerhafte Daten? Hier ist nach wie vor noch ein sehr großer Forschungsbedarf vorhanden, damit solche Ansätze nicht allein durch die Empirie bestätigt werden.

5 KI als Blackbox

In spezifischen Teilbereichen übertrifft KI bereits die menschlichen kognitiven Fähigkeiten. Es werden Mengen an Trainingsdaten verarbeitet, die ein menschliches Gehirn in einem ganzen Leben weder speichern noch verarbeiten könnte. So beeindruckend solche Ergebnisse auch sind, es stellt sich doch die Frage, ob wichtige oder gar kritische Entscheidungen auf Basis solcher Ergebnisse überhaupt getroffen werden dürfen.

Kann man bei wichtigen Entscheidungen der KI-Technologie grundsätzlich vertrauen?

Diese Frage, die in solcher Deutlichkeit aufgrund der gefeierten Erfolge von KI oft vermieden wird, ist eigentlich von signifikanter Bedeutung. Die lapidare Aussage, dass Rechner keinen Fehler machen, gilt in diesen Teilbereichen eben nicht mehr. Für viele Anwendungen weiß niemand – nicht einmal der Programmierer, der die KI-Software geschrieben hat – wie der Algorithmus seine Entscheidung überhaupt getroffen hat. Dieses Phänomen wird als Blackbox-Problem bezeichnet [15]. In der Praxis scheinen die modernen Lernalgorithmen größtenteils zu funktionieren. Fakt ist jedoch, dass man diese Mechanismen, die seitens der KI zu einer Entscheidung führen, oft schlichtweg nicht versteht. In der Interaktion Mensch/Maschine handelt es sich hier sicherlich um einen in dieser Form nie da gewesenen Paradigmenwechsel. Auch gesellschaftlich betrachtet wirft diese Herausforderung weitere Fragestellungen auf:

- Wer besitzt und versteht die Trainingsalgorithmen bzw. -software?
- Wer besitzt und versteht die trainierten neuronalen Netze bzw. KI-Modelle?

Wohlwissend, dass die KI-Technologie etliche Bereiche des täglichen Lebens erreicht hat, sind solche Schwerpunkte nicht nur gesellschaftlich relevant, sondern auch von politischer Bedeutung [19,20].

In der Annahme, dass die KI-Algorithmen in einer vorgegebenen Metrik optimal lernen, was wir – Stand heute – mathematisch nicht beweisen können, stellt sich weiterhin die Frage, ob die Trainingsdaten für die anvisierten Entscheidungen überhaupt geeignet sind. In diesem Zusammenhang sind mehrere Aspekte zu berücksichtigen. In vielen Anwendungsbereichen entstehen die Daten aus Messungen, die allgemein nicht exakt bestimmt werden können. Messfehler sind für Sensoren an der Tagesordnung. Hieraus ergibt sich nun die Frage, wie ein KI-System sowohl auf Trainings- als auch Eingabedaten reagiert, die nicht ganz fehlerfrei sein können. Die Thematik der Sensitivität von solchen Systemen im Hinblick auf Unschärfen in den Daten ist nach wie vor Forschungsgegenstand und allgemein bis heute nicht richtig durchdrungen [16,17].

Ein weiterer Aspekt ist womöglich noch schwerwiegender: Was passiert, wenn die Trainingsdaten unvollständig sind und das KI-System nur partiell den Datenraum erlernen kann? Die Gefahr der vorprogrammierten Diskriminierung lauert genau an dieser Stelle. Die Zeitschrift Focus vom 12.10.2018 [31] brachte diesen Sachverhalt am Beispiel von KI-basierten Bewerbungsbewertungen auf den

Punkt: „Künstliche Intelligenz erachtet Bewerbungen von Frauen als minderwertig“. Nach (menschlicher) Analyse des gesamten Prozesses wurde festgestellt, dass die Trainingsdaten überwiegend von männlichen Personen stammten. Somit nahm das KI-System eine Bewertung vor, die eher aus Unkenntnis heraus entstand als aus einer objektiven Betrachtung. Es handelt sich hier leider nicht um eine Randerscheinung, sondern um eine Herausforderung, die stets beleuchtet werden muss: Eine konstante und transparente Prüfung von KI-Systemen im Hinblick auf mögliche Diskriminierung/Bias ist sicherlich eine zentrale Aufgabe nicht nur der Wissenschaft, sondern auch aller Kernakteure der Gesellschaft [18,22,21,23].

