St. Pölten University of Applied Sciences

/informatik & security

/fh/// st.pölten

Erklärbare Künstliche Intelligenz

Tools und Methoden für erklärbare KI-Modelle

Diplomarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

eingereicht von

Katrin Andrae
is181845

im Rahmen des
Studiengangs Information Security an der Fachhochschule St. Pölten

Betreuung
Betreuer/Betreuerin: FH-Prof. Mag. Dr. Simon Tjoa

St. Pölten, 18.01.2021

(Unterschrift Autorin)

(Unterschrift Betreuer)

Ehrenwörtliche Erklärung

Ich versichere, dass

- ich diese Diplomarbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich sonst keiner unerlaubten Hilfe bedient habe.
- ich dieses Diplomarbeitsthema bisher weder im Inland noch im Ausland einem Begutachter/einer Begutachterin zur Beurteilung oder in irgendeiner Form als Prüfungsarbeit vorgelegt habe.
- diese Arbeit mit der vom Begutachter/von der Begutachterin beurteilten Arbeit übereinstimmt.

Der Studierende/Absolvent räumt der FH St. Pölten das Recht ein, die Diplomarbeit für Lehre- und Forschungstätigkeiten zu verwenden und damit zu werben (z.B. bei der Projektevernissage, in Publikationen, auf der Homepage), wobei der Absolvent als Urheber zu nennen ist. Jegliche kommerzielle Verwertung/Nutzung bedarf einer weiteren Vereinbarung zwischen dem Studierenden/Absolventen und der FH St. Pölten.

St. Pölten, 18.01.2021

(Unterschrift Autorin)

Zusammenfassung

Künstliche Intelligenz (KI) hat sich in den letzten Jahren weltweit zu einem strategisch und wirtschaftlich relevanten Faktor entwickelt. So existiert kaum ein Bereich des modernen Lebens, der nicht von KI-basierten Technologien transformiert wird: sei es die Güterproduktion im Kontext der Industrie 4.0, die Prozessautomatisierung in Fertigung und Montage mit selbstregulierenden Steuerungsparametern, die Diagnostik mit KI-Assistenten im Gesundheitswesen oder die Steuerung von autonomen Fahrzeugen. Vor allem durch die gestiegene Rechenleistung und die Verfügbarkeit großer Datenbanken sind diese KI-Systeme in der Lage, eine gute Leistung bei zunehmend komplexen Aufgaben zu erzielen (teilweise auch das menschliche Niveau übertreffend).
An der vordersten Front bei dieser Entwicklung stehen die sogenannten Deep Learning Modelle. Aufgrund ihrer nicht linearen und verschachtelten Struktur werden diese leistungsstarken Modelle jedoch im Allgemeinen als "Black-Boxes" betrachtet, die keine Informationen über ihr Verhalten und ihre Entscheidungsfindung liefern. Da eine solche Intransparenz bei zahlreichen Anwendungen, wie z. B. im medizinischen Bereich, nicht akzeptabel ist, hat die Entwicklung von Tools und Methoden zur Erklärung von Deep Learning Modellen in den letzten Jahren zunehmend an Bedeutung gewonnen.
Das Ziel der vorliegenden Arbeit ist es, die relevantesten dieser Verfahren und Tools zur Sicherstellung der Erklärbarkeit, Interpretation und Visualisierung von Deep Learning Modellen und anderen KI-Modellen vorzustellen. So werden einerseits Erklärbarkeitstechniken präsentiert und näher beschrieben, durch die das Verhalten bzw. die Entscheidungsfindung von KI-Modellen des überwachten Lernens post hoc visualisiert bzw. interpretiert werden kann. Andererseits wird im Rahmen der Arbeit ein neuartiges Prüfrahmenwerk für erklärbare KI vorgestellt. Mit Hilfe dieses Rahmenwerks können Auditor/-innen bewerten, ob angemessene Verfahren, Prozesse und Kontrollen im Unternehmen implementiert sind, um die Erklärbarkeit des KI-Systems über dessen gesamten Lebenszyklus hinweg sicherzustellen.

