Preprint – die finale und zitierfähige Fassung wird publiziert als:
Buchholz, O. (2025): Maschinelles Lernen in der Wissenschaft. In: Noller, J., Reinhardt, K. (Hrsg.): *Handbuch Philosophie der Digitalität*. Heidelberg: Metzler.

Maschinelles Lernen in der Wissenschaft

Oliver Buchholz[*]

Abstract

Die Technologie des maschinellen Lernens spielt inzwischen eine bedeutende Rolle im Forschungsprozess verschiedener wissenschaftlicher Disziplinen. Das vorliegende Kapitel untersucht diese Rolle des maschinellen Lernens, vor allem hinsichtlich ihrer methodologischen und wissenschaftstheoretischen Implikationen. Ein besonderes Augenmerk liegt dabei auf der Frage, inwiefern eine Kontinuität zwischen der bislang in der empirischen Forschung dominierenden klassischen Statistik und dem maschinellen Lernen besteht.

Keywords: Maschinelles Lernen in der Wissenschaft; Wissenschaftstheorie; Empirische Forschungsmethoden; Prognosemodelle; Opazität; Statistische Lerntheorie

Einleitung

Seit jeher wird der Wissenschaft – zumindest seitens der Wissenschaftstheorie – eine herausragende Rolle bei der methodischen und systematischen Überprüfung von Hypothesen zugeschrieben (vgl. Popper 1935; 1963; Poser 2012). Unter einer hinreichend weiten Wissensdefinition könnte folglich argumentiert werden, dass die Wissenschaft in gewisser Weise einen wichtigen Beitrag zur Überprüfung bestehenden sowie zur Hervorbringung neuen Wissens leistet. Wissenschaftliche Resultate, denen die Eigenschaft neuen Wissens zugeschrieben wird, entstehen dabei vor allem in empirisch arbeitenden Disziplinen durch die Erhebung von Daten und deren Analyse mittels Methoden der beschreibenden oder schließenden Statistik. Der Datenanalyse kommt folglich eine verbindende Funktion zu, die bestenfalls von der Empirie zur Erkenntnis führt. Aufgrund dieser zentralen Stellung im Forschungsprozess waren und sind die Datenanalyse sowie die dazugehörigen Methoden der Statistik stets Gegenstand intensiver wissenschaftstheoretischer Debatten (vgl. Romeijn 2022).

Vor dem Hintergrund wachsender Datenmengen und Rechnerkapazitäten, einer Entwicklung, die neben anderen Gesellschaftsbereichen auch die Wissenschaft erfasst, zeichnet sich derzeit allerdings in vielen Disziplinen ein methodologischer Wandel ab: Methoden der klassischen Statistik werden zunehmend durch Methoden des maschinellen Lernens ergänzt, stellenweise sogar ersetzt. Dieser Wandel wirft eine Reihe dringlicher wissenschaftstheoretischer Fragen auf, allen voran jene nach der Kontinuität zwischen klassischer Statistik und maschinellem Lernen: Inwiefern unterscheiden sich die beiden und handelt es sich demnach um grundsätzlich ähnliche oder verschiedene Paradigmen der Datenanalyse? Eine Antwort auf diese Frage ist von entscheidender Bedeutung, würde sie doch zugleich eine Einschätzung hinsichtlich der Übertragbarkeit wissenschaftstheoretischer Erkenntnisse über die klassische Statistik auf das maschinelle Lernen erlauben – und damit auch eine Charakterisierung der mittels Methoden des maschinellen Lernens gewonnenen Resultate. Denn darum muss es bei der wissenschaftstheoretischen Untersuchung des maschinellen Lernens schlussendlich gehen: Zu

[*] ETH Zürich, Professur für Bioethik, Hottingerstrasse 10, CH-8032 Zürich. E-Mail: oliver.buchholz@hest.ethz.ch.
Die Arbeit an diesem Text wurde durch ein ETH Zürich Postdoctoral Fellowship unterstützt.

klären, ob wissenschaftlichen Resultaten, die durch dessen Einsatz erzielt wurden, der gleiche epistemische Status zukommt wie jenen, die durch den Einsatz klassischer Statistik erzielt wurden; ob die Wissenschaft also auch neues Wissen hervorbringen kann, wenn sie sich im Forschungsprozess auf Methoden des maschinellen Lernens stützt.

Der Diskurs zu diesen Fragestellungen hat im Laufe der letzten Jahre eine beträchtliche Dynamik entwickelt und sich in eine Vielzahl von Teildebatten ausdifferenziert. Ziel dieses Beitrags ist es, einen Überblick über die hierbei zentralen Themen zu geben. Dazu umreiße ich zunächst, wo und in welcher Form maschinelles Lernen aktuell im Forschungsprozess eingesetzt wird. Im Anschluss widme ich mich dem Problem der Opazität des maschinellen Lernens und skizziere abschließend verschiedene Lösungsansätze.

