Technisch, analytisch und ethisch robuste KI-Systeme: Alternative Konzepte und Implementierungen

Oliver Maspfuhl

1 Einleitung

Künstliche Intelligenz (KI) ist längst keine Zukunftstechnologie mehr, wohl aber eine Technologie mit großer Zukunft. Es wäre wohl nicht übertrieben, sie in eine Reihe mit bahnbrechenden Erfindungen wie etwa die Dampfmaschine oder der Radiowellen zu stellen. Ihre Auswirkungen auf die menschliche Gesellschaft und insbesondere deren politische und wirtschaftliche Fundamente zeichnen sich bereits heute deutlich ab.

Wie in vielen anderen Fällen, erleben wir ein exponentielles Wachstum der Entwicklung und Anwendung von KI-Systemen. Mit der Verfügbarkeit geeigneter Big-Data-Plattformen sowie von Mechanismen für Nutzerfeedbacks und damit von Menschen kategorisierter („gelabelter") Daten in gigantischen Mengen, sind in den letzten zehn Jahren zwei wesentliche Hindernisse dafür aus dem Weg geräumt worden.

Ein besonderer Aspekt von KI-Anwendungen besteht darin, mitunter tief in die Privatsphäre ihrer menschlichen Nutzer einzugreifen, ohne dass diese sich der zugrunde liegenden Mechanismen wirklich bewusst wären. Denken wir nur an Suchmaschinen, Kauf- und Konsumempfehlungen, Partnervermittlung oder medizinische Diagnosen. Das vielerorts geringe Verständnis für die Natur von KI und ihre technische Funktionsweise sollte daher nicht leichtfertig hingenommen werden. Denn vielleicht noch mehr als bei anderen Technologien ist

ein solches Verständnis unumgänglich, wenn Künstliche Intelligenz auf lange Sicht einen Beitrag zur Wohlfahrt unserer freiheitlichen Gesellschaft leisten soll. Oberflächliche Bilder und Methaphern sind hier eher schädlich und führen allenfalls zu Vorurteilen.

Als von je her ihrer Rolle nach datengetriebene Finanzunternehmen müssen Banken die neue Technologie vollumfänglich beherrschen und nutzen können. Ein korrektes Verständnis ist dafür die notwendige Voraussetzung und sollte zum festen Bestandteil (in abgestuftem Detailgrad selbstverständlich) der Bankausbildung gehören.

Der vorliegende Beitrag verfolgt daher zwei Ziele. Zum einen möchte er die grundlegenden Konzepte so darlegen, dass sie Anwenderinnen und Anwendern ein solides Verständnis für KI-Anwendungen ermöglichen. Sie sollen insbesondere in die Lage versetzt werden, sich in den auf von Halbwissen und Missverständnissen geprägten Debatten um Nutzen und Gefahren der modernen Datenanalyse zu positionieren – nicht zuletzt im Dialog mit den Aufsichtsbehörden.

Letzteren fällt derzeit die ungemein komplexe Aufgabe zu, ein Rahmenwerk für die Nutzung von KI zu schaffen. Die beste Unterstützung, die die einzelnen Institute auch im eigenen Interesse für ein gelungenes Regelwerk leisten können, besteht sicher im Einüben klarer, korrekter Begriffe und Standards. Wir kennen dies bereits etwa vom Begriff des Kreditrisikos aus den Baseler Papieren[1].

Auf Grundlage des vorgeschlagenen Begriffsrahmens soll im zweiten Teil des Beitrags dann dargestellt werden, wie ein Rahmen für die Entwicklung und Nutzung von KI-Anwendungen aussehen könnte. Dabei werden technische, analytische und ethische Aspekte berücksichtigt.

Das Ziel soll dabei nicht sein, existierende systematische Darstellungen zu den Grundlagen Künstlicher Intelligenz, der statistischen Datenanalyse, Datenschutz oder Ethik zu ersetzen (Stuart Russel, Peter Norvig, 2010). Das Augenmerk liegt vielmehr darauf, gezielt diejenigen Aspekte anzusprechen, die am

1 Auf die Verwendung des mathematischen Formalismus muss hier natürlich vollständig verzichtet werden in der Hoffnung, die kritischen Punkte werden auch durch beispielhafte bildliche Darstellungen verdeutlicht.

häufigsten missverstanden oder falsch beziehungsweise ungeschickt interpretiert werden, um ein alternatives Verständnis zu motivieren und zu ermöglichen.

Die Absicht des Beitrags wäre erreicht, wenn er dadurch all jenen Instituten, die sich mit dem Zugang zu KI bisher schwertun, beim richtigen Start und den ersten Schritten behilflich wäre und zur Findung eines gemeinsamen Begriffs- und Normensystems über alle Institute hinweg beitrüge.

2 Aspekte der Künstlichen Intelligenz in der modernen Datenanalyse – ein alternativer Zugang

Überall stößt man auf diverse Definitionen für den komplexen Begriff der Künstlichen Intelligenz. Wir wollen hier einige Aspekte einer Definition ansprechen und damit eine Begriffsbestimmung vornehmen, die für unsere praktischen Zwecke sinnvoll ist, aber ohne den Anspruch auf Allgemeingültigkeit auskommt.

Die hier angesprochenen Aspekte stellen (dem Begriff „Aspekt“ entsprechend) keine Systematik dar und sind nicht überschneidungsfrei.

2.1 Aspekt 1: Lösungen selber finden

Die klassische, tiefliegende Fähigkeit von KI-Systemen besteht darin, für ein gestelltes Problem selbstständig eine Lösung zu finden, also nur die Aufgabe, aber keinen Lösungsweg gestellt zu bekommen. Diese Fähigkeit wird mit Intelligenz bzw. intelligentem Verhalten gleichgesetzt. Neben der Ausführung logischer Operationen stehen für die Lösungsfindung Erfahrung und eventuell auch die Möglichkeit zum Experimentieren zur Verfügung.

In der Bankpraxis ist aus diesem Blickwinkel festzuhalten, dass KI-Systeme in aller Regel zur Automatisierung von Aufgaben genutzt werden, für die eine algorithmische Lösungsbeschreibung nicht, nur unvollkommen oder nur mit unvernünftigem Aufwand erstellt werden kann.

Als Beispiel ließe sich die automatische Sortierung von Dokumenten gemäß Typ und Inhalt anführen. Eine allgemeine Beschreibung des nötigen Vorgehens

ist in aller Regel nicht mit sinnvollem Aufwand möglich. Die Übersetzung von Dokumenten in eine andere Sprache ist ein weiteres Beispiel.

2.2 Aspekt 2: Replikation intelligenten Verhaltens

Geübte Menschen können die aufgeführten Beispielaufgaben i. d. R. sehr gut bewältigen. Daher ist eine weitere klassische Eigenschaft ebenfalls eine naheliegende Definition für KI: Die Fähigkeit, Tätigkeiten auszuführen, die normalerweise menschliche Intelligenz erfordern.

Für den Einsatz von KI in der Bankpraxis ist dieser Aspekt von großer Bedeutung, wie später erläutert wird. Mit dieser Definition können z. B. regelbasierte und durch Maschinelles Lernen erzeugte Entscheidungssysteme sowie vor allem die häufig anzutreffenden Mischformen daraus, unter gemeinsamen Richtlinien behandelt werden, was sich als das einzig sinnvolle Vorgehen erweisen wird. Selbstverständlich werden wir es in aller Regel mit KI-Systemen zu tun haben werden, die gemäß Aspekt 1 selbstständig Lösungen finden.

2.3 Aspekt 3: Einzelproblem vs. Typenproblem

Man erkennt bei den genannten Beispielen bereits, dass sowohl für den Menschen als auch für ein KI-System, das eigenständig Lösungen findet (Aspekt 1), offenbar zwei Phasen der Problemlösung zu unterscheiden sind: Das Auffinden der allgemeinen Lösung und das Anwenden der gefundenen allgemeinen Lösung auf konkrete Aufgaben des vorher beschriebenen Typs. Wir betrachten in der Praxis per heute nur Aufgaben dieser Art.

Die Lösung von erstmalig auftretenden Problemen liegt dagegen außerhalb des derzeit Erreichbaren. In einigen Bereichen der Robotik versucht man, derartige Probleme durch eine Zerlegung in bekannte Teilprobleme zu lösen. Dieser Ansatz ist möglich und steht auch Anwendungen in Finanzunternehmen zur Verfügung. Ein Beispiel ist die Erstellung eines Chatbots, der in der Lage ist, sehr frei mit Kunden zu kommunizieren. In der Praxis sind diese Systeme derzeit aber gerade in der Kommunikation mit Menschen darauf beschränkt, bestimmte bekannte Kommunikationsabläufe anzusteuern (etwas eine Kontoeröffnung) und dann einen normierten Informationsabruf für eine Auswahl bekannter Prozesse in die Wege zu leiten.

2.4 Aspekt 4: Lernen

Bei beiden Arten der Problemstellung in Aspekt 3 ist es offensichtlich notwendig, dass das System seine Leistungsfähigkeit in gewissem Maße eigenständig verbessert. Bei der Lösung von Typenproblemen passiert dies (ggf. iterativ) in der erste Phase. Für Systeme mit allgemeiner Intelligenz ist dies dagegen ein kontinuierlicher Prozess, in dem allgemeine, immer abstraktere Grundlagen für die Erkenntnis der Umwelt und die Entwicklung von Problemlösungsstrategien gelegt werden. Beide Prozesse bezeichnet man i. d. R. als „Lernen".

Es ist wichtig zu verstehen, dass dieser Begriff im Fall einer allgemeinen Intelligenz durchaus zutreffend sein könnte, da er tatsächlich ein grundlegendes Verständnis für die abstrakten Gegenstände hervorbringt, ähnlich dem Prozess des menschlichen Lernens.