Abstract

In recent years, artificial intelligence (AI) has become a strategically and economically relevant factor worldwide. Thus, hardly any area of modern life exists that is not transformed by AI-based technologies: for example the production of goods in the context of Industry 4.0, process automation in manufacturing and assembly with self-regulating control parameters, diagnostics with AI assistants in healthcare, or autonomous driving. Primarily due to increase in computing power and the availability of large databases, these AI are able to perform very well on increasingly complex tasks.
Leading this development are so-called deep learning models. Due to their non-linear and nested structure, these models are largely considered “black boxes” that do not provide information about their inner behaviour and decision making. Such intransparency is inacceptable in some applications and has led to an increase in the importance of developing tools and methods to explain, interpret and visualize deep learning models over the past few years.
The goal of this work is to present the most relevant of these techniques and tools for ensuring the explainability, interpretation, and visualization of deep learning models and other AI models. On the one hand, explainability techniques are presented through which the behaviour or decision making of AI models of supervised learning can be visualized or interpreted post hoc. On the other hand, a novel audit framework for explainable AI is presented in the following work. Using this framework, auditors can assess whether appropriate procedures, processes, and controls are implemented in the organization to ensure the explainability of the AI system throughout its lifecycle.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

1.1. Problemstellung und Zielsetzung

Künstliche Intelligenz (KI) hat sich in den letzten Jahren zu einer Schlüsseltechnologie für Industrie und Wirtschaft entwickelt [1]. Es wurden zahlreiche neue autonome Systeme in allen Wirtschafts- und Industriebereichen geschaffen, die selbst wahrnehmen, lernen, handeln und Entscheidungen treffen können [2]. Die großen Erfolge im Bereich der KI konnten insbesondere durch eine Verbesserung der KI-Methodik [3], eine erhöhte Rechenleistung [4] und die Verfügbarkeit stetig wachsender Datenbanken [5] erzielt werden. Diese Faktoren ermöglichen es den heutigen Algorithmen, eine gute Leistung bei zunehmend komplexen Aufgaben zu erzielen [1].

Trotz des revolutionären Charakters dieser Technologie bestehen zahlreiche Herausforderungen, die eine Verbreitung von KI in Anwendungen verlangsamen oder zum Teil sogar behindern [1]. Bespielhafte Herausforderungen sind die hohe Komplexität aktueller Deep-Learning-Modelle und der damit verbundene erhöhte Energiebedarf, der einen Einsatz in ressourcenbeschränkten Umgebungen verhindert, bzw. die Anfälligkeit von KI gegenüber gegnerischen Angriffen. Letztere können bei zahlreichen Anwendungsfällen von KI, z. B. im Bereich des autonomen Fahrens, zu einem erhöhten Sicherheitsrisiko für die Anwender/-innen führen [6].

Die mitunter größte Herausforderung von KI stellen jedoch die mangelnde Transparenz und Erklärbarkeit von KI-Modellen dar, die sowohl die Überprüfbarkeit als auch das Vertrauen in die von KI-Systemen getroffenen Entscheidungen verringern können [3]. Die Anforderung an Transparenz und Erklärbarkeit von KI-Systemen ist dabei abhängig von der Kritikalität des Anwendungsfalles sowie dem Grad der autonomen Entscheidung. Während das mit einer einzigen falschen Vorhersage verbundene Risiko bei unkritischen Systemen (z. B. Gesichtserkennungsdienste von Handykameras oder Empfehlungssysteme, die Artikel zum Kauf vorschlagen) gering ist, kann die Intransparenz von KI-Techniken bei sicherheitskritischen Anwendungen ein begrenzender bzw. sogar disqualifizierender Faktor sein [1] [7]. Dies trifft insbesondere zu, wenn eine einzelne falsche Entscheidung durch ein KI-System zu einer Gefährdung von Leben und Gesundheit der Anwender/-innen, wie z. B. im medizinischen Bereich oder beim autonomen Fahren, oder zu erheblichen Geldverlusten, wie z. B. im algorithmischen Handel, führen kann [1].