Methodologische Grundlagen und Grundprobleme

Innerhalb weniger Jahre haben Methoden des maschinellen Lernens ihren Weg in die wissenschaftliche Praxis vieler Disziplinen gefunden. So fallen beispielsweise in verschiedenen Teilbereichen der Physik Datenmengen an, deren Umfang eine sinnvolle Analyse durch Forschende praktisch unmöglich macht. Methoden des maschinellen Lernens werden daher dazu verwendet, wissenschaftlich relevante Objekte – von kleinsten Partikeln bis hin zu Exoplaneten – in diesen Datenmengen zu identifizieren (vgl. Baldi *et al.* 2014; Foreman-Mackey *et al.* 2015).

In der Biologie wiederum machte kürzlich eine Anwendung von sich reden, die einen großen Schritt zur Lösung des sogenannten Proteinfaltungsproblems darstellt. Dieses Problem geht auf das sogenannte Anfinsen-Dogma zurück, in welchem der gleichnamige Biochemiker die Vermutung formulierte, dass sich auf Basis der Abfolge von Aminosäuren innerhalb eines Proteins dessen spezifische dreidimensionale Struktur bestimmen lässt. Letztere ist insbesondere deshalb von Interesse, da sie maßgeblich die spezifische Funktion eines Proteins festlegt. Seit der Formulierung des Anfinsen-Dogmas in den 1960er-Jahren drehten sich die Arbeiten zum Proteinfaltungsproblem im Kern darum, die Präzision zu verbessern, mit der sich Proteinstrukturen aus Aminosäuresequenzen vorhersagen lassen. Während lange Zeit lediglich kleine Verbesserungen erzielt wurden, konnten Forschende durch die Anwendung maschinellen Lernens schließlich eine beträchtliche Steigerung der Präzision erzielen. Der von ihnen entwickelte Algorithmus war tatsächlich in der Lage, lediglich auf Basis von Aminosäuresequenzen mit hoher Präzision die dreidimensionale Struktur des jeweils daraus resultierenden Proteins vorherzusagen, was weithin als bahnbrechender Fortschritt in der Arbeit am Proteinfaltungsproblem gewertet wurde (vgl. Jumper *et al.* 2021).

Breite Anwendung finden Methoden des maschinellen Lernens zudem in der Chemie, nicht zuletzt in der sogenannten Retrosynthese. In diesem Verfahren entwerfen Forschende neue Moleküle mit spezifischen, wünschenswerten Eigenschaften und spalten sie dann in immer kleinere Bestandteile auf, um sie auf bereits verfügbare Chemikalien zurückzuführen, aus denen das Wunschmolekül also gewonnen werden kann. Generell ist dieses Verfahren sehr kompliziert und funktioniert in der Theorie weitaus besser als in der praktischen Umsetzung. Methoden des maschinellen Lernens werden allerdings erfolgreich dazu eingesetzt, die theoretisch sinnvollen und praktisch umsetzbaren Pfade von grundlegenden Chemikalien zu neuartigen Molekülen ausfindig zu machen sowie die tatsächliche Herstellung der Letzteren zu steuern (vgl. Coley *et al.* 2019; Segler *et al.* 2018).

Die vielfältigen und zweifelsohne häufig erfolgreichen Anwendungen des maschinellen Lernens in der Wissenschaft werfen Fragen nach dem wechselseitigen Verhältnis der jeweils verwendeten Methoden auf: Teilen sie bestimmte Eigenschaften oder gar das grundlegende Erfolgsrezept? Beides wäre für eine wissenschaftstheoretische Untersuchung des maschinellen Lernens zu hoffen, sofern sie sich nicht in einer Analyse von Einzelfällen erschöpfen, sondern zu

allgemeinen Einsichten gelangen soll. Werfen wir vor dieser Untersuchung also zunächst einen kurzen Blick auf die Grundprinzipien des maschinellen Lernens.

Gemäß der paradigmatischen Definition von Mitchell (1997, S. 2) lassen sich Methoden des maschinellen Lernens als Algorithmen beschreiben, deren Leistungsfähigkeit bei der Erledigung einer Klasse von Aufgaben sich gemessen an einem bestimmten Erfolgsmaß durch Erfahrung verbessert. Insbesondere im Rahmen des sogenannten überwachten („supervised") maschinellen Lernens, auf welches ich mich im Folgenden konzentriere und das sowohl vom unüberwachten („unsupervised") als auch vom verstärkten („reinforcement") maschinellen Lernen abzugrenzen ist, handelt es sich bei der relevanten Klasse von Aufgaben um das Erstellen von Vorhersagen, während Daten die Rolle der Erfahrung zukommt. Letztere haben in der Regel die Struktur eines Datensatzes, in dem jede Zeile eine Beobachtung enthält, welche durch verschiedene, in den Spalten des Datensatzes aufgeführte Merkmale charakterisiert wird. Eine zusätzliche Spalte enthält außerdem für jede Beobachtung ein dazugehöriges Label, welches entweder als diskrete Kategorie („Hund", „Katze") oder reelle Zahl codiert sein kann. Ziel des maschinellen Lernens ist es, einen allgemeinen Zusammenhang zwischen Merkmalen und Labels, welcher häufig mit dem Begriff ‚Muster' umschrieben wird, aus den vorliegenden Daten abzuleiten und diesen für die möglichst präzise Vorhersage der Labels neuer Beobachtungen zu nutzen, über die lediglich deren Merkmale bekannt sind. Um dieses Ziel der Mustererkennung zu erreichen, wird ein aus mehreren Schritten bestehender, sogenannter Lern- oder Trainingsprozess durchgeführt: Der Algorithmus erhält Zugang zu den Merkmalen der vorliegenden Daten und soll die dazugehörigen Labels vorhersagen. Mittels eines zuvor festgelegten Maßes wird im Anschluss die Präzision dieser Vorhersagen evaluiert. Typischerweise wird die Präzision zunächst gering sein, da die vorhergesagten stark von den tatsächlichen Labels abweichen. Der Algorithmus verändert daher seine Spezifikation in der Weise, welche die größte Verbesserung seiner Präzision erwarten lässt. Anschließend beginnt der Prozess von vorn und der Algorithmus erstellt Vorhersagen für die Labels der vorliegenden Daten, nun jedoch basierend auf der veränderten Spezifikation. Sobald durch das Verändern der Spezifikation keine weitere Verbesserung der Präzision erzielt werden kann, endet der Prozess und der Algorithmus wird in der dann vorliegenden Spezifikation zur tatsächlichen Vorhersage der Labels neuer Daten eingesetzt. Diesem Vorgehen liegt die Annahme zugrunde, dass die vorliegenden Daten repräsentativ für die relevante Grundgesamtheit sind und die Spezifikation, welche die besten Ergebnisse auf den ersteren erzielt deshalb auch am wahrscheinlichsten zu den besten Ergebnissen für neue Daten führt (vgl. Buchholz und Raidl 2022).