6 Interpretierbare KI als möglicher Lösungsweg

Mit zunehmendem Einsatz von KI ist die Frage nach der Interpretierbarkeit und Erklärbarkeit von KI-Entscheidungen essenziell geworden. Im Englischen beschreibt der Ausdruck der „Explainable Artificial Intelligence (XAI)“ den Bereich, der sich die Erklärbarkeit von künstlicher Intelligenz zum Ziel gesetzt hat [26,27]. Dabei geht es darum, zu verstehen, wie und warum Entscheidungen von KI-Systemen getroffen worden sind. Der Blackbox-Charakter von vielen KI-Systemen soll somit aufgebrochen werden [28]. Wissenschaftlich stellen solche Fragestellungen nach wie vor eine große Herausforderung dar. Bei mehrschichtigen Deep-Learning-Modellen beispielsweise können diese Aspekte auf Basis der heutigen wissenschaftlichen Erkenntnisse nicht beantwortet werden. In der letzten Dekade sind innovative Konzepte entstanden, die in diesem Kontext neue Perspektiven eröffnen. Dabei wird zwischen Ante-Hoc- und Post-Hoc-Ansätzen unterschieden [29,30,25]. Die Ante-Hoc-Methodik konzentriert sich auf Modelle, die per se und a priori – d. h. vorher – interpretierbar sind. Der Post-Hoc-Ansatz untersucht, inwieweit Blackbox-Modelle a posteriori interpretierbar analysiert werden können. Diese Themen sind in vielen Anwendungsbereichen noch Forschungsgegenstand.

In der zwischenmenschlichen Interaktion können wir nicht immer erklären, warum Mitmenschen die eine oder andere Entscheidung treffen. Unser Vertrauen, dass eine Entscheidung richtig ist oder nicht, basiert auf einer Vielfalt an Faktoren, die wir bereits in der Kindheit im Austausch mit Mitmenschen erlernt haben. Das Vertrauen in einen Algorithmus, den man ggf. nicht richtig versteht, ragt in diesem Zusammenhang natürlich sehr weit aus dem üblichen menschlichen Erfahrungsschatz heraus. Somit lässt sich das Spannungsfeld zwischen einem

wahren Fortschritt und einer digitalen Alchemie, dem die KI unterliegt, nur dann auflösen, wenn mutige, transparente und innovative Wege weiterhin beschritten werden.

Referenzen

[1] Walch K. Rethinking Weak vs. Strong AI. 2019. https://www.forbes.com/sites/cognitiveworld/2019/10/04/rethinking-weak-vs-strong-ai/.

[2] Nikolic, D. Why deep neural nets cannot ever match biological intelligence and what to do about it?, International Journal of Automation and Computing, volume 14, pages 532–541, 2017.

[3] Flowers J. Strong and Weak AI – Deweyan Considerations, AAAI Spring Symposium, 2019.

[4] Fjelland R. Why generalized artificial intelligence will not be realized, Humanities and Social Sciences Communications, 7:10 2020.

[5] Liu B. Weak AI is likely to never become Strong AI, so what is its greatest value for us?, Computer Science, 2021.

[6] Shaffi. AI and Mathematics. 2020. https://medium.com/swlh/ai-mathematics-699a9ea2a0d6

[7] Garrido A. Mathematics and AI, two branches of the same tree, Procedia – Social and Behavorial Sciences, Vol. 2, Issue 2, 2010.

[8] Sun Z, Wang P.P. A Mathematical Foundation of Big Data, New Mathematics and Natural Computation, Vol. 13, No. 02, 2017.

[9] Saxton D, Grefenstette E, Hill F, Kohli P. Analysing Mathematical Reasoning Abilities of Neural Models, ICLR conference, 2019.

[10] Thesing L, Autun V, Hansen A.C. What do AI algorithms actually learn? – On false structures in deep learning, arXiv, 2019.

[11] Brownlee J. A Tour of Machine Learning Algorithms. 2019. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[12] Yosinki J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?, Advances in neural information processing systems, 2014.