Für kritische KI-Systeme bzw. KI-Systeme, die Entscheidungen autonom treffen, ist es daher von grundlegender Bedeutung, das interne Verhalten des Systems verstehen und validieren zu können [7]. Um dieses Ziel erreichen zu können, müssen Verfahren zur Interpretation und Erklärung von kritischen KI-Modellen eingesetzt werden [8] [9].
In den letzten Jahren wurden bereits unterschiedliche solcher Erklärbarkeitstechniken für Modelle des überwachten Lernens vorgestellt und zum Teil auch der Praxis angewendet (z. B. [10] [11] [12] [13]). Prüfbare AI-spezifische Sicherheitsvorschriften und -standards wurden bis dato jedoch nicht etabliert und ebenso wurde noch kein Prüfrahmenwerk für erklärbare KI-Systeme publiziert.

Die vorliegende Arbeit soll hier ansetzen und einerseits die renommiertesten Erklärbarkeitstechniken vorstellen, die die Nachvollziehbarkeit der automatisierten KI-Entscheidungsprozesse verbessern bzw. überhaupt erst ermöglichen können. Zudem soll im Rahmen dieser Arbeit ein neuartiges Prüfrahmenwerk definiert werden, mit Hilfe dessen Auditor/-innen bewerten können, ob durch die im Unternehmen implementierten Prozesse und Kontrollen eine kontinuierliche Erklärbarkeit des KI-Systems sichergestellt werden kann [14] [15] [10].
Konkret lässt sich daraus folgende Forschungsfrage ableiten:
Welche Methoden und Tools können im Unternehmensumfeld zum Einsatz kommen, um die Entscheidungen von Künstlicher Intelligenz zu bewerten bzw. zu überprüfen?

1.2. Aufbau und Methodik

Die vorliegende Arbeit ist in folgende vier Hauptkapitel gegliedert:
Im nachfolgenden Kapitel werden unterschiedliche Ansätze vorgestellt, wie KI definiert werden kann, und es werden die wesentlichen KI-Subfelder, mit speziellem Fokus auf maschinellem Lernen, beschrieben.
Im dritten Kapitel wird die Notwendigkeit von erklärbarer KI aufgezeigt. Dabei wird ein Überblick über wenig komplexe und somit intrinsisch erklärbare bzw. transparente Modelle sowie über Black-Box-Modelle, die nur mit Hilfe spezieller Post-hoc-Techniken erklärt werden können, gegeben. Der Fokus liegt hierbei auf Modellen des überwachten Lernens, da diese in realen Anwendungen am häufigsten eingesetzt werden [16]. Anschließend werden die relevantesten Post-hoc-Techniken zur Verbesserung der Erklärbarkeit vorgestellt.
Darauf aufbauend wird im vierten Kapitel ein Prüfrahmenwerk zur Bewertung der Umsetzung von Maßnahmen für erklärbare KI in Unternehmen präsentiert und näher erläutert. Den Abschluss der Arbeit bildet die Schlussbetrachtung, in der die wesentlichen Erkenntnisse prägnant zusammengefasst und interpretiert werden.

Um den aktuellen Forschungsstand auf dem Gebiet der Erklärbaren KI zu erarbeiten und daraus Ansätze für die Beantwortung der Forschungsfrage abzuleiten, wurde eine umfassende Literaturrecherche durchgeführt. Bei dieser Recherche wurden diverse Qualitätskriterien beachtet; so wurden ausschließlich die aktuellen Forschungsergebnisse und bekannte IT-Standards für die Erarbeitung des Prüfrahmenwerks (z. B. ISO 2700x und IT-Grundschutz-Kompendium des BSI) herangezogen und die Seriosität der verwendeten Fachzeitschriften wurde mittels Journal-Ranking überprüft.

2. Grundlagen Künstlicher Intelligenz

Nachfolgend werden die für diese Arbeit relevanten Begriffsdefinitionen vorgestellt und es wird ein Überblick über die Funktionsweise der unterschiedlichen maschinellen Lernverfahren gegeben. Da die diversen Verfahren sowie deren Vor- und Nachteile nicht im Fokus dieser Arbeit stehen, wird dieser Überblick bewusst kurz gehalten.