Bereits dieser kurze Abriss zeigt, dass die Funktionsweise aller Methoden des maschinellen Lernens maßgeblich auf der Nutzung von Daten beruht. Diese Einsicht stellt somit eine erste, die verschiedenen Methoden verbindende Gemeinsamkeit dar. Zudem zeigt sich daran, dass Methoden des maschinellen Lernens eine ausgeprägte statistische Komponente und somit eine Ähnlichkeit mit Methoden der klassischen Statistik aufweisen, deren Funktionsweise ebenfalls auf der Nutzung von Daten beruht. Diese Ähnlichkeit scheint es folglich wie oben erwähnt zu erlauben, bestimmte wissenschaftstheoretische Erkenntnisse über die eine Klasse von Methoden auf die andere Klasse zu übertragen. Eine Reihe methodologischer Probleme der klassischen Statistik sind also offenbar auch Probleme des maschinellen Lernens. Tatsächlich unterstreicht die einschlägige Literatur diesen Umstand, wie sich gut an zwei paradigmatischen Problemen illustrieren lässt.

Zunächst wäre in diesem Zusammenhang das altbekannte Diktum zu nennen, demzufolge bloße *Korrelation* keine *Kausalität* impliziert, ein gemeinsames Auftreten bestimmter Ereignisse also nicht ohne weiteres einen Schluss über deren ursächlichen Zusammenhang erlaubt. Wie oben beschrieben werden Methoden des maschinellen Lernens meist mit dem Ziel präziser Vorhersagen

eingesetzt, welches sie durch die Ableitung allgemeiner Zusammenhänge zwischen Merkmalen und Labels in vorliegenden Daten zu erreichen versuchen. Ihr Fokus liegt also auf der Identifizierung von typischerweise auftretenden Merkmals-Label-Kombinationen beziehungsweise – im Fachjargon – auf der Identifizierung von Korrelationen zwischen bestimmten Merkmalen und Labels. Wie in der klassischen Statistik wird dieser Umstand zum Problem, sobald die Methoden in Bereichen eingesetzt werden, in denen es gar nicht um bloße Korrelationen, sondern um die Identifizierung kausaler Beziehungen geht. Als Beispiel hierfür wäre etwa die Epidemiologie zu nennen, deren Zielsetzung häufig darin besteht, Handlungsempfehlungen zu geben, die bestimmte negative Resultate vermeiden oder positive Resultate hervorbringen – etwa eine Maskenpflicht zur Reduzierung von Infektionen. Für das Erreichen dieser Zielsetzung sind allerdings Kenntnisse der relevanten Kausalbeziehungen notwendig, da nur so geklärt werden kann, welche Merkmale durch individuelles oder kollektives Handeln verändert werden müssten, um ein bestimmtes Resultat zu erzielen. Werden in diesem Kontext nun Methoden wie jene des maschinellen Lernens eingesetzt, die lediglich zur Identifizierung von Korrelationen fähig sind, so können im Allgemeinen keine derartigen Handlungsempfehlungen abgeleitet werden, da der jeweilige Algorithmus nicht zwingend die dafür relevanten Kausalbeziehungen erfasst. Im schlimmsten Fall könnten daraus sogar falsche Handlungsempfehlungen entstehen, wenn der methodologische Fokus des maschinellen Lernens nicht reflektiert und die identifizierten Korrelationen zwischen Merkmalen und Labels fälschlicherweise als Kausalbeziehungen interpretiert werden (vgl. Grote und Buchholz i.E.).