Im Fall eines KI-Systems zum Lösen spezifischer Probleme ist er jedoch unbedingt als Metapher zu verstehen, die nichts mit dem Erwerb von Verständnis zu tun hat. Ein Beispiel mag dies verdeutlichen: Für ein maschinelles gelerntes Bilderkennungssystem, das Katzen auf Bildern erkennen kann, gibt es keinen Unterschied zwischen dem Foto einer Katze und der Katze selbst. Katze und Katzenbild sind aber unterschiedliche Konzepte. Die Ursache vieler prominenter Schwachstellen von KI-Systemen liegt genau hier[2].

2.5 Aspekt 5: Bewusstsein

Häufig wird in der Diskussion um die Problemtypen des Aspektes 3 mit den Begriffen starke/schwache KI gearbeitet, dies sei der Vollständigkeit halber erwähnt, ohne näher auf diese Unterscheidung einzugehen. Wichtig ist allerdings zu wissen, dass mit hoher Wahrscheinlichkeit eine allgemeine Problemlösungskompetenz bzw. eine „starke" KI mit abstrakter Lernfähigkeit im Sinne von Aspekt 4 mit der Erzeugung eines künstlichen Bewusstseins einhergehen müsste.

Es existieren heute komplexe Theorien des Bewusstseins, keine jedoch kann allgemein unumstrittene Gültigkeit beanspruchen.

2 Richtig ist jedoch, dass während der Phase des Auffindens einer Lösung für ein Typenproblem bei der Verwendung eines datengetriebenen Lernalgorithmus durchaus Strukturen in den Trainingsdaten erkannt werden, die nicht offensichtlich sind, etwa relevante Bereiche für eine Bilderkennung.

Als gesichert kann gelten, dass das natürliche Bewusstsein ebenso wie die Fähigkeit zu abstrakter Problemlösung physiologisch im Großhirn verortet ist, während im Kleinhirn standardisierte, sich wiederholende Abläufe gesteuert werden. Wir werden hierauf im Abschnitt Maschinelles Lernen zurückkommen.

2.6 Aspekt 6: Übermenschliche Fähigkeiten – Big Data

Vor dem Hintergrund der gerade beschriebenen Grenzen ist es wichtig zu verstehen, dass damit in keiner Weise ausgeschlossen ist, dass KI-Systeme bei spezifischen Aufgaben eine der menschlichen weit überlegenen Leistungsfähigkeit aufweisen – ganz prominent wiederum in den beiden Königsdisziplinen, der Bild- und Texterkennung. Die Unterscheidung hunderter Hunderassen auf Fotos wird nur für Experten mit geringer Fehlerquote möglich sein. Dass sie aber – mit entsprechender Erfahrung – offenbar für grundsätzlich lösbar ist und einem KI-System, das i. d. R. auf Maschinellem Lernen beruht, praktisch unbegrenzt Lerndaten und „Gedächtnis" zur Verfügung steht, wird letztlich sogar den einzelnen Experten übertreffen. Ähnliche Beispiele kennen wir aus der Karzinomerkennung oder etwa dem Aufspüren von Anomalien in Maschinen- oder Transaktionsdaten.

Der entscheidende Aspekt hierbei ist der „Big-Data-Effekt", die Quantität der verfügbaren Information, die in eine neue Qualität bei der Problemlösung umschlägt.

2.7 Aspekt 7: KI ist kein Teilgebiet der Informatik

Wie ist KI abschließend technisch als Fachgebiet einzuordnen? Sehr weit verbreitet ist die Ansicht, es handele sich um ein Teilgebiet der Informatik. Unstrittig ist die Tatsache, dass sehr viele KI-Anwendungen von Informatikern entwickelt werden und unter ihnen ein hohes Interesse an diesem Gebiet herrscht. Man kann auch grundsätzlich zustimmen, dass die Erschaffung Künstlicher Intelligenz – als Ziel – in den Bereich der Theoretischen Informatik fällt.

Wir werden im Folgenden sehen, dass es für die technische Umsetzung nicht so sehr auf das Ziel der Schaffung „intelligenter" Programme ankommt, sondern darauf, wie dies geschieht. Die entsprechenden Methoden entstammen sämtlich

der Mathematik, vor allem deren Teilgebieten der Statistik, Optimierung und Geometrie.

Die Entwicklung von KI-Systemen ohne hinreichenden Hintergrund – vor allem in statistischer Testtheorie und Optimierungsmethoden – führt immer wieder zu spektakulären Misserfolgen selbst bei großen Anbietern. Ein Großteil der Diskussionen um die sogenannte „Erklärbarkeit" von KI hat ihren Ursprung auch darin, dass KI-Methoden von den Anwendern – wie in der Informatik allgemein üblich – als fertige Tools und somit methodische Black Boxes angewendet werden. Leider wird dieses Etikett oft – zu Unrecht – der Methodik selbst angeheftet.

Es ist daher dringend anzuraten, KI-Entwicklung als eine interdisziplinäre Aufgabe zwischen angewandter Mathematik, Informatik und Fachexpertise zu verstehen. Praktisch alle spezialisierten Ausbildungsangebote für Data Science erfüllen diese Anforderung derzeit nicht. Sie vermitteln ein grundsätzliches Verständnis, aber z. B. nicht die Fähigkeit technisch zu prüfen, ob ein bestimmtes Lernverfahren tatsächlich konvergiert ist. Daher ist beim Aufbau von Datenanalytik-Teams unbedingt darauf zu achten, alle erwähnten Bereiche abzudecken.

3 Maschinelles Lernen in der modernen Datenanalyse

Dieser Abschnitt ist bewusst kein Unterabschnitt des vorherigen, so wie Maschinelles Lernen (ML) in unserem Kontext nicht als Teilgebiet der KI gesehen werden sollte. Zunächst sollen aber die grundlegenden Begriffe ohne mathematischen Formalismus vorgestellt werden. Für eine ausführliche Einführung sei auf die Standardwerke verwiesen (Bishop, 2006), (Trevor Hastie et al., 2009), (Ian Witten et al., 2017), (Murphy, 2012).

3.1 „Lernen" als Allegorie

Der Lernprozess eines ML-Algorithmus ähnelt nicht dem menschlichen Lernen in Abstraktionen, die eine Übertragung von Konzepten in neuen Zusammenhängen ermöglichen. Neuronale Netze ähneln dem Kleinhirn, das sich für die Verarbeitung komplexer Signale ständig optimiert, aber nicht denken kann.

Ebenso besteht die Funktionsweise von ML-Algorithmen in der Optimierung von Modellparametern

3.2 Grundsätzliche Funktion: Optimierung von Modellparametern

Ein ML-Lernalgorithmus ist eine i. d. R. in Programmcodes implementierte Vorschrift, nach der ein mathematischer Modelltyp, also eine (i. d. R. komplexe) Beziehung zwischen Variablen repräsentiert durch eine geometrische Figur (Kurve, Fläche etc.) im Datenraum, das mit freien Parametern an gegebene Daten zu den Variablen optimal angepasst werden kann. Das Ergebnis bezeichnet man als ML-Modell, also der Modelltyp mit den optimalen Parametern.

Ein ML-Modell ist also immer ein vereinfachtes Abbild der realen Daten – und enthält damit insbesondere auch einen großen Teil der Informationen, die in den Daten stecken. Dies ist vor allem im Hinblick auf den Datenschutz ein wichtiger Fakt, der oft übersehen wird.

Sehr häufig werden Lernalgorithmus, Modelltyp und Modell verwechselt, mit teilweise radikalen Folgen für die Gültigkeit von Argumentationslinien. Die Unterscheidung dieser Begriffe ist von grundlegender Bedeutung: Beispielsweise lässt sich das fertige Modell im Grunde nicht mehr von einem menschlichen Regelsystem unterscheiden, mit der Konsequenz, dass man für beide die gleichen Methoden zur Beurteilung der Leistungsfähigkeit anwenden sollte.

Sehr wichtig ist das Verständnis, dass die „gegebenen Daten“ zwar i. d. R. durch eine Auswahl Trainingsdaten repräsentiert werden, grundsätzlich aber alle potenziell verfügbaren Daten gemeint sind, auch zukünftig zu erzeugende. Das bedeutet, dass die optimalen Parameter nicht unbedingt ein optimales Modell der Trainingsdaten liefern, sondern eines, das unter statistischen Gesichtspunkten die beste Chance hat, auch unbekannte Datenpunkte gut zu beschreiben. Um eine Aussage darüber treffen zu können, nutzt man bekanntermaßen Validierungsstichproben, die erst nach Fertigstellung des Modells zur finalen Beurteilung seiner Leistung herangezogen werden.

3.3 Automatische Featuregenerierung

An dieser Stelle ist es wichtig zu erwähnen, dass es neben den Parametern selbst auch deren Struktur ist, die die Lösung der Optimierungsaufgabe auf eine neue Stufe hebt. Im Gegensatz zu klassischen Regressionsverfahren, bei denen die Eingabewerte in aller Regel eine klare Bedeutung haben (bzw. so gebildet werden können, etwas als Verhältniszahlen), ist dies gerade bei Deep-Learning-Verfahren, also der Verwendung tiefer neuronaler Netze nicht der Fall: Bei der Text- oder Bilderkennung hat der einzelne Grauwert oder der einzelne Buchstabe keine eigenständige Bedeutung. Erst durch komplexe Kombinationen von Eingabewerten entstehen sinnvolle Einheiten. Das automatische Bilden dieser Kombinationen ist die größte Stärke der fortgeschrittenen Verfahren und kann daher zu Recht, im oben beschriebenen Sinne, als Lernen bezeichnet werden (Ian Goodfellow, Yoshua Bengio, Aaron Courville, 2016).

3.4 Anwendung von Modellen

Verfügt man über ein Modell für die Daten, kann man es nutzen, um aus einer teilweisen Kenntnis relevanter Werte zu einem Datenpunkt Vorhersagen oder Wahrscheinlichkeitsaussagen für die übrigen Werte abzuleiten. Häufig enthält das Modell per se schon eine sogenannte Zielvariable – also eine Größe, die man anhand anderer Werte vorhersagen möchte, etwa ein Kreditausfallindikator. Es ist jedoch grundsätzlich auch möglich, etwa in einem Ratingmodell, eine Schätzung für den Umsatz eines Unternehmens aus einer gegebenen Ausfallwahrscheinlichkeit abzuleiten. Dies ist vor allem unter Datenschutzaspekten bemerkenswert.