2.1. Definition von Künstlicher Intelligenz

Die genaue Bedeutung und Definition des Begriffs „Künstliche Intelligenz“ sind Gegenstand vieler Diskussionen [9] [17]. Auch Wörterbücher enthalten zum Teil mehrere Definitionen des Begriffs, wobei z. B. das Webster Dictionary KI als „Fähigkeit einer Maschine, intelligentes menschliches Verhalten nachzuahmen“, definiert [18]. „Intelligentes menschliches Verhalten“ kann dabei unterschiedliche Ausprägungen annehmen, konkret fällt darunter jedoch u. a. die Fähigkeit, zu denken, zu lernen, zu planen, zu argumentieren, zu manipulieren bzw. Sprache zu sprechen und zu verstehen [7].
Andere Quellen verwenden nahezu die gleiche Begriffsdefinition wie das Webster Dictionary und beschreiben KI als „Theorie und Entwicklung von Computersystemen, die Aufgaben ausführen können, für die traditionell menschliche Intelligenz erforderlich ist“ [19] [20].

Unabhängig von der Begriffsdefinition können die meisten KI-Systeme in vier Kategorien eingeteilt werden:

- „Systeme, die wie Menschen denken;
- Systeme, die sich wie Menschen verhalten;
- Systeme, die rational denken und
- Systeme, die rational handeln“ [17, p. 2].

Es kann davon ausgegangen werden, dass der im Rahmen der vorliegenden Arbeit verwendete Begriff „Künstliche Intelligenz“ all jene Systeme umfasst, die mindestens einer dieser vier Kategorien zugeordnet werden können.

Arend Hintze, Assistenzprofessor an der Michigan State University, definierte in seiner Publikation im Jahr 2016 ebenfalls vier Arten von KI, die sich jedoch von den oben angeführten Kategorien unterscheiden.
Bei Typ I KI nach Hintze handelt es sich um *Reaktive Maschinen*, das bedeutet „KI in ihrer einfachsten Form“. Reaktive Maschinen reagieren in denselben Situationen immer auf die gleiche Art und Weise, wie z. B. eine Maschine, die Brettspielfiguren erkennt bzw. die Spielzüge der Gegner/-innen kennt und die nächsten Spielzüge vorhersagen kann.
Typ II KI sind Maschinen mit begrenztem Speicher, d. h. Maschinen, die weder Erinnerungen aufbauen können noch aus früheren Erfahrungen „lernen“ können. Ein Beispiel hierfür wäre ein autonomes Fahrzeug, das aufgrund eines Hindernisses jedes Mal erneut entscheidet einen Spurwechsel durchzuführen.
Bei Typ III KI handelt es sich Hintze zufolge um *Native Theorie*, was sich auf die Idee bezieht, dass Maschinen die Gedanken, Erwartungen und Gefühle ihres Gegenübers verstehen und ihr eigenes Verhalten daran anpassen können.
Eine Erweiterung dieser *Nativen Theorie* stellt die Typ IV KI, die *Selbsterkenntnis*, dar. KI-Systeme dieser Kategorie verfügen über ein Selbstbewusstsein bzw. ein Bewusstsein und können somit ihren aktuellen Zustand verstehen und aus diesen Informationen auch den Gefühlszustand ihres Gegenübers ableiten [21] [22].
In anderen Worten ausgedrückt würde ein autonomes Fahrzeug von Typ II beim Erkennen einer Fußgängerin oder eines Fußgängers die Fahrspur wechseln, während ein Fahrzeug vom Typ III auch die Wünsche und Bedürfnisse der Fußgänger/-innen in seine Entscheidungsfindung einfließen ließe. Im Gegensatz dazu würde ein Typ-IV-Fahrzeug die Fahrspur wechseln, da es sich das gleiche Verhalten von einem ihm entgegenkommenden Auto erwarten würde und diese Erwartungshaltung auf die Fußgängerin bzw. den Fußgänger übertrüge. Die meisten KI-Systeme, die heutzutage im Einsatz sind, stellen Ausführungen von Typ I oder Typ II KI dar. Die laufenden Entwicklungs- und Forschungsinitiativen werden es, aus Sicht der Expert/-innen, den Organisationen jedoch zukünftig ermöglichen, praktische Anwendungen der Typ III und Typ IV KI voranzutreiben [22].