Als weiteres methodologisches Problem, welches aus der Datenbezogenheit des maschinellen Lernens resultiert und daher ebenfalls in der klassischen Statistik auftritt, wäre beispielhaft das sogenannte *Referenzklassenproblem* zu nennen. In seiner Grundform besteht dieses Problem darin, Einzelfällen bedingte Wahrscheinlichkeiten zuzuweisen. Eine solche Zuweisung ist problematisch, da jeder Einzelfall einer Vielzahl unterschiedlicher Klassen zugewiesen werden kann, die bedingten Wahrscheinlichkeiten über unterschiedliche Klassen hinweg unter Umständen jedoch stark variieren können. Es ist also unklar, welche die *richtige* Klasse ist, der ein Einzelfall für die Bestimmung der bedingten Wahrscheinlichkeit zugewiesen werden sollte (vgl. Reichenbach 1935; Venn 1876, S. 194). Eine leicht abgewandelte Form dieses Problems entsteht, wenn auf Basis statistischer Informationen, oder kurz: Daten, Schlussfolgerungen über Einzelfälle gezogen werden sollen. In Situationen dieser Art ist ebenfalls unklar, welche Merkmale innerhalb der vorliegenden Daten zur Erstellung einer individuellen Schlussfolgerung herangezogen werden und welche keine Berücksichtigung finden sollten. Mit ebenjener Unklarheit sind Methoden des maschinellen Lernens jedoch konfrontiert: Im Laufe des Lernprozesses müssen sie zunächst auf Basis vorliegender Daten die für präzise Vorhersagen relevanten Merkmale identifizieren. Ist die finale Spezifikation des Algorithmus bestimmt, wird diese wie oben beschrieben zur Vorhersage neuer Labels eingesetzt. Es findet also ein Schluss von den vorliegenden Daten auf Einzelfälle statt, der unweigerlich mit einer Form des Referenzklassenproblems einhergeht – und wie im Bereich der klassischen Statistik existiert bislang keine allgemein akzeptierte Strategie zur Lösung des Problems (vgl. Buchholz 2023b).

Präzise Vorhersagen als spezifisches Ziel des maschinellen Lernens

Die vorangegangenen Bemerkungen legen den Schluss nahe, dass in der Tat eine ausgeprägte Kontinuität zwischen Statistik und maschinellem Lernen besteht. Diese ergibt sich hauptsächlich aus der Funktionsweise der Methoden, welche in beiden Fällen entscheidend auf der Verwendung von Daten beruht und den Methoden sowie den mit ihnen gewonnenen Resultaten daher einen probabilistischen Charakter verleiht. Jenseits ihrer ähnlichen Funktionsweise lassen sich Statistik und maschinelles Lernen jedoch anhand der jeweils verfolgten Zielsetzung unterscheiden, denn

während diese im Fall des maschinellen Lernens meist im Erstellen möglichst präziser Vorhersagen besteht, versucht die Statistik häufig, bestimmte Hypothesen über eine Grundgesamtheit auf Basis spezifischer wahrscheinlichkeitstheoretischer Annahmen zu überprüfen. Es lassen sich folglich *zwei Kulturen der Datenanalyse* unterscheiden, die – teils sogar unter Verwendung ähnlicher Methoden – zwei verschiedene Ziele verfolgen (vgl. Breiman 2001). Diese Unterscheidung wirft die Frage auf, ob sich daraus methodologische Probleme ergeben, die aufgrund seines besonderen Fokus auf dem Erstellen von Vorhersagen vorrangig das maschinelle Lernen betreffen.

Wie oben beschrieben bringt es dieser Fokus zunächst mit sich, dass Methoden des maschinellen Lernens einen Lernprozess unter Verwendung vorliegender Daten durchlaufen, die eigentlichen Vorhersagen jedoch daran anschließend für neue Daten erstellen. Sie sind folglich in gewisser Weise mit dem bereits bei Hume diskutierten Induktionsproblem konfrontiert. Die Fähigkeit eines Algorithmus zum Erstellen präziser Vorhersagen für neue Daten kann zwar immerhin getestet werden, indem ein Teil der ursprünglich vorliegenden Daten (die sogenannten Testdaten) nicht im Lernprozess, sondern lediglich für einen solchen Test verwendet wird – eine vergleichbare Präzision jenseits der kontrollierten Testbedingungen kann dadurch jedoch nicht garantiert werden. Insbesondere für Disziplinen, die auf die Analyse von Beobachtungsdaten aus dynamischen Systemen (etwa sozioökonomische oder klimatische Daten) angewiesen sind, stellt dieser Umstand eine große Herausforderung dar. Darüber hinaus wäre es generell wünschenswert, von der Leistungsfähigkeit eines Algorithmus im Lernprozess sowie auf den Testdaten auf dessen allgemeine Leistungsfähigkeit zur Erstellung präziser Vorhersagen schließen zu können, wenn aus mehreren Kandidaten der optimale Algorithmus für eine bestimmte Anwendung ausgewählt werden soll.