Ein wichtiger Unterschied in der Anwendung von Modellen besteht darin, ob es sich um eine echte zeitliche Prognose, also eine Aussage über die zeitliche Zukunft handelt (etwa der Ausfall eines Schuldners innerhalb eines Jahres), oder eine Aussage, die im Prinzip sofort geprüft werden könnte, wie die korrekte Übersetzung eines Fachbegriffes.

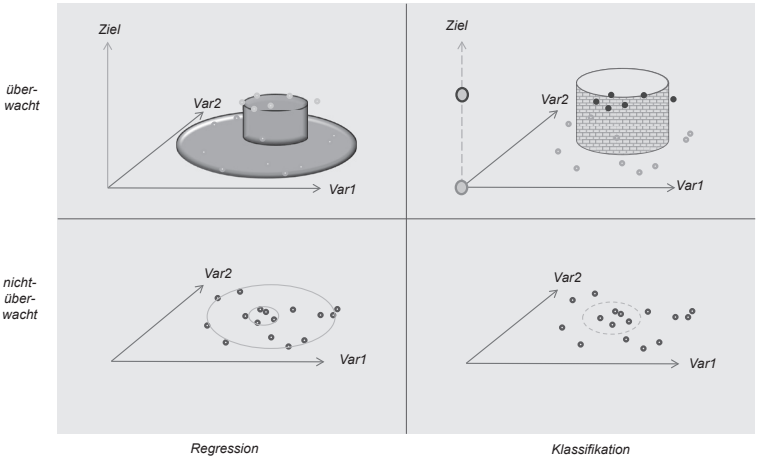
Echte Prognosemodelle stellen wesentlich höhere Anforderungen an die Modellierung und unterscheiden sich daher bezüglich der relevanten Verfahren zur Erstellung und Validierung grundlegend von Modellen, die einen gleichzeitigen Zusammenhang abbilden.

Insbesondere spielen für Prognosemodelle kausale Zusammenhänge eine zentrale Rolle, die ihrer Natur nach zeitlich entfernte Ereignisse verbinden, während gleichzeitig Prognosen natürlicherweise auf Korrelationen beruhen.

3.5 Geometrischer Vergleich von Typen ML-Modelltypen

Zur Veranschaulichung einiger (übergreifender) ML-Modelltypen sollen hier nur vier sehr einfache Beispiele dienen – falls nötig, finden sich alle Details in der Standardliteratur.

Abbildung 1: Einteilung von ML-Modelltypen



Quelle: Commerzbank AG.

Sinnvolle Unterscheidungen von Modelltypen liefern die Einteilung nach überwachten und nicht-überwachten sowie nach Regressions- und Klassifikationsmodellen. Abb. 1 zeigt links oben schematisch eine Datenpunktmenge, die in einem überwachten Regressionsverfahren durch eine Fläche beschrieben wurde. Jeder Datenpunkt besteht neben zwei (sogenannten erklärenden oder Feature-) Variablen Var1 und Var2 aus einem Zielwert Ziel, die angepasste Fläche be-

schreibt hier also den geometrischen Zusammenhang zwischen den Variablen Var1 und Var2 und Ziel. Ein Beispiel für diese Modellklasse ist die klassische Multivariate Regression.

In der links unten dargestellten nichtüberwachten Version wird dieselbe Datenmenge ohne die Dimension der Zielvariablen betrachtet. Das angepasste Modell hat nun eine Dimension weniger und beschreibt lediglich die Beziehung zwischen Var1 und Var2 in Form einer (zweiteiligen) Kurve. Trotzdem kann man erkennen, dass auch in dem nicht-überwachten Modell eine ähnliche Struktur der Daten erfasst wurde wie in dem Modell mit bekannter Zielvariable. Dies ist durchaus typisch und eine wichtige Erkenntnis für die Modellierung: Wenn die innere Struktur der Variablen keine Beziehung zur Zielvariablen aufweist, ist i. d. R. auch ein überwachtes Verfahren nicht erfolgreich. Trotzdem gilt selbstverständlich auch, dass nur überwachte Verfahren die Chance bieten, die Feinheiten der Beziehung zwischen Zielvariable und Features zu erfassen. Ein Beispiel für diesen Modelltyp wäre etwa die klassische Hauptkomponentenanalyse bzw. im nichtlinearen Fall ein neuronales Autoencoder-Netzwerk.

Während die Regressionsverfahren die Daten mithilfe mathematischer Beziehungen, also Gleichungen zu erfassen versuchen, beruhen Klassifikationsverfahren eher auf Ungleichungen: Sie teilen den Datenraum in unterschiedliche Bereiche, i. d. R. über die Definition von begrenzenden Teilmengen. Die gestrichelten Flächen bzw. Kurven auf den Bildern auf der rechten Seite von Abb. 1 sollen dies für den überwachten und nicht-überwachten Fall darstellen. Die Zielvariable ist nun auf einen diskreten, hier binären Wert reduziert, der zusätzlich farblich hervorgehoben wurde. Anstelle eines vereinfachten geometrischen Modells für die Datenmenge liefern diese Verfahren also ein duales Modell für die „Leerräume“ zwischen den Datenpunkten. Logistische Regression und 1-Klassen-Support-Vector-Maschinen liefern Beispiele für ein überwachtes bzw. nicht-überwachtes Klassifikationsverfahren.

Auch aus der Gegenüberstellung der Regressions- und Klassifikationsansätze kann man erkennen, dass beide Modelltypen in enger Beziehung stehen und geometrisch eng verwandt sind.

3.6 Statistischer Vergleich – Robuste Lernalgorithmen

Worin bestehen die wesentlichen Unterschiede zwischen den Modelltypen? Wie oben erwähnt, führt die geometrische Verwandtschaft in vielen Fällen zu ähnlichen Ergebnissen in der Modellgüte, also dem Maß, indem das Modell die Daten richtig beschreibt. Typischerweise wird sie – etwa im Fall einer Klassifikation – mit einer Konfusionsmatrix über einer Validierungsdatenmenge gemessen. Diese Matrix stellt tatsächliche Klassen und vom Modell vorhergesagte Klassen gegenüber (Abb. 2).

Abbildung 2: Konfusionsmatrix für eine (0,1)-Klassifikationsmodell

Anzahl Validierungsdatenpunkte		*Vorhergesagte Klasse* 0	1
Tatsächliche Klasse	0	**T**rue **N**egatives	**F**alse **P**ositives
	1	**F**alse **N**egatives	**T**rue **P**ositives

Quelle: Commerzbank AG.

Häufig werden Modelle vor allem über solche Maße verglichen, die ein gutes Gefühl für die Leistungsfähigkeit als Ganzen geben.

Was in dieser Sicht jedoch nicht enthalten ist, sind Aussagen darüber, mit welcher statistischen Unsicherheit die Einzelprognosen des Modells behaftet sind. Diese sind jedoch in der Anwendung in der Praxis häufig von großer Bedeutung, insbesondere für die Beurteilung der Nachvollziehbarkeit und für das Modellrisikomanagement.

Die wesentlichen Unterschiede zwischen den einzelnen Modellklassen und den darin enthaltenen Modelltypen zeigen sich genau an dieser Stelle, und vor allem bei den für ihre Erstellung verwendeten Lernalgorithmen. Einige Algorithmen liefern Wahrscheinlichkeiten für die Richtigkeit der Zuordnung zu einer Klasse, manche können Unsicherheiten der Parameter quantifizieren. Lernverfahren mit sogenannten Penalties erlauben es, zu komplexe Modellwahlen zu verhindern. Viele moderne Algorithmen besitzen heute Bayes'sche Varianten, die es erlauben, Vorwissen in die Modellparameter zu integrieren.

Zu den Details sei wiederum auf die Literatur verwiesen, hier soll nur festgehalten werden, dass die Wahl des Lernalgorithmus wesentlichen Einfluss auf die statistisch robuste Validierbarkeit und Anwendbarkeit des ML-Modells auf neue Daten hat. Oft liefern ähnlich aussehende Modelle sehr ähnliche Ergebnisse für viele bekannte Datenpunkte, liegen aber in speziellen Einzelfällen dramatisch auseinander. Die Wahl des richtigen Lernalgorithmus hilft, Modell zu wählen, die auch in diesen Einzelfällen eher richtige Entscheidungen fällen oder in der Lage sind zu erkennen, dass eine Entscheidung nicht möglich ist.

3.7 Ist ML eine Teilmenge von KI?

Zum Abschluss dieser kurzen Sammlung von Aspekten zum Maschinellen Lernen soll noch sein Verhältnis zur KI beleuchtet werden.

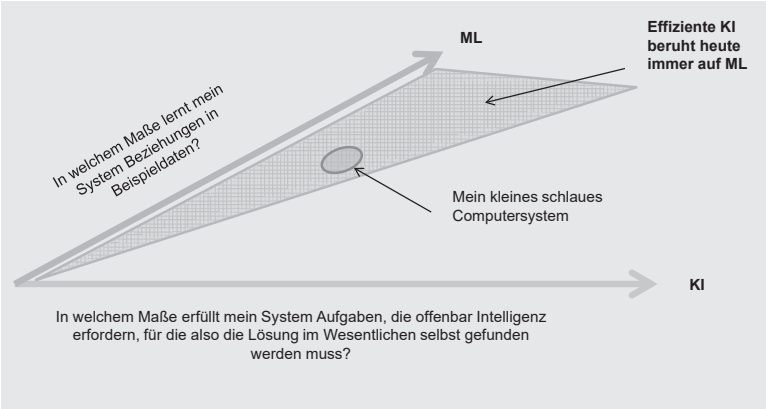
Sehr häufig findet man Darstellungen, in denen ML als Teilmenge der KI präsentiert wird. Diese drücken durchaus eine richtige Sicht aus – wenn man KI als „KI-Verfahren" liest. Legen wir jedoch die im ersten Abschnitt verwendeten Begriffe von KI zugrunde, wird klar, dass diese Darstellung verkürzt ist.