2.2. Aktueller Kontext

KI weist eine lange Geschichte auf, die von sich abwechselnden Hoch- und Tiefphasen geprägt ist [23]. Durch das Verständnis der verteilten neuronalen Prozesse im menschlichen Gehirn angeregt, entstanden die ersten Konzepte von Künstlichen Neuronalen Netzen (KNN) bereits in den späten 1940er Jahren und fanden circa zehn Jahre später erste Implementierungen [24]. Beflügelt durch die vielversprechenden Leistungen der ersten KI-Systeme wurden in den 1950er und frühen 1960er Jahren optimistische Prognosen aufgestellt, darunter auch die Voraussage, KI könne in „absehbarer Zukunft“ die gleichen Aufgabenstellungen bewältigen wie der menschliche Verstand [23].
Dass diese Erwartungen u. a. aufgrund unzureichender Rechenressourcen und fehlender Datenmengen stark überzogen waren, stellte sich etwa zehn Jahre später heraus. So schlussfolgerte der KI-Experte Lighthill 1973, dass „die gemachten Entdeckungen in keinem Bereich die zuvor versprochenen bedeutenden Auswirkungen haben“ [25]. Auch die beiden bekannten KI-Experten Minsky und Papert zeigten Ende der 1960er Jahre in ihrer Publikation die starke Beschränktheit der Rechenleistung von größeren Neuronalen Netzen auf [26]. Diese Erkenntnisse führten Anfang der 1970er Jahre zum ersten sogenannten „KI-Winter“, der mit einer massiven Kürzung der KI-Forschungsfinanzierung sowie einer daraus resultierenden Stagnation der KI-Forschung einherging [23] [24].
In den 1980er Jahren richtete sich der Blick der KI-Expert/-innen auf symbolische Expertensysteme, deren Wissensbasis aus manuell einzugebenden logischen Regeln bestand. Die Erkenntnis, dass der konsistente Aufbau größerer Wissensbasen immer komplexer wurde und niemals alle möglichen Vorbedingungen für

jede Aktion explizit angegeben werden konnten, leitete Ende der 1980er Jahre den zweiten „KI-Winter“ ein. Erneut nahm das öffentliche und wirtschaftliche Interesse an KI-relevanten Themen und an KI-Forschungs- und Entwicklungstätigkeiten stark ab [23] [24] [27].

Erst um die Jahrtausendwende ermöglichten technologische Fortschritte, darunter u. a. die Entwicklung schnellerer und leistungsfähigerer Prozessoren, die gesunkenen Speicherkosten und die Verfügbarkeit von Hochleistungs-Parallel-Computing (z. B. bei GPUs), ein deutliches Wiederaufleben von KI [4] [7] [28]. Weiters trugen auch die weit verbreitete Verfügbarkeit von Cloud-Computing-Lösungen und der damit verbundene Zugriff auf billigere, flexible, skalierbare und leicht miteinander verbundene Computerumgebungen sowie das Aufkommen von Big Data zum Fortschritt von KI bei [7] [24]. Letzteres ermöglichte eine starke Zunahme an verfügbaren Datenmengen, einschließlich großer Datenmengen, die zum Trainieren von Algorithmen verwendet werden können [7].
Heutzutage sind KI-Lösungen in vielen praktischen Anwendungen erfolgreich (u. a. Gesichtserkennung, Spracherkennung, Empfehlungssysteme, Bildklassifizierung, autonomes Fahren, Verarbeitung natürlicher Sprache, automatisierte Diagnose) und sind somit zu einem unverzichtbaren Bestandteil des täglichen Lebens geworden [1] [29]. Mit Fortschreiten der Digitalisierung werden auch die Anwendungsmöglichkeiten von KI weiter zunehmen [26]. Einer Prognose von Tratica zufolge wird der Umsatz mit Unternehmensanwendungen im Bereich der KI in den nächsten Jahren weltweit kontinuierlich zunehmen. Während die Umsätze im Jahr 2016 noch 357 Millionen US-Dollar betrugen, sollen diese im Jahr 2025 auf 31,2 Milliarden US-Dollar ansteigen (siehe Abbildung 1) [30].