Die *statistische Lerntheorie* verfolgt daher das Ziel, derartige Schlussfolgerungen zu ermöglichen. Mit verschiedenen, hauptsächlich wahrscheinlichkeitstheoretischen Werkzeugen der Mathematik versuchen Forschende auf diesem Gebiet, *statistische Garantien* für die Leistungsfähigkeit von Methoden des maschinellen Lernens zu geben. Ihrer Bezeichnung folgend sind diese Garantien so aufgebaut, dass sie zum einen garantiert gelten, sofern bestimmte grundlegende Annahmen über die Struktur der Daten erfüllt sind. Zum anderen haben sie jedoch einen statistischen Charakter, da sie lediglich eine Aussage darüber treffen, mit welcher Wahrscheinlichkeit die allgemeine Leistungsfähigkeit eines Algorithmus zur Erstellung präziser Vorhersagen schlechter als ein für den schlechtesten Fall angenommener Grenzwert ausfällt (vgl. von Luxburg und Schölkopf 2011). Die statistische Lerntheorie liefert folglich probabilistische Aussagen über die schlechteste zu erwartende Leistungsfähigkeit von Algorithmen und damit immerhin eine konservative Einschätzung darüber, wie diese sich jenseits des Lernprozesses verhalten werden. Ein großes und bislang ungelöstes Problem besteht jedoch darin, dass derartige statistische Garantien bestimmte in der Praxis recht häufig auftretende Bedingungen nicht erfassen. Viele der populärsten Methoden des maschinellen Lernens, insbesondere künstliche neuronale Netze, entziehen sich deshalb einer abschließenden Analyse durch die statistische Lerntheorie, sodass sich ihre Nutzung oftmals lediglich durch den (ebenfalls induktiven) Verweis auf erfolgreiche Anwendungen in der Vergangenheit rechtfertigen lässt (vgl. Bartlett *et al.* 2021).

Ein weiteres in der Literatur diskutiertes Problem des maschinellen Lernens ist dessen mangelnde *Robustheit*: Meist können Algorithmen im Anschluss an den Lernprozess nahezu fehlerfrei den Merkmalen in den vorliegenden Daten die korrekten Labels zuordnen und darüber hinaus äußerst präzise Vorhersagen für die Testdaten erstellen. In der eigentlichen Anwendung auf gänzlich neue Daten zeigt sich dann allerdings häufig eine deutlich verschlechterte Leistungsfähigkeit (vgl. Freiesleben und Grote 2023). Dies lässt sich unter anderem darauf zurückführen, dass Methoden des maschinellen Lernens ihre Spezifikation in manchen Fällen an bestimmte Eigenheiten der im Lernprozess verwendeten Daten anpassen, welche jedoch keinen

allgemeinen Zusammenhang zwischen Merkmalen und Labels darstellen und somit für die Vorhersage neuer Daten irrelevant sind („*overfitting*"). Unter Umständen kann sich zudem die Struktur der Daten grundlegend ändern, sodass ein innerhalb des Lernprozesses identifizierter Zusammenhang zwischen Merkmalen und Labels in neuen Daten nicht länger vorhanden ist. Ein Beispiel für derartige Änderungen ist die vor allem in den Sozialwissenschaften diskutierte *Performativität von Vorhersagen*: Vorhersagen können individuelles oder kollektives Verhalten beeinflussen und so dazu führen, dass bestehende Zusammenhänge zwischen Merkmalen und Labels verändert oder vollständig aufgehoben werden. Methoden des maschinellen Lernens sind in solchen Situationen nicht mehr in der Lage, präzise Vorhersagen zu erstellen, da sie ihre Spezifikation im Rahmen des Lernprozesses wie oben beschrieben an ebendiese bestehenden Zusammenhänge anpassen (vgl. Perdomo *et al.* 2020; Buchholz und Grote 2023).

Opazität und Erklärbarkeit des maschinellen Lernens

Die bisherigen Ausführungen haben zweierlei gezeigt: Zum einen beruht die Funktionsweise des maschinellen Lernens ganz entscheidend auf der Nutzung von Daten, weshalb eine ausgeprägte Kontinuität mit Methoden der klassischen Statistik und den aus diesem Kontext bekannten Problemen besteht. Zum anderen besteht die spezifische Zielsetzung des maschinellen Lernens im Erstellen präziser Vorhersagen, was eine Reihe weiterer, in der Statistik nicht zwingend relevanter methodologischer Probleme mit sich bringt. Methoden des maschinellen Lernens stehen allerdings noch einem weiteren, zutiefst erkenntnistheoretischen Problem gegenüber, das sich nicht aus der Nutzung von Daten oder der dabei verfolgten Zielsetzung, sondern direkt aus der Natur der Methoden selbst ergibt. Es handelt sich dabei um das nahezu ausschließlich in der Literatur über das maschinelle Lernen behandelte Problem der Opazität.

Das *Problem der Opazität*, welches in der einschlägigen Literatur teils synonym als „black-box problem" diskutiert wird, beschreibt den Umstand, dass in vielen Fällen unklar ist, *wie* und *weshalb* Methoden des maschinellen Lernens funktionieren. So ist beispielsweise häufig nicht ersichtlich, wie bestimmte Vorhersagen auf technischer Ebene durch die im Lernprozess entstandene Spezifikation hervorgebracht werden. Zudem bleiben häufig die Gründe verborgen, die Methoden des maschinellen Lernens auf Basis bestimmter Eingabedaten zur Erstellung einer spezifischen Vorhersage verleiten: Wird ein Bild etwa als Abbildung eines Huskys klassifiziert, weil darauf in der Tat ein Husky zu sehen ist oder aufgrund der Tatsache, dass darauf auch Schnee zu sehen ist und der Algorithmus im Lernprozess einen Zusammenhang zwischen Huskys und Schnee identifiziert hat? Letzteres wäre eine Spielart der in der Literatur eingehend diskutierten Verzerrungen oder – anthropomorphisch formuliert – Vorurteile (‚biases'), die Algorithmen aus verschiedenen Gründen aufweisen können (Fazelpour und Danks 2021) – und welche durch das Problem der Opazität oft unentdeckt bleiben. Es ist offensichtlich, dass solche Situationen insbesondere in wissenschaftlichen Anwendungen des maschinellen Lernens äußerst problematisch sind, da sich gewonnene Resultate unter diesen Voraussetzungen nur sehr begrenzt überprüfen lassen.