KI beschreibt zunächst ein Ziel, die Automatisierung intelligenten Verhaltens. Ein KI-System ist durch eine Antwort auf die Frage nach seinem Zweck („Wozu?") charakterisiert.

ML beschreibt ein methodisch-technisches Verfahren, mit dem KI-Anwendungen besser als auf jedem anderen Weg realisiert werden können[3]. Ein ML-System ist durch eine Antwort auf die Frage nach seiner Funktionsweise („Wie?") charakterisiert. In Abb. 3 ist dies schematisch dargestellt.

3 Gemäß den Erfahrungen der letzten zehn Jahre.

Abbildung 3: Zum komplementären Verhältnis von ML und KI



Quelle: Commerzbank AG.

Insbesondere lassen sich Beispiele aufzählen von ML-Systemen, die nicht zu den KI-Anwendungen gezählt werden sollten, etwa bei der Verwendung von ML zum effizienten Lösen von Optimierungs- und Simulationsaufgaben in der Finanzmathematik oder der Vorhersage chemischer Bindungsstrukturen. Die zu lösende Aufgabe erfordert hier vermutlich kein semantisch tiefes Erkennen von neuen Konzepten, sondern nur eine effiziente Interpolation von Datenpunkten.

Die Diskussion der Begriffe KI und ML hat nicht nur eine philosophische Dimension. Sie ist vielmehr auch die Grundlage für eine in den nächsten Jahren zu entwickelnde Regulierung. Während sich die besonderen Anforderungen an eine KI-Entwicklung und Anwendung aus der Frage nach dem Ziel (dem „Wozu?“) ergeben müssen – nämlich dem Ersetzen menschlichen Handelns und Entscheidens durch automatisierte Entscheidungen – muss sich ein Rahmenwerk für die Entwicklung von sogenannten „Vertrauenswürdigen KI-Anwendungen“ (Europäische Komission, 2019) anhand von objektive Kriterien formulieren lassen und damit an der technischen Umsetzung (das „Wie?“) über ML orientieren, wobei alle Verfahren, selbst regelbasierte Entscheidungsverfahren wiede-

rum gleich behandelt werden müssen. Die klare Unterscheidung und adäquate Berücksichtigung beider Aspekte wird grundlegend sein für das Gelingen einer KI-Regulierung. Im letzten Abschnitt werden wir hierauf noch näher eingehen.

4 Technische Grundlage: Datenhaushalte für KI-Anwendungen

Wir haben bereits gesehen, wie wichtig der „Big-Data"-Effekt für die Entwicklung von KI ist. Da der Begriff „Lernen" im Wesentlichen ein Synonym für die statistisch-geometrische Interpolation von Beispieldaten ist, ist intuitiv klar, dass der Schlüssel zu erfolgreichen KI- bzw. ML-Anwendungen in guten Daten in großer Menge liegt. In diesem Abschnitt sollen einige Hinweise für die Einrichtung eines die praktische Anwendung von KI grundlegenden Big-Data-Haushalts gegeben werden.

Wiederum können hier keine für den technisch versierten Leser interessanten Detail vermittelt werden, trotzdem ist es auf der anderen Seite für den fachlich Interessierten sehr hilfreich, die entscheidenden, nicht immer intuitiven Begriffe aus dem Bereich der nicht-SQL-basierten Datenaufbereitung zu kennen und grundsätzlich richtig zu verstehen – und vor allen zu erkennen, wie eng KI und Big Data verknüpft sind.

4.1 Was ist Big Data – für Finanzunternehmen?

Eine klassische Definition für Big Data besagt, sie läge vor „wenn Daten Teil des Problems werden". Schlichter ausgedrückt kann man sagen, Big Data sind Daten, aus denen man die in ihnen enthaltene Information nicht direkt ablesen kann – und somit ist auch die Brücke zur Datenanalytik und KI unmittelbar geschlagen, denn letztere sind genau die Verfahren, die relevante Informationen und Beziehungen in Daten selbstständig extrahieren können.

Zur Beschreibung der Herausforderungen bei der Nutzung von Big Data wird häufig die (vom Englischen abgeleitete) Systematik der „Vs" genutzt. Drei davon beschreiben technische Aspekte, die für Finanzunternehmen auf jeden Fall relevant sind: Große Datenmenge (Volume), hohe Erzeugungs-Frequenz (Ve-

locity) und vielfältige Datenformate (Variety). Abb. 4 listet diese Kriterien und Beispiele aus der Bankenwelt auf.

Abbildung 4: Technische Charakteristika von Big Data in Banken

- **V**olume *(PB an Kunden-, Transaktions- und Finanzdaten)*
- **V**elocity *(z. B. viele Millionen Zahlungstransaktionen täglich)*
- **V**ariety *(Kreditverträge, Kunden-Datenbanken, Finanzdaten,...)*

Quelle: Commerzbank AG.

Daneben gibt es auch inhaltliche Aspekte, die häufig mit den verbundenen Begriffen Veracity für die Korrektheit und Value für die Wertigkeit der enthaltenen Informationen angegeben werden. Auch auf der inhaltlichen Seite ist jedoch ein dritter Aspekt unbedingt zu den Charakteristika von Big Data zu zählen: Der der Vernetzung (englisch wäre Verticality passend) (s. Abb. 5).

Abbildung 5: Inhaltliche Charakteristika von Big Data in Banken

- **V**eracity *(veraltete Dokumente, Schlüsseltabellen...)*
- **V**alue *(Informationen in unstrukturierten Gesprächsprotokollen)*
- **V**erticality *(KYC-Informationen zu einem Kunden in diversen Quellen)*

Quelle: Commerzbank AG.

Während in klassischen Data Warehouses für Datensätze, die in einer Beziehung stehen, in aller Regel explizite Beziehungstabellen mit sogenannten Primär- und Fremdschlüsseln zur eindeutigen automatischen Verknüpfung vorhanden sind, enthalten Big-Data-Haushalte oft eine Vielzahl unterschiedlicher Datenquellen mit Informationen zu den gleichen unterliegenden Objekten, ohne dass eine

Zusammenführung dieser Informationen ohne weiteres möglich wäre, da sie z. B. nur durch Namen in unterschiedlichen Schreibweisen identifiziert werden können. Dies ist insbesondere bei unstrukturierten Daten (siehe unten) der Fall.

Die Extraktion und Nutzbarmachung von solchen „vertikalen" Beziehungen[4], also Beziehungen zwischen Datensätzen ohne eindeutige Schlüssel, ist entscheidend für den Erfolg von Big-Data-Analysen und deren Nutzung zur Entwicklung von KI-Anwendungen mit Maschinellen Lernen.

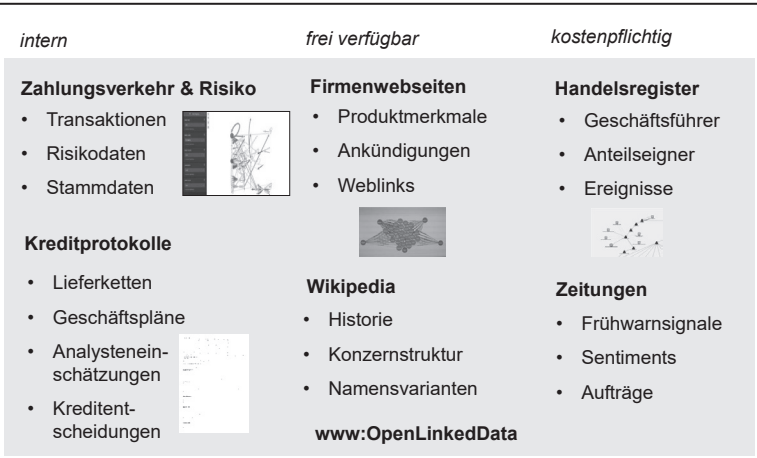
4.2 Interne und externe Datenquellen

Auch heute wird immer noch gelegentlich die Frage gestellt, ob es im eigenen Haus überhaupt genug Daten gibt, als dass eine Beschäftigung mit Themen wie Big Data oder KI lohnen würden. Daher findet sich in Abb. 6 eine Auswahl von internen und externen Datensätzen, die in jedem Finanzinstitut von Nutzen sein sollen.

Der Leitgedanke bei der Auswahl von Quellen sollte dabei immer sein, bewusst Informationslücken zu schließen und Aspekte beleuchten zu können, die typischerweise schwer – oder erst spät – zugänglich sind.

4 In Datenbanktabellen werden Datensätze zu unterschiedlichen fachlichen Objekten i. d. R. untereinander (in Zeilen) dargestellt, während die einzelnen Attribute zu den Objekten horizontal nebeneinander (in Spalten) stehen.

Abbildung 6: Beispiele für interne und externe Datenquellen



Quelle: Commerzbank AG.

Als klassisches Beispiel seien hier Zeitungsmeldungen erwähnt, die über Geschäftsvorhaben, Wirtschaftsbeziehungen oder auch mögliche Verwicklungen in kriminelle Aktivitäten zeitnah Informationen enthalten können. Solche Daten können i. d. R. käuflich erworben werden. Als Beispiel für interne Quellen ließen sich Protokolle von Beratungsgesprächen anführen, die oft risikorelevante Informationen in unstrukturierter Form enthalten, aber i. d. R. nur wenigen Personen (Beratern) zugänglich sind.

4.3 Unstrukturierte Datenquellen

Allen genannten Beispielen ist gemein, dass es sich um sogenannte unstrukturierte Daten handelt, die im Big-Data-Bereich eine sehr große Rolle spielen und letztlich auch für die enge Beziehung von Big Data und KI verantwortlich sind. Was genau ist darunter zu verstehen?