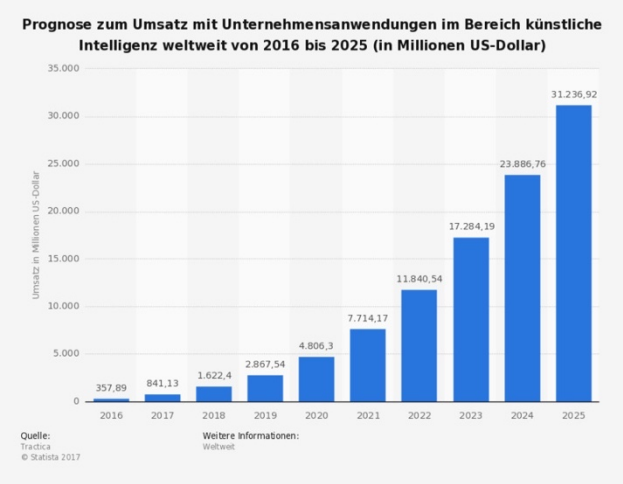

Abbildung 1: Prognose zum Umsatz mit Unternehmensanwendungen im Bereich KI **[30]**

Die Förderung der Verbreitung von KI hat sich in den letzten Jahren auch zu einem bedeutenden Handlungsfeld der Politik entwickelt. Von den Mitgliedstaaten der EU sowie von der Europäischen Kommission wurde bereits ein Aktionsplan erarbeitet, mit dem die Entwicklung und der Einsatz innovativer, ethischer und sicherer KI-Lösungen gefördert werden sollen. So ist im Aktionsplan definiert, dass Investitionen im Bereich der KI innerhalb der EU zukünftig gesteigert werden sollen und eine stärkere Koordinierung der Mitgliedstaaten anzustreben ist [31]. Angesichts der Tatsache, dass in der Zukunft neue Fähigkeiten bzw. spezifischeres Wissen für die Entwicklung und den Betrieb von KI-Lösungen erforderlich sein werden, sollen sich die Regierungen in der EU auch auf Initiativen zur Aus- und Weiterbildung in KI-Disziplinen konzentrieren [31] [7].
Die im Rahmen des Aktionsplans umzusetzenden Maßnahmen sollen der EU eine weltweite Vorreiterrolle bei der Entwicklung und dem Einsatz von modernen KI-Lösungen sichern [22]. Derzeit entfallen noch über 70 % der seit 2006 im KI-Bereich erfassten Patente auf die USA, China und Südkorea, „mit den Unternehmen Google, Amazon, Microsoft, Facebook, Samsung und Huawei an der Spitze“ [24, p. 6].

2.3. Subfelder von Künstlicher Intelligenz

Während KI vor etwa 40 Jahren noch als einheitliches Forschungsfeld betrachtet wurde, entwickelten sich in den späten 1980er Jahren die ersten Subfelder (z. B. maschinelles Lernen) mit unterschiedlichen Forschungsinhalten und Zielen [32]. Seit damals hat sich dieser Trend unvermindert fortgesetzt und es sind zahlreiche neue Subfelder von KI entstanden [1]. In der nachfolgenden Abbildung sind jene Subfelder abgebildet, die technisch bereits weit fortgeschritten und in Industrie und Wirtschaft weltweit im Einsatz sind.