Jenseits derlei anschaulicher Beispiele besteht in der Literatur zwar Konsens über die Existenz des Problems der Opazität, eine einheitliche Begriffsdefinition existiert jedoch nicht, obwohl diese für eine zielgerichtete Lösung des Problems eigentlich unabdingbar wäre (vgl. Buchholz 2023a). Stattdessen werden verschiedene Ursachen angeführt, die das Problem einzeln oder gemeinsam hervorbringen. An erster Stelle ist in diesem Zusammenhang die Komplexität der Methoden zu nennen, die jene der aus der klassischen Statistik bekannten Methoden meist weit übersteigt: Methoden des maschinellen Lernens sind unter anderem deshalb so erfolgreich bei der Erstellung präziser Vorhersagen, weil ihre Spezifikation nicht selten Tausende oder gar Millionen von Parametern umfasst – Stellschrauben also, mit deren Hilfe Algorithmen im Lernprozess

komplexe Zusammenhänge aus den Daten abbilden können. Das Zusammenspiel dieser Parameter, insbesondere die Art und Weise, in der sie Eingabedaten in eine Vorhersage transformieren, sind angesichts ihrer schieren Zahl und der Komplexität ihrer Wechselwirkungen allerdings kaum durchschaubar (vgl. Boge 2022, S. 59f.). Eine zweite Hauptursache für das Problem der Opazität ergibt sich aus dem Umstand, dass die Funktionsweise des maschinellen Lernens nicht bereits zu Beginn einer Anwendung feststeht, sondern erst im Rahmen eines Lernprozesses durch den Algorithmus selbst abschließend festgelegt wird. Dies stellt vermutlich den größten Unterschied zwischen Methoden des maschinellen Lernens und anderen in der Wissenschaft verwendeten technischen Artefakten von Mikroskopen bis hin zu Satelliten dar. Denn während es im Falle der Letzteren für ein Individuum zwar *faktisch*, beispielsweise aufgrund eines Mangels an Zeit, unmöglich sein kann, die Funktionsweise vollständig zu durchschauen, wurde in der Literatur darauf verwiesen, dass dies im Falle des maschinellen Lernens aufgrund der dynamisch entstehenden Funktionsweise unter Umständen *prinzipiell* unmöglich ist (vgl. Schubbach 2021).

Das Problem der Opazität wirft die drängende Frage auf, ob sich die Verwendung von Methoden des maschinellen Lernens in der Wissenschaft überhaupt rechtfertigen lässt. Eine solche Rechtfertigung wäre ein wichtiger Schritt dazu, die Vorhersagen des maschinellen Lernens analog zu Ergebnissen, die mit Methoden der klassischen Statistik gewonnen wurden, und somit als Beiträge zur Überprüfung bestehenden oder Hervorbringung neuen Wissens betrachten zu können, gilt die Rechtfertigung doch als zentrale Eigenschaft, die Wissen von bloßer Meinung unterscheidet. Wie sollte der erkenntnistheoretischen Herausforderung durch das Problem der Opazität also begegnet werden? In der Literatur lassen sich dazu unter den Schlagworten Interpretierbarkeit, Erklärbarkeit und Reliabilismus drei einflussreiche Strategien unterscheiden.

Die Strategie der *Interpretierbarkeit* besteht im Wesentlichen darin, das Problem der Opazität durch eine geeignete Auswahl der Methoden gar nicht erst entstehen zu lassen. Statt äußerst komplexen und daher in ihrer Funktionsweise kaum durchschaubaren Methoden sollten deshalb sogenannte inhärent interpretierbare Methoden des maschinellen Lernens verwendet werden (vgl. Rudin 2019). Diese enthalten unter anderem weniger Parameter, die zudem auf einfachere Art miteinander verknüpft werden. Die gesamte Funktionsweise ist unter diesen Voraussetzungen durchschaubarer sodass beispielsweise besser ersichtlich ist, wie sich eine bestimmte Vorhersage aus der Spezifikation des Algorithmus ergibt. Für inhärent interpretierbare Methoden des maschinellen Lernens ist das Problem der Opazität also offenbar weniger gravierend als oben beschrieben. Allerdings hat der epistemische Vorteil geringerer Opazität einen – ebenfalls epistemischen – Preis: Wenngleich diese Schlussfolgerung umstritten ist, wurde wiederholt eingewendet, dass inhärent interpretierbare Methoden aufgrund ihrer geringeren Komplexität schlechter als andere Methoden des maschinellen Lernens in der Lage sind, allgemeine Zusammenhänge aus Daten abzuleiten und präzise Vorhersagen zu erstellen.