Erinnern wir uns zunächst daran, was gemeinhin unter strukturierten Daten verstanden wird. Diese sind typischerweise in Tabellen mit Schlüsselspalten zur

Identifikation der Datenobjekte und Attributspalten für die einzelnen Merkmale abgelegt. Die Wertausprägungen in jeder der Spalten haben jeder für sich eine individuelle Bedeutung, die im Prinzip maschinell verarbeitet werden kann. Insbesondere haben nicht-numerische Attribute nur endlich viele Ausprägungen (Kategorien) (vgl. Abb. 7).

Abbildung 7: Typische strukturierte Daten[6]

Strukturierte Daten

- Tabellen mit Zeilen (Datenobjekte) und Spalten (IDs und semantische Attribute)
- Attributs**ausprägungen** haben individuelle Semantik („endlicher" Wertebereich)

Kunden-ID	Name	Rating	Wohnort
Kunde_123	Schulze, Karl	2.4	Berlin

Quelle: Commerzbank AG.

Im Gegensatz dazu ist es für unstrukturierte Daten typisch, dass sie numerische oder nicht-numerische Attribute enthalten, die keine individuelle, maschinell interpretierbare Bedeutung haben, weder die einzelnen Ausprägungen noch (in einigen Fällen) die einzelnen Spalten. In der Regel ist die Anzahl möglicher Ausprägungen nicht-numerischer Attribute auch praktisch unbeschränkt, so dass es unmöglich ist, eine unmittelbare maschinelle Interpretation durch eine entsprechende Zuordnungsstabelle (Ausprägung -> Bedeutung) vorzunehmen, s. Abb. 8.

In sehr vielen Fällen handelt es sich bei diesen Attributen entweder um nichtnumerischen Freitext, oder aber um durch Pixeldaten beschriebene Bilddaten. Jedoch können auch Sensormassendaten darunterfallen oder in gewisser Hinsicht z. B. auch Telefonverbindungsdaten. Die entscheidende Frage ist, ob man sowohl dem einzelnen Attribut als auch der einzelnen Ausprägung unmittelbar eine Bedeutung zuordnen kann. Damit hängt die Einstufung also immer vom Analyseziel ab. Nutzungsstatistiken lassen sich i. d. R. maschinell erheben, die Erkennung ungewöhnlicher (vielleicht auf einen Defekt oder Betrug hindeutenden) Datensätze dagegen nicht.

Die Vorstellung, unstrukturierte Daten könnten nur in flachen Dateisystemen und nicht in Datenbanken gespeichert sein, ist dagegen nicht korrekt und nicht wesentlich. Allerdings empfiehlt es sich durchaus, spezielle Speichersysteme zu nutzen.

Abbildung 8: Typische unstrukturierte Daten

Unstrukturierte Daten

- Sammlung von Dokumenten mit Keys (Document ID) und Values (Dokument)
- Attributs**ausprägungen** einzigartig (Text, Bilder) („unendlicher" Wertebereich) ohne maschinell zuordenbare Semantik

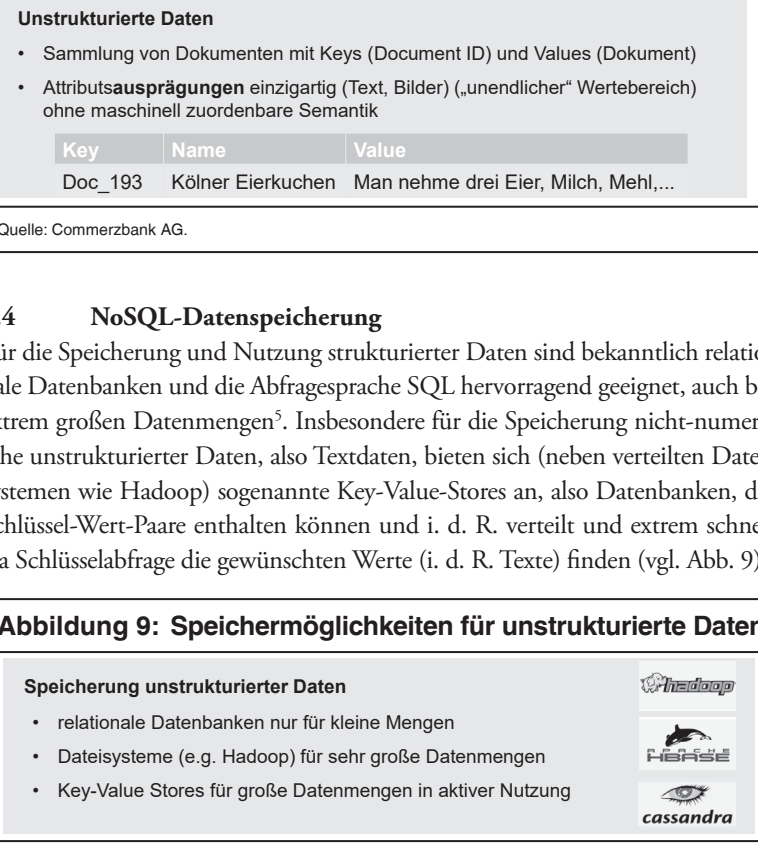
Key	Name	Value
Doc_193	Kölner Eierkuchen	Man nehme drei Eier, Milch, Mehl,...

Quelle: Commerzbank AG.

4.4 NoSQL-Datenspeicherung

Für die Speicherung und Nutzung strukturierter Daten sind bekanntlich relationale Datenbanken und die Abfragesprache SQL hervorragend geeignet, auch bei extrem großen Datenmengen[5]. Insbesondere für die Speicherung nicht-numerische unstrukturierter Daten, also Textdaten, bieten sich (neben verteilten Dateisystemen wie Hadoop) sogenannte Key-Value-Stores an, also Datenbanken, die Schlüssel-Wert-Paare enthalten können und i. d. R. verteilt und extrem schnell via Schlüsselabfrage die gewünschten Werte (i. d. R. Texte) finden (vgl. Abb. 9).

Abbildung 9: Speichermöglichkeiten für unstrukturierte Daten



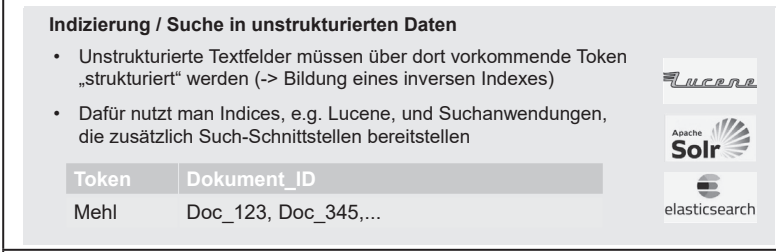
Quelle: Commerzbank AG.

5 Selbst native Big-Data-Technologien wie Apache Spark oder Apache Hive nutzen zur Analyse strukturierter Daten SQL-Dialekte.

Typischerweise werden (Text-)Daten aber, wenn sie bezüglich der zu lösenden Analyseaufgabe als unstrukturiert im obigen Sinne gelten sollen, nicht nur über wenig aussagekräftige Schlüsselwerte gesucht werden können. Die Analyse wird vielmehr drauf beruhen, Inhalte in den Texten aufzuspüren[6], indem den unendlich vielen möglichen Ausprägungen eines Textes (z. B. eines Zeitungsartikels) eine oder mehrere von endlich vielen, wohldefinierten Kategorien zugeordnet werden.

Als Vorstufe hierzu können die sogenannten Suchindizes gelten, die zur Speicherung unstrukturierter Daten unbedingt dazugehören. Sie können mithilfe gängiger Anwendungen (Suchmaschinen) erstellt werden (s. Abb. 10).

Abbildung 10: Indizierung unstrukturierter Daten: Gängige Suchmaschinen



Quelle: Commerzbank AG.

Die zentrale Rolle spielt hierbei ein sogenannter inverser Index, der im Prinzip nichts anderes darstellt als ein Stichwortverzeichnis für den gesamten Textdatenkorpus. Das heißt, für jedes in den indizierten Dokumenten vorkommende Wort wird einmal vorab die Liste aller Dokumente (bzw. deren Schlüssel) ermittelt, die dieses Wort enthalten. So können im Anschluss Suchbegriffe und logische Kombinationen aus Suchbegriffen schnell über den Index und die Dokumentenschlüssel gefunden werden.

6 Zur Vereinfachung bleiben wir hier beim Beispiel der Textanalyse, für Bild- oder Transaktionsdaten könnte man ähnlich vorgehen.

Abstrakt ausgedrückt, findet über den inversen Index eine Prä-Kategorisierung und damit Prä-Strukturierung der Textdaten statt: Für jedes in den Texten vorkommende Wort w werden die zwei Kategorien „enthält w“ und „enthält w nicht“ gebildet, und über den Index wird jedem Dokument für jedes Wort eine von diesen Kategorien zugeordnet.

Fasst man später diese Basis-Wortattribute zu sinnvollen Einheiten zusammen (also etwa „enthält ‚Betrug‘ oder ‚Verbrechen‘ oder ‚kriminell‘“), kann man sehr bequem sinnvolle Kategorien zusammensetzen, die für das zu lösende Problem (hier Signale für Straftatenbezug) relevant sind und einen eindeutige Bedeutung haben. Insbesondere können solche Kombinationen durch ML automatisch gefunden werden. Voraussetzung dafür ist jedoch, die Begriffe in den Texten über den Index schnell lokalisieren zu können.

4.5 Entitiy Recognition, Entity Linkage und Datenintegration

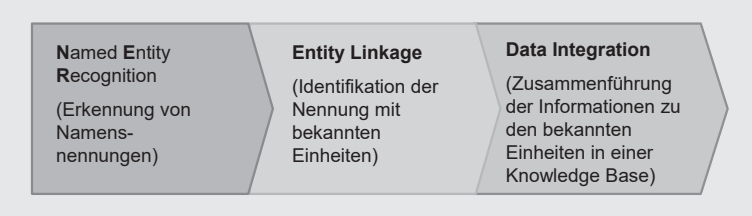
Wie bereits erwähnt, spielen Namen oder allgemeiner gesprochen, benannte Einheiten (Named Entities) sowie Datenfelder, die Information zu diesen enthalten (etwa Namensfelder), in unstrukturierten Daten eine besondere Rolle.