[13] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv, 2015.

[14] Deisenroth A, Faisal A, Soon Ong C. Mathematics for Machine Learning, Cambridge University Press, 2020.

[15] Bleicher A. Demystifying the Black Box that is AI. 2017. https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

[16] Lim H. 7 Types of Data Bias in Machine Learning. 2020. https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/

[17] Angwin J, Larson J. Machine Bias. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[18] Yapo A, Weiss J.W. Ethical Implications of Bias in Machine Learning, HICSS, 2018.

[19] Ntoutsi E et al. Bias in data-driven artificial intelligence systems - an introductory survey, Wires, 2020.

[20] Mehrabi N et al. A Survey on Bias and Fairness in Machine Learning, arXiv, 2019.

[21] Fu R, Huang Y, Singh P.V. AI and Algorithmic Bias: Source, Detection, Mitigation and Implications, SSRN, 2020.

[22] O´neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threaten Democracy, Crown Edition, 2016.

[23] Baer T. Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists, Apress, 2019.

[24] Li O, Liu H, Chen C, Rudin C. Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions, arXiv, 2017.

[25] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1,206–215, 2019.

[26] Arrieta A.B et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Vol. 58, 2020.

[27] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods, Entropy 23(1), 2021.

[28] Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning – A brief history, state-of-the-art and challenges, PKDD/ECML Workshops, 2020.

[29] Escalante H.J et al. Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018.

[30] Samek W et al. Explainable AI – Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019.

[31] Die Tücken der Intelligenz: Künstliche Intelligenz erachtet Bewerbungen von Frauen als minderwertig – Amazon muss reagieren, Focus, 12.10.2018.

[32] https://250.dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth

Über die Autoren

Vincent Heuveline (geboren in Paris, 1968) ist seit Dezember 2018 Chief Information Officer der Universität Heidelberg. Ebenso ist er Direktor des Rechenzentrums der Universität Heidelberg. Als Professor leitet er das „Engineering Mathematics and Computing Lab“ (EMCL) am Interdisziplinären Zentrum für Wissenschaftliches Rechnen (IWR) der Universität Heidelberg. Darüber hinaus leitet er die Arbeitsgruppe „Data Mining and Uncertainty Quantification“ am Heidelberger Institut für Theoretische Studien (HITS gGmbH). Seine Forschungsinteressen umfassen das Hochleistungs- und datenintensive Rechnen sowie die Softwareentwicklung mit besonderem Fokus auf Anwendungsgebiete in der Medizin. In der Lehre beschäftigt er sich – neben dem wissenschaftlichen Rechnen – im Rahmen dedizierter Vorlesungen und Seminare intensiv mit dem Thema IT-Sicherheit und KI. Prof. Heuveline ist Mitglied in den Programmkomitees zahlreicher internationaler Konferenzen zum Hochleistungsrechnen. Als Experte und Ansprechpartner berät er Vertreter der Industrie zu Themen der Digitalisierung, dem anwendungsbezogenen Einsatz numerischer Simulationen, Big and Smart Data, KI und der IT-Sicherheit im industriellen Umfeld.

Viola Stiefel hat Romanistik (Französisch und Italienisch) und Geschichte an der Universität Heidelberg studiert und wurde 2018 in französischer Literaturwissenschaft promoviert. Seit März 2021 arbeitet sie als Referentin am Universitätsrechenzentrum Heidelberg und beschäftigt sich hier u. a. mit Themen an der Schnittstelle zwischen Geisteswissenschaften, Digitalisierung und KI.

Korrespondenzadresse:
Prof. Dr. Vincent Heuveline
IWR - Engineering Mathematics and Computing Lab (EMCL)
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg

E-Mail: vincent.heuveline@iwr.uni-heidelberg.de
Homepage:
https://emcl.iwr.uni-heidelberg.de/people/heuveline-vincent

Dr. Viola Stiefel
Universitätsrechenzentrum
Universität Heidelberg
Im Neuenheimer Feld 330
69120 Heidelberg

E-Mail: viola.stiefel@urz.uni-heidelberg.de
Homepage:
https://www.urz.uni-heidelberg.de/en/staff/stiefel-viola