Die Strategie der *Erklärbarkeit* setzt deshalb darauf, nicht von der Verwendung komplexer Methoden abzurücken, sondern das Problem der Opazität als Preis für das Erstellen präziser Vorhersagen hinzunehmen und Instrumente vorzuschlagen, die verschiedene Aspekte der Methoden dennoch nachvollziehbar machen. Vor allem das Feld der sogenannten erklärbaren KI (auch „explainable AI“ oder „XAI“) beschäftigt sich mit der Entwicklung solcher Instrumente. Sie werden auf Methoden des maschinellen Lernens angewendet, nachdem diese den Lernprozess bereits durchlaufen und Vorhersagen erstellt haben. Aus diesem Grund werden die von Instrumenten der erklärbaren KI bereitgestellten Informationen auch als *post hoc*-Erklärungen bezeichnet. Diese liefern Einsichten in die Entstehung einzelner Vorhersagen, indem sie etwa durch Visualisierungen oder Quantifizierungen die für die Vorhersage entscheidenden Aspekte der Modellspezifikation oder ausschlaggebende Merkmale in den Daten hervorheben. Aufgrund der Verwendung des Begriffs ‚Erklärungen‘ im Zusammenhang mit Instrumenten der erklärbaren KI

hat die Debatte zur Erklärbarkeit in den vergangenen Jahren zunehmendes Interesse seitens der Wissenschaftstheorie erfahren, die bereits seit Jahrzehnten konzeptuelle und normative Aspekte des Erklärungsbegriffs verhandelt (vgl. Salmon 1989; Schurz 1990). Wissenschaftstheoretische Erkenntnisse, etwa über notwendige Eigenschaften guter Erklärungen, könnten daher das Feld der erklärbaren KI möglicherweise dabei unterstützen, die Opazität von Methoden des maschinellen Lernens zu verringern und diese so nachvollziehbarer zu machen. Ob eine solche Übertragung des wissenschaftstheoretischen Diskurses zum Erklärungsbegriff auf die Debatte zur erklärbaren KI sinnvoll und vor allem nützlich im Umgang mit dem Problem der Opazität sein kann, ist derzeit allerdings noch umstritten.

Was die *post hoc*-Erklärungen der erklärbaren KI gegenwärtig bereits zu leisten im Stande sind, ist dagegen deutlich klarer: Ihrem Namen folgend liefern sie Informationen über spezifische Aspekte der jeweils betrachteten Methode des maschinellen Lernens, und zwar wie oben beschrieben *nachdem* diese den Lernprozess bereits durchlaufen und Vorhersagen erstellt hat. Für Anwendungen des maschinellen Lernens in der Wissenschaft ist dies eine wichtige Feststellung, da der Fokus hier meist auf einem grundlegenden Phänomen liegt, das näher erforscht werden soll. Methoden des maschinellen Lernens erfassen jedoch unter Umständen lediglich Korrelationen und nicht notwendigerweise die kausale Struktur des Zielphänomens. Zudem ist weitgehend ungeklärt, inwiefern Methoden des maschinellen Lernens überhaupt gute Modelle für das jeweils untersuchte Zielphänomen sind, ob sie dessen zentrale Eigenschaften in ihrer Spezifikation berücksichtigen und also repräsentieren (vgl. Knüsel und Baumberger 2020; Sullivan 2022). In dem Maße, in dem die Beziehung zwischen Methoden des maschinellen Lernens und den jeweiligen Zielphänomenen ungeklärt bleibt, muss folglich auch die Aussagekraft von Instrumenten der erklärbaren KI über diese Zielphänomene unklar bleiben, schließlich stehen sie lediglich in einer indirekten Beziehung zu den Letzteren. Insgesamt ist der Nutzen der erklärbaren KI im Forschungsprozess deshalb begrenzt.

Die Strategie des *Reliabilismus* besteht schließlich darin, dem Problem der Opazität durch mathematisch beweisbare Resultate über die allgemeine Leistungsfähigkeit von Methoden des maschinellen Lernens entgegenzuwirken (vgl. Grote *et al.* 2024). Auch wenn die Funktionsweise der Letzteren oder die Erstellung einzelner Vorhersagen undurchschaubar sein mag, ist auf diese Weise eine Einschätzung über die zu erwartende Präzision bestimmter Methoden unter bestimmten Bedingungen möglich. Diese Strategie, die maßgeblich von der oben vorgestellten statistischen Lerntheorie verfolgt wird, erlaubt es somit, die Auswahl bestimmter Methoden *a priori* zu rechtfertigen. In der klassischen Statistik existiert eine Vielzahl solcher *a priori*-Resultate, weshalb die mit diesen Methoden erlangten Ergebnisse – unter gewissen Annahmen – auf einem soliden methodologischen Fundament stehen. Trotz ihrer probabilistischen Natur lassen sich diese Ergebnisse deshalb oft als Beiträge zur Hervorbringung neuen Wissens beschreiben und machen die klassische Statistik so zu einem unverzichtbaren Werkzeug der empirischen Forschung. In dem Maße, in dem das Projekt des Reliabilismus und also die statistische Lerntheorie Fortschritte macht und beispielsweise statistische Erfolgsgarantien für neuartige Algorithmen zu geben vermag, wird dies, insbesondere im Lichte seiner ausgeprägten Kontinuität mit der klassischen Statistik, auch für das maschinelle Lernen gelten können.