Das hat damit zu tun, dass es – zumindest im Finanzkontext – sehr häufig darum geht, aus Big Data Informationen aus unterschiedlichen Quellen, z. B. internen und externen Quellen, zusammenzuführen. Häufig sind diese unstrukturiert, sodass Bezüge zu den im Text behandelten Einheiten (Personen, Firmen, Orte, Ereignisse, Organisationen) nur durch nicht normierte Namen gegeben sind.

Die im Text erwähnten Entitäten können als spezielle Kategorien angesehen werden, sie sind jedoch weit wichtiger, da sie Beziehungen zwischen den einzelnen Dokumenten herstellen oder, in umgekehrter Betrachtungsweise, die Dokumente Beziehungen zwischen den Einheiten herstellen. Beziehungen zwischen Personen, Firmen usw. und Informationen darüber sind im Finanzsektor von enormer Bedeutung. Selbst die Feststellung, dass zwei Personennennungen dieselbe natürliche Person betreffen, kann ein wichtiger Befund in eine KYC-Prüfung sein.

Die Erkennung von Entitäten (Named Entity Recognition, NER), die korrekte Zuordnung zu bekannten Identitäten (Entity Linkage) und die damit mögliche Integration von Datenquellen in einer identitätsbasierten Wissensbasis (Knowledge Base) stellen zusammen zentrale Techniken für die Erschließung von Big Data dar (s. Abb. 11).

Abbildung 11: Schematischer (Teil-)Prozess der Informationsextraktion

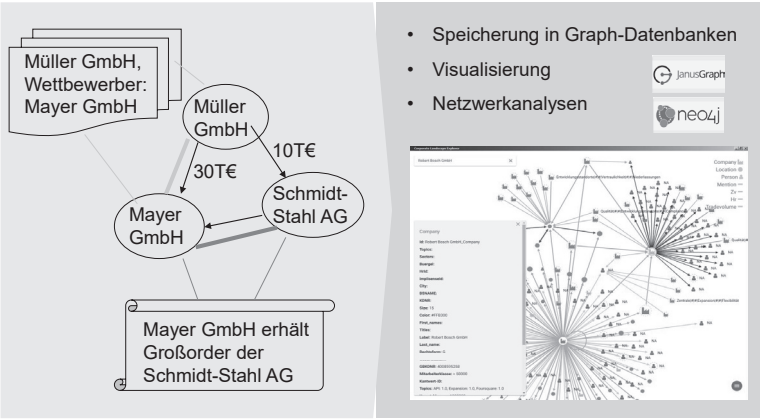


Quelle: Commerzbank AG.

Für die Lösung dieser Aufgaben steht heutzutage ein umfangreiches Arsenal an Methoden und quelloffenen Implementierungen zur Verfügung, die, wenig überraschend, so gut wie immer auf Maschinellem Lernen oder bereits vortrainierten ML-Modellen z. B. für die Textanalyse oder Themenerkennung beruhen. Eine nähere Beschreibung dieser Verfahren geht über den Rahmen dieses Betrags selbstverständlich weit hinaus, wir verweisen daher auf die Literatur.

Es soll aber unterstrichen werden, dass es sich dabei nicht etwa um hochkomplexe oder teure Lösungen handelt, sondern in aller Regel jeder gute Absolvent einer quantitativen Fachrichtung in der Lage ist, die entsprechenden quelloffenen Programme in wenigen Wochen selbstständig einzusetzen.

Abbildung 12: Datenintegration in Graph-Datenbanken



Quelle: Commerzbank AG.

Abschließend sei noch einmal darauf hingewiesen, wie die enge Verbindung zwischen Big Data und KI am Beispiel der Datenintegration zum Ausdruck kommt: KI ermöglicht erst die Zusammenstellung umfangreicher Big-Data-Datensätze, die viele Informationen zusammenführen, aus denen dann wiederum mithilfe von KI Informationen gewonnen werden können, die so nicht direkt vorliegen.

Dies kann man sich durch die Veranschaulichung der Datenstrukturen in einer Wissensdatenbank vor Augen führen (s. Abb. 12): Aus den Daten werden neben Attributen auch eine große Menge an Beziehungen abgeleitet, die erst durch die Vernetzung sinnvolle Einheiten bilden – z. B. ganze Lieferketten, die sich aus einzelnen Handelsbeziehungen zusammensetzen. Solche relevanten Strukturen können mit KI aufgespürt werden.

4.6 Exkurs: Datenschutz bei unstrukturierten Daten und durch Modelle generierten Daten

Ein letzter Aspekt zum Thema Datennutzung soll an dieser Stelle aufgrund seiner enormen Bedeutung noch kurz erwähnt werden: Die Rolle des Datenschutzes bei ML-Anwendungen mit Big Data, die PII (Personenbezogene Daten) enthalten. Da auch hier eine umfassende Behandlung den Rahmen sprengen würde, beschränken wir uns auf die folgende Aufzählung von besonders kritischen Aspekten, die in diesem Zusammenhang zu beachten sind.

- Um Prozesse wie NER und Entity Linkage effizient durchführen zu können, ist eine verwendungsunabhängige Vorverarbeitung (Preprocessing) unerlässlich. Dies kann im Gegensatz zur Zweckgebundenheit der Verarbeitung von PII stehen.
- Die Datenflusskette (Data Lineage) muss auch für die analytischen Verfahren lückenlos aufgezeigt werden – es ist zulässig, eine Entity Linkage für den KYC-Process ohne weiteres auch für die Vertriebsunterstützung zu nutzen.
- Beziehungen gehören zu den sensibelsten Daten, sie betreffen immer mehrere Beteiligte.
- Identifikations- und andere Beziehungen, die durch analytische Verfahren gewonnen werden, gelten oft nur mit einer gewissen Wahrscheinlichkeit. Diese sollte als Attribut der Beziehung verfügbar sein. Es ist nicht immer klar, wie dies datenschutzrechtlich zu behandeln ist. Es ist nicht klar, ob diese z. B. unter das Auskunftsrecht einer natürlichen Person fallen.
- Durch Verknüpfungen kann eine Anonymisierung leicht aufgehoben werden, insbesondere bei Zeitreihendaten.
- Bei Transaktionsdaten sind zwar nur die Buchungstexte besonders geschützt, aber die Transaktionspartner enthalten unter Umständen ebenso viel Information.

- Auch Daten, die keine PII enthalten, sind personenbeziehbar, z. B. durch Anwendung eines Scoring-Modells auf eine Person zur Ermittlung einer Ausfallwahrscheinlichkeit.

5 Ethische Richtlinien

Mit den Anmerkungen zum Thema Datenschutz im Kontext KI haben wir bereits den Blick auf das letzte Thema dieser Übersicht gerichtet: Die Leitplanken für eine nicht nur technisch erfolgreiche, sondern auch ethisch einwandfreie Nutzung von KI. Wie kommt es, dass immer wieder Zweifel an der Konformität von KI mit ethischen Grundsätzen geäußert werden?

5.1 Der Ursprung besonderer ethischer Anforderungen an KI

Für die Beantwortung dieser Frage ist es wieder ganz entscheidend, das Problem aus dem richtigen Blickwinkel zu betrachten. Man kann und muss sich sicher auch über die technische Verlässlichkeit von ML-Modellen Gedanken machen oder diese durch Messgrößen nachweisen, ohne damit aber den Kern des Problems zu erfassen.

Es ist die Natur von KI-Systemen, menschliche Entscheidungen zu automatisieren, die die Bedenken hervorruft. Und diese sind zumindest solange berechtigt, wie wir es mit „schwachen", nicht-allgemeinen KI-Systemen zu tun haben – mit anderen Worten, auf unabsehbare Zeit. Denn diesen Systemen fehlt die (prinzipielle) menschliche Möglichkeit, kurzfristig aus den gelernten „Modellstrukturen[7]" auszubrechen und neue Aspekte in die Entscheidung einfließen zu lassen, die das „Modell" nicht vorsieht.

Ob dies tatsächlich immer oder auch im langfristigen Mittel überhaupt wünschenswert ist, kann angezweifelt werden[8]. Fakt ist jedoch, dass es viele Beispiele gibt für Situationen, die ein kurzfristiges Umdenken erfordern, beispielsweise auch beim Autonomen Fahren in unbekanntem Gelände, bei Einstellungsverfahren oder Kreditentscheidungen. Das bedeutet für KI-Anwendungen, dass sie

7 Wir verwenden den Begriff hier einfach etwas unpräzise als das, was dem KI-System zugrunde liegt, z. B. (aber nicht zwingend) ein ML-Modell.

8 Beispielsweise haben Piloten schon seit Jahren nicht mehr die Möglichkeit, sich über die durch den Autopiloten ermittelten Grenzen für zulässige Flugparameter hinwegzusetzen.

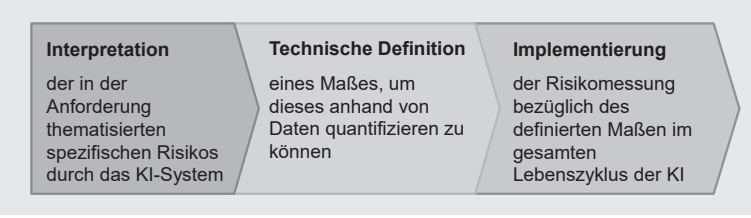
die Grenzen ihrer Anwendbarkeit kennen und deren Überschreitung erkennen müssen. Dies kann nur auf der Ebene der technischen Umsetzung, also der ML-Modellierung erfolgen.