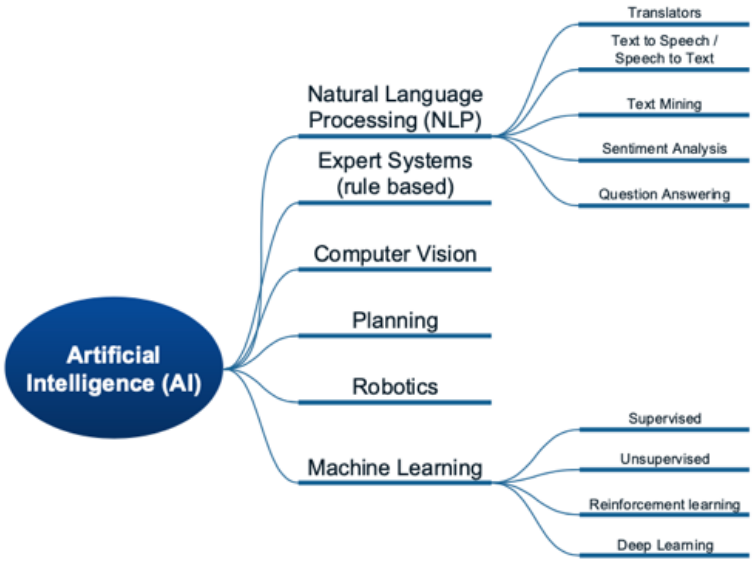

Abbildung 2: Wesentliche Subfelder von Künstlicher Intelligenz **[7, p. 10]**

Maschinelles Lernen

Die meisten im Einsatz befindlichen innovativen KI-Lösungen basieren auf maschinellem Lernen, so dass die beiden Begriffe „KI“ und „maschinelles Lernen“ häufig synonym verwendet werden [7] [24]. Laut der Norm ISO/IEC 38505-1:2017 zur IT-Betriebsführung handelt es sich bei maschinellem Lernen um einen Prozess, der „Algorithmen anstelle von prozeduraler Codierung verwendet und das Lernen aus vorhandenen Daten ermöglicht, um zukünftige Ergebnisse vorherzusagen“ [33]. Anders ausgedrückt ist maschinelles Lernen eine Reihe von Methoden, die Muster in verfügbaren Daten automatisch identifizieren und das daraus generierte Wissen auf neue Daten anwenden, um Vorhersagen zu treffen bzw. andere Arten der Entscheidungsfindung unter Unsicherheit durchzuführen. Somit kann aus Beispielen ein komplexes Modell ohne im Vorhinein festgelegte Regeln oder Berechnungsvorschriften entwickelt werden [34] [35]. Maschinelles Lernen bietet sich immer dann an, wenn die analytische Beschreibung eines Prozesses zu kompliziert ist, aber genügend Beispieldaten, wie etwa Bilder, Texte oder Sensordaten, zur Verfügung stehen [24].

Es existieren verschiedene Lernstile für maschinelles Lernen, die einen unterschiedlichen Grad an menschlicher Intervention benötigen, um die Daten angemessen zu kennzeichnen [19] [7]. Beim prädiktiven oder überwachten Lernen sind die Ergebnisse der zu lernenden Ausgabe bereits bekannt und die richtigen Antworten werden als sogenannte Labels mitgeliefert [24] [36]. Das Ziel besteht somit darin, die Ausgaben „y“ aus den Eingaben „x“ zu lernen, wenn ein gelabelter Satz von Eingabe-Ausgabe-Paaren $D = \{(x_i, y_i)\}_{i=1}^{N}$ vorliegt. D wird hier als Trainingssatz bezeichnet und N entspricht der Anzahl der Trainingsbeispiele [34].

In der einfachsten Einstellung ist jede Trainingseingabe xi ein Zahlenvektor, der beispielsweise das Gewicht bzw. die Größe einer Person darstellt. Diese Zahlenvektoren werden als Merkmale, Attribute oder Kovariaten bezeichnet.[1] Bei x_i kann es sich jedoch auch um ein komplex strukturiertes Objekt handeln, wie beispielsweise ein Bild, eine Zeitreihe, einen Satz, ein Diagramm oder eine E-Mail-Nachricht [34]. Die Ausgabevariable kann ebenso nahezu jede Form annehmen, wobei die meisten Methoden davon ausgehen, dass es sich bei y_i um eine nominelle oder kategoriale Variable aus einer endlichen Menge (z. B. weiblich oder männlich) oder einen realen Skalar (z. B. Einkommensniveau) handelt. Sofern die Zielvariable y_i einen kategorialen Wert darstellt, wird das Problem als Klassifizierung bzw. Mustererkennung bezeichnet. Dahingegen charakterisiert sich ein Regressionsproblem durch die Vorhersage einer reellen Zahl als Zielvariable [34] [37]. Mit Hilfe dieser beiden Verfahren des überwachten maschinellen Lernens, der Klassifikation und der Regression, kann bereits eine Vielzahl der für das Unternehmensumfeld relevanten Lernaufgaben bewältigt werden (siehe Tabelle 1) [38].