Literaturverzeichnis

Baldi, Pierre, Peter Sadowski, und Daniel Whiteson. 2014. Searching for Exotic Particles in High-energy Physics with Deep Learning. *Nature Communications* 5(1): 4308.

Bartlett, Peter L., Andrea Montanari, und Alexander Rakhlin. 2021. Deep Learning: A Statistical Viewpoint. *Acta Numerica* 30: 87-201.

Boge, Florian. 2022. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines* 32(1): 43-75.

Breiman, Leo. 2001. Statistical Modeling: The Two Cultures. *Statistical Science* 16(3): 199-231.

Buchholz, Oliver. 2023a. A Means-End Account of Explainable Artificial Intelligence. *Synthese* 202(2): 33.

Buchholz, Oliver. 2023b. The Deep Neural Network Approach to the Reference Class Problem. *Synthese* 201(3): 111.

Buchholz, Oliver, und Thomas Grote. 2023. Predicting and Explaining With Machine Learning Models: Social Science as a Touchstone. *Studies in History and Philosophy of Science* 102: 60-69.

Buchholz, Oliver, und Eric Raidl. 2022. A Falsificationist Account of Artificial Neural Networks. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/721797.

Coley, Connor W., Dale A. Thomas III, Justin A.M. Lummiss, Jonathan Jaworski, Christopher P. Breen, Victor Schultz, Travis Hart, Joshua S. Fishman, Luke Rogers, Hanyu Gao, Robert W. Hicklin, Pieter P. Plehiers, Joshua Byington, John S. Piotti, William Green, A. John Hart, Timothy F. Jamison, Klavs F. Jensen. 2019. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* 365(6453): eaax1566.

Fazelpour, Sina und David Danks. 2021. Algorithmic Bias: Senses, Sources, Solutions. *Philosophy Compass* 16(8): e12760.

Foreman-Mackey, Daniel, Benjamin T. Montet, David W. Hogg, Timothy D. Morton, Dun Wang, und Bernhard Schölkopf. 2015. A Systematic Search for Transiting Planets in the K2 Data. *The Astrophysical Journal* 806(2): 215.

Freiesleben, Timo, und Thomas Grote. 2023. Beyond Generalization: A Theory of Robustness in Machine Learning. *Synthese* 202(4): 109.

Grote, Thomas, und Oliver Buchholz. i.E. Machine Learning in Public Health and the Prediction-Intervention Gap. In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, Hrsg. Juan Durán und Giorgia Pozzi. Cham: Springer.

Grote, Thomas, Konstantin Genin, und Emily Sullivan. 2024. Reliability in Machine Learning. *Philosophy Compass* 19(5): e12974.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, … und Demis Hassabis. 2021. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596(7873): 583-589.

Knüsel, Benedikt, und Christoph Baumberger. 2020. Understanding Climate Phenomena with Data-driven Models. *Studies in History and Philosophy of Science* 84: 46-56.

Mitchell, Tom M. 1997. *Machine Learning*. New York und London: McGraw Hill.

Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner, und Moritz Hardt. 2020. Performative Prediction. In *Proceedings of the 37th Conference on Machine Learning* (PMLR) 119, 7599-7609.

Popper, Karl. 1935. *Logik der Forschung*, Wien: Julius Springer Verlag.

Popper, Karl. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.

Poser, Hans. 2012. *Wissenschaftstheorie. Eine philosophische Einführung*. Stuttgart: Reclam.

Reichenbach, Hans. 1935. *Wahrscheinlichkeitslehre. Eine Untersuchung über die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*, Leiden: A. W. Sijthoff.

Romeijn, Jan-Willem. 2022. Philosophy of Statistics. In *The Stanford Encyclopedia of Philosophy*, Hrsg. Edward N. Zalta und Uri Nodelman. https://plato.stanford.edu/archives/fall2022/entries/statistics/. Zugegriffen am 20.03.2024.

Rudin, Cynthia. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1(5): 206-215.

Salmon, Wesley C. 1989. Four Decades of Explanation. In *Scientific Explanation*, Hrsg. Philip Kitcher und Wesley C. Salmon, 3-219. Minneapolis: University of Minnesota Press.

Schubbach, Arno. 2021. Judging Machines: Philosophical Aspects of Deep Learning. *Synthese* 198(2): 1807-1827.

Schurz, Gerhard, Hrsg. 1990. *Erklären und Verstehen in der Wissenschaft*. Berlin und Boston: De Gruyter.

Segler, Marwin, Mike Preuss und Mark P. Waller. 2018. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* 555(7698): 604-610.

Sullivan, Emily. 2022. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science* 73(1): 109-133.

Venn, John. 1876. *The Logic of Chance*, 2. Aufl. London: Macmillan.

von Luxburg, Ulrike und Bernhard Schölkopf. 2011. Statistical Learning Theory: Models, Concepts, and Results. In *Handbook of the History of Logic*, Bd. 10, Hrsg. Dov M. Gabbay, Stephan Hartmann und John Woods, 651-706. Amsterdam und Boston: North Holland.