Neben dieser ersten konkreten Schwachstelle heutiger KI ist deren Anwendung auch mit dem unbewussten Wissen darum konfrontiert, dass sich Konzepte wie Fairness oder Transparenz gar nicht eindeutig objektivieren lassen – obwohl sie als objektiv unstrittig angesehen werden! Dies ist ein tieferliegendes philosophisches Dilemma, das natürlich durch KI nicht aus der Welt geschafft werden kann. Hier kann eine Verteidigungsstrategie nur darin bestehen, die Objektivierung der Anforderungen bewusst einzufordern – und damit wiederum auf die Ebene der Umsetzung von KI in Form von ML-Modellen zu wechseln, auf der quantitative Kriterien formuliert und objektiv geprüft werden können.

5.2 Von den Ethik-Richtlinien für Vertrauenswürdige KI zu implementierbaren KI-Policies

Setzt man sich intensiv mit den von der hochrangigen Expertengruppe der Europäischen Kommission erarbeiteten Leitlinien für vertrauenswürdige KI (Europäische Komission, 2019) auseinander, wird man die soeben aufgezählten zwei Motivationspunkte für ein Misstrauen gegenüber KI-Anwendungen mühelos nachweisen können. Die Arbeit gibt zwar keinerlei Hinweise für eine praktische Lösung der aufgezeigten Herausforderungen, sie bildet aber einen exzellenten Ausgangspunkt für die Entwicklung von KI-Rahmenwerken mit Fokus auf objektive Messbarkeit und Implementierbarkeit – ganz im Sinne der Erkenntnisse aus dem letzten Anschnitt.

Abbildung 13: Mögliches Vorgehen zur Erstellung einer KI-Policy



Quelle: Commerzbank AG.

Die Expertengruppe hat ihre Leitlinien anhand von vier Grundprinzipien entwickelt, die in Abb. 14 dargestellt sind.

Abbildung 14: Grundprinzipien Vertrauenswürdiger KI der hochrangigen Expertengruppe der Europäischen Kommission

I. Achtung der menschlichen Autonomie
II. Schadensverhütung
III. Fairness
IV. Erklärbarkeit

Quelle: Commerzbank AG.

Das erste Prinzip wurde implizit bereits in den einleitenden Bemerkungen des letzten Abschnitts kommentiert. Die wichtigste Maßnahme für die Erhaltung der menschlichen Autonomie besteht im Grunde darin, umfassendes Wissen zur Natur und den Grenzen von KI-Systemen bei Entwicklern, Betreibern und Nutzern aufzubauen, ein Ziel, dem dieser Beitrag verpflichtet ist.

Im Folgenden sollen daher noch die drei anderen Prinzipien erörtert werden, insbesondere ihre oben eingeforderte Objektivierung.

5.3 Schadensverhütung – Validierung und Modellrisiko

Unter Schadensverhinderung im Sinne der Richtlinien kann im Wesentlichen verstanden werden, dass ein KI-System durch falsche Entscheidungen beziehungsweise falsche Entscheidungsempfehlungen seinem Nutzer – oder auch der Gesellschaft als Ganzes oder der natürlichen Umwelt – Schaden zugefügt wird.

Wiederum wird an dieser Stelle deutlich, wie wichtig die Unterscheidung des KI-Begriffs und ihre technischen Umsetzung ist. Da alle leistungsfähigen KI-Systeme heute auf ML, also statistischen Verfahren beruhen, ist offensichtlich, dass eine hundertprozentige Schadensvermeidung auch bei perfekter Funktion der KI-Anwendung nicht möglich ist.

Ganz neu ist dieser Umstand nicht: So gibt es etwa im medizinischen Bereich viele Therapien und Medikamente, die unter Umständen beim Einzelnen wirkungslos oder sogar schädlich sein können. Der Grundsatz der Schadensvermeidung kann daher auch für KI-Systeme nur sinnvoll interpretiert werden im Sinne einer klassischen Risikobewertung als Kombination aus deutlich positivem erwarteten Nutzen und gleichzeitiger Begrenzung von zu definierenden maximalen Schäden auf einem bestimmten Vertrauensniveau[9]. Mit anderen Worten, es handelt sich um klassische Modellrisikosteuerung – und eine gute Modellvalidierung sollte dafür die Grundlage bilden.

5.4 Fairness und Bias

Nachdem mit den beiden ersten Prinzipien sichergestellt werden soll, dass KI weder katastrophale Kontrollverluste noch überhaupt negative Auswirkungen für die Einzelnen haben kann, werden mit dem dritten Prinzip die Anforderungen noch einmal höher geschraubt: Es reicht hier nicht mehr aus, dass KI für alle Vorteile bringt, es sollten auch alle in gleichem Maße profitieren.

9 Beispielsweise kann man (vereinfacht) anstreben, dass eine Behandlung für 70 % der Patienten einen Heileffekt haben sollte, wobei starke negative Nebenwirkungen in höchstens 10–4 % aller Fälle auftreten dürfen.

Eine ähnliche Forderung ist bei kaum einer neuen Technologie je erhoben worden[10], und wird bereits früher angesprochen. Es ist ein ausgesprochen tiefgehendes philosophisches Problem, Fairness zu definieren. Daher ist es für professionelle KI-Anwender unbedingt notwendig, diesen Begriff im Rahmen einer eigenen KI-Strategie zunächst quantitativ zu fassen. Solange hierfür keine Vorgaben existieren, sind durchaus mehrere sinnvolle, aber i. d. R. unvereinbare Festlegungen möglich[11].

Wir wollen hier wieder nur einige grundlegende Begriffe vorstellen, um die Komplexität des Fairnessbegriffs zu illustrieren.

KI-Systeme sind dafür gebaut, jede statistisch relevante Differenzierungsfähigkeit in Bezug auf die Zielvariable in den verfügbaren Features zu finden und für die Prognose zu nutzen. Dies kann dazu führen, dass ein – handwerklich einwandfrei gebautes – Modell[12] ungleiche Prognosen für Datensubjekte liefert, die sich nur durch Attribute unterscheiden, die aus ethischen oder logischen Gründen keine Rolle spielen sollten.

- Ist der Grund logischer Natur, spricht man typischerweise von Verzerrung oder Bias im Modell.
- Handelt es sich ethische Gründe, wird man eher von mangelnder (bedingter) Fairness des Modells sprechen.
- Sowohl ein Bias als auch mangelnde Fairness können eine Reihe von Gründen haben und müssen vermieden werden.
- Manchmal wird der Begriff Fairness auch ohne die Voraussetzung verwendet, dass die Datensubjekte in allen anderen Attributen gleich sind. In diesem Zusammenhang sollte man von mangelnder unbedingter Fairness sprechen[13].

10 Hier soll lediglich die Größe der Herausforderung zum Ausdruck gebracht werden und keineswegs die Notwendigkeit bestritten werden, sich dieser zu stellen.
11 Siehe z. B. Fußnote 3.
12 Typischerweise wird man sich hier auf Modelle beziehen, die Argumentation gilt aber ebenso für alle anderen Verfahren, auch regelbasierte.
13 Die Unterscheidung von bedingter und unbedingter Fairness wird selten gemacht und führt regelmäßig zu Missverständnissen.

- Bias und mangelnde Fairness können bereits in den Trainingsdaten vorhanden sein, etwa dadurch, dass diese nicht repräsentativ sind oder auf menschlichen Entscheidungen unter Vorurteilen beruhen. Dies kann als empirischer Bias bzw. mangelnde empirische Fairness bezeichnet werden. Ein auf empirisch verzerrten oder mangelnder empirischer Fairness reflektierenden Daten gebauten Modell kann technisch einwandfrei sein und in aller Regel trotzdem verzerrt oder unfair.

Wichtig ist: Nur wenn logische oder ethische Gründe die Erwartung gleicher Modellergebnisse rechtfertigen, ist es legitim, von Bias oder mangelnder Fairness zu sprechen, sonst wäre statistische Modellierung nicht sinnvoll. Es kann sehr gute Gründe dafür geben, beispielsweise Produktangebote nach Alter zu differenzieren.

Ebenso muss darauf geachtet werden, dass sich Verzerrung und Fairness nicht nur auf die Gleichheit der Modellergebnisse beziehen darf, sondern auch auf Maße wir z. B. die statistische Genauigkeit des Modells, die für verschiedene Teilgruppen unterschiedlich sein kann. Hier muss entschieden werden, ab wann eine Ungleichbehandlung vorliegt, die zulasten einer Gruppe geht und gegebenenfalls die Verwendung unterschiedlicher Modelle erforderlich macht.

Grundsätzlich ist davon auszugehen, dass Bias und Fairness bei datengetriebenen Anwendungen nicht statisch sind, sondern sich über die Zeit entwickeln können.

Ein permanentes Monitoring ist daher neben einer Messung als Bestandteil der initialen Validierung notwendig. Dabei sollten nicht nur potenziell diskriminierende Attribute, sondern so viel Variablen wie möglich untersucht werden, für die sich logische Bedingungen zu ihrer Wirkung formulieren lassen. Dies wir eine großen Beitrag zur Robustheit des Modells leisten.

Um sicherzustellen, dass Bias und mangelnde Fairness vermieden werden, gibt es unterschiedliche Verfahren, und es sollte streng darauf geachtet werden, die jeweils für die aufgeführten Typen passenden Verfahren zu wählen[14]. Beispiels-

14 Es kann sogar passieren, dass verschiedene Fairnessziele inkompatibel sind. Vgl. John Kleinberg, Sendhil Mullainathan, Manish Raghavan (2016).

weise sollte mangelnde unbedingte Fairness, die nicht auf Modellschwächen oder Bias beruht, niemals durch eine Manipulation des Modells behoben werden, sondern vielmehr dadurch, dass die Modellergebnisse grundsätzlich über alle Ausprägungen potenziell diskriminierender Attribute gemittelt werden.

Eine offensichtliche Konsequenz dieses Verfahrens ist, quasi im Umkehrschluss, dass eine Diskriminierung bezüglich (z. B. aus Datenschutzgründen) nicht bekannter Attribute grundsätzlich nicht erkannt und verhindert werden kann (etwa bezüglich eines Migrationshintergrundes).

5.5 Erklärbarkeit und Transparenz

Der Erklärbarkeitsbegriff ist der vielleicht meistdiskutierte im Rahmen der Diskussion um den Einsatz moderner Datenanalyse. In ihm bündeln sich wie in einem Brennglas alle berechtigten und unberechtigten Bedenken, in Hoffnungen und philosophischen Rätsel um die Schaffung und Nutzung Künstlicher Intelligenz. Ein tiefes, korrektes Verständnis dieses Begriffs ist in der Tat grundlegend.

Ein ganz entscheidender Aspekt von Erklärbarkeit ist dabei, dass er sich notwendigerweise auf kausale Zusammenhänge beziehen muss. In der Tat ist ein rein statistischer Erklärbarkeitsbegriff problematisch und verwirrend. Dies wird sehr oft übersehen – in den Richtlinien der Expertengruppe passiert dies ebenfalls, allerdings wird hier die grundsätzlich statistische Natur von KI-Systemen ebenfalls übersehen.

Oft trifft man auf die Ansicht, dass beispielsweise ein regelbasiertes Entscheidungssystem im Gegensatz zu einem, ML-Modell deterministisch wäre. Dies ist ein Trugschluss.

Entscheidend dafür, ob ein System deterministisch ist oder stochastisch, ist zunächst nur die Frage, ob die Ausgabewerte reproduzierbar sind, beziehungsweise, ob zu gleicher Eingabe immer die gleichen Ausgabewerte produziert werden, was für beide Systemtypen der Fall ist. Wichtiger ist aber die Frage danach, ob das gelieferte Ergebnis, die vom System getroffene Entscheidung, systematisch korrekt ist. Und hier verhalten sich beide Systemtypen statistisch, sobald wir es

mit einem nicht trivial berechenbaren Ergebnis zu tun haben: Ihre Entscheidungen sind mal richtig, mal falsch, je nach Modellgüte.

Auch wenn in der Folge noch Möglichkeiten aufgezeigt werden, sich der Kausalität zu nähern, ist es aufgrund der für automatische Entscheidungssysteme heute noch immer begrenzten Informationslage unmöglich, in allen Situationen perfekt zu entscheiden. Genau dies war ja auch einer der Hauptbedenken gegen den Einsatz solcher Systeme: Sie sind nicht in der Lage, sich flexibel neue Informationen für den Einzelfall zu beschaffen.

In dieser Situation für ein KI-System zu fordern, dass die Entscheidungen erklärbar sind, ist praktisch unmöglich: Es würde bedeuten, die zum Ergebnis führenden Kausalketten gefunden zu haben. Damit würde jede falsche Entscheidung des Modells einen logischen Widerspruch bedeuten.

Alternativ könnte man akzeptieren, dass eine Erklärung in einem Fall eine richtige Entscheidung erklärt, aber in einem anderen Fall mit gleichen Eingangswerten eine falsche Entscheidung – das wäre offensichtlich absurd. Anders ausgedrückt: Wenn es, für einen gegebenen Satz an Eingabewerten, eine vollständige Erklärung für die Entscheidung gibt, muss diese in allen anderen Fällen ebenfalls zutreffen – die Prognosen des Modells wäre also perfekt.

Es ist daher sehr ratsam, den Begriff der Erklärbarkeit, wenn nicht ganz aus der Debatte zu verbannen, so doch wenigstens nur als Synonym zu benutzen für wesentlich treffendere Begriffe wie Transparenz und Rechtfertigung.

Transparenz beschreibt allgemein die Möglichkeit, die Berechnungen der Entscheidung nachzuvollziehen, rein formal, aber vor allem im Sinne beschränkter Komplexität. Dies ermöglicht es, von Modellen getroffene Entscheidungen zu plausibilisieren und bei Bedarf zu rechtfertigen. Im Gegensatz zur Erklärbarkeit setzt eine Rechtfertigung nicht voraus, dass die getroffene Entscheidung richtig war: Es reicht, dass es aus statistischer Sicht (und selbstverständlich unter ethischen Aspekten) naheliegend war, die Entscheidung so zu fällen.

Als klassisches Beispiel soll hier wieder ein Kreditwürdigkeits-Scoringmodell dienen, dass eine Prognose für den Ausfall eines Kreditnehmers berechnet und Kredite ablehnt, bei denen diese über einem Schwellwert liegt. Es wäre falsch

zu behaupten, man könne das Ergebnis einer Ablehnung dadurch erklären, dass z. B. der Kreditnehmer ein bestimmtes Alter hat. Ein direkter kausaler Zusammenhang zur höheren Ausfallwahrscheinlichkeit kann hier in im Einzelfall nicht hergestellt werden. Sehr wohl aber ist dieser Zusammenhang auf einer Gruppe ähnlicher Fälle statistisch nachweisbar, plausibel und darf daher als Rechtfertigung dienen, auch im Interesse junger Kunden höhere – und transparente – Hürden für eine Verschuldung aufzubauen.

Etwas anders liegt der Fall, wenn man das Monatsgehalt als „erklärenden" Faktor betrachtet. Kann der Kreditnehmer praktisch keine Einkünfte nachweisen, kann man durchaus von einem direkten kausalen Zusammenhang sprechen, auch wenn dieser (z. B. aufgrund fehlender Informationen) nicht zu 100 Prozent perfekt sein muss, denn die Rückzahlung des Kredites ist logisch unmöglich, wenn keine Einnahmen vorliegen.

Dieses Beispiel zeigt, dass es trotz der grundsätzlich (letztendlich aufgrund der immer unvollständigen Informationslage) statistischen Natur von KI-Systemen möglich ist, eine kausale Modellierung anzustreben. In der klassischen Ökonometrie ist es eine Grundregel der Modellaufstellung, nur Faktoren mit einem (plausiblen) kausalen Zusammenhang in das Modell aufzunehmen. Beim Einsatz von ML-Methoden, die die Featureerzeugung selbst durchführen, ist dies komplexer und die seit über zehn Jahren bestehenden Forschungsanstrengungen im Bereich des kausalen Lernens (Pearl, 2009) (Jonas Peters, Dominik Janzing, Bernhard Schölkopf, 2017) sind noch nicht am Ziel. Dennoch konnten einige aufschlussreiche Ergebnisse erzielt werden (Yoshua Bengio et al., 2019).

Die Entwicklung kausaler Methoden wird für die Nutzung von KI-Systemen für kritische Bereiche und insbesondere für zeitliche Prognosemodelle von zentraler Bedeutung sein. Ein sicheres Indiz dafür ist die steigende Zahl von wissenschaftlichen Nachweisen, dass insbesondere ML-Modelle gezielt angegriffen werden können, um das Verhalten des Systems durch minimale Änderungen der Eingabedaten zu manipulieren. So konnte etwa die Erkennung eines Verkehrsschildes durch einen an der richtigen Stelle angebrachten kleinen Aufkleber beeinflusst werden.

Viele Verfahren der sogenannten „Erklärbaren KI“ beruhen darauf, die Features mit dem größten Einfluss auf das Ergebnis herauszufinden (z. B. Variable Importances, Shapley values etc.). Bei sehr komplexen Modellen kann versucht werden, ein einfacheres Modell in der Umgebung eines konkreten Eingabewertsatzes zu finden, dass das komplexe Modell approximiert, aber einfacher und leichter zu interpretieren ist, z. B. nur lineare Zusammenhänge benutzt, wie etwa LIME (Ribiero).

Diese und viele weitere Verfahren sind sehr gut geeignet, Modelle zu plausibilisieren, aber vor allem Schwächen aufzudecken, etwa indem die Teile eines Bildes oder Textes, die für ein Bild- oder Texterkennungssystem besonders wichtig waren, hervorgehoben werden. Damit tragen sie zur Transparenz der Modelle bei.

Bei der Interpretation ist dennoch höchste Vorsicht geboten: In aller Regel ergeben sich nur selten wirklich eindeutige Muster. Sehr häufig entsteht die Sinnhaftigkeit der Interpretation erst durch die selektive Betrachtung des menschlichen Betrachters, also unter Zwischenschaltung einer subjektiven menschlichen Intelligenz. Die öffentlich bekannten Bilder sind i. d. R. sorgfältig ausgewählt und nicht repräsentativ für die Praxis. Letztlich ist auch hier wieder der Mensch selbst die größte Fehlerquelle. Solange man sich jedoch dessen bewusst ist, wird die Kombination aus maschineller Objektivität, menschlicher Intuition und Flexibilität sehr erfolgreich sein.

Bibliographie

Bishop, C. M. (2006): Pattern recognition and Machine Learning, Springer Science.

Europäische Kommission (2019): Ethik-Leitlinien für Vertrauenswürdige KI, Europäische Kommission, Brüssel.

Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016): Deep Learning, in: The MIT Press.

Ian Witten et al. (2017): Data Mining, Practical Machine Learning, Tools and Techniques (4th edition), Morgan Kaufmann.

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan (2016): Inherent Trade-Offs in the Fair Determination of Risk Scores, abrufbar unter http://www.arxiv.org/abs/1609.05807v2.

Jonas Peters, Dominik Janzing, Bernhard Schölkopf (2017): Elements of Causal Learning, in: The MIT Press.

Murphy, K. P. (2012): Machine Learning, A Probabilistic Perspective, in: The MIT Press.

Pearl, J. (2009): Causality, 2nd edition, in: Cambridge University Press.

Ribiero, M. (ohne Datum): LIME, abrufbar unter https://www.github.com/marcotcr/lime.

Stuart Russel, Peter Norvig (2010): Artificial Intelligence, A modern approach (3rd edition), Prentice Hall.

Trevor Hastie et al. (2009): The Elements of Statistical Learning (2nd edition), Springer Science.

Yoshua Bengio et al. (2019): A Meta-Transfer Objective for Learning to Disentangle Causal, abrufbar unter https://www.arxiv.org/abs/1901.10912.