Tabelle 1: Anwendungsfälle für überwachte Lernverfahren **[38, p. 372]**

Input	**Output**	**Applikation**	**Lernalgorithmen**
Kreditantrag Informationen	Annahme/Ablehnung	Kreditvergabe	Entscheidungswälder k nächste Nachbarn Neuronale Netze Support Vector Machine
Störungen	Vorhersage des Fehlertyps	Maschinen-komponente	k nächste Nachbarn Support Vector Machine
Bilder von Personen	Identifikation der Personen	Gesichtserkennung	Neuronale Netze
Text	Klassifizierung	Spam-Mails/ Bewertungen	Entscheidungswälder k nächste Nachbarn
Medizinische Bilder	Krankheitsdiagnose	Medizin	Neuronale Netze
Tonaufzeichnungen	Text	Sprachassistent	Neuronale Netze

Die zweite Hauptart des maschinellen Lernens stellt das unüberwachte Lernen dar, bei dem nur die rohen Beispieldaten zur Erkennung von grundlegenden Mustern in den Datenbeständen verwendet werden [24] [34]. Dies bedeutet, dass der Algorithmus nur anhand der Eingaben, $D = \{x_i\}_{i=1}^{N}$, nach „interessanten Mustern“ in den Daten sucht. Daten, die sich durch charakteristische Muster voneinander unterscheiden, werden hierbei in unterschiedliche Kategorien bzw. Cluster zusammengefasst [7] [34]. Ein anschauliches Beispiel für die Arbeitsweise von unüberwachtem Lernen ist ein Algorithmus zur Bewertung von illiquiden Wertpapieren. Der Algorithmus würde die Wertpapiere basierend auf ähnlichen Merkmalen gruppieren (clustern) und zur Schätzung des Preises eines neuen illiquiden Wertpapiers die Preisgestaltung anderer Wertpapiere aus einem geeigneten Cluster für illiquide Wertpapiere heranziehen [7] [19].

[1] Diese Merkmale werden durch eine Abfolge von Datentransformationsschritten erstellt. Zu den Datentransformationsoperationen gehören u. a. Normalisierung, Datenmapping, Aggregationen, Neuskalierung, Diskretisierung [7].

Beim unüberwachten Lernen ist somit weder definiert, nach welcher Art von Mustern gesucht werden soll, noch können Fehler gemessen werden (dies steht im Gegensatz zum überwachten Lernen, bei dem die Ergebnisse des Lernprozesses mit den bekannten, richtigen Ergebnissen verglichen werden können) [34] [36].

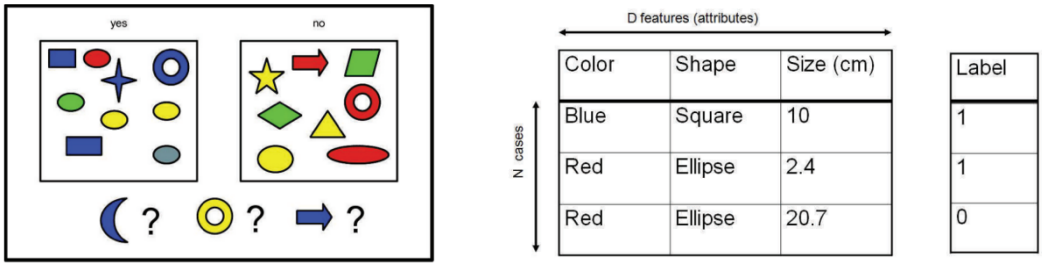

Abbildung 3: Links: Einige gelabelte sowie drei ungelabelte Trainingsbeispiele für farbige Formen. Rechts: Darstellung der Trainingsdaten als N×D-Matrix. Die erste Zeile repräsentiert hierbei den Eingabevektor xi. Die letzte Spalte stellt das Label yi ∈ {0, 1} dar. **[34, p. 3]**

Die dritte Hauptart, das bestärkende Lernen, wird derzeit im Unternehmens- sowie industriellen Umfeld seltener angewendet als überwachtes bzw. unüberwachtes Lernen [34]. Beim bestärkenden Lernen wird nicht aus einem Beispieldatensatz gelernt, sondern das Feedback der Interaktion der Maschine mit ihrer Umgebung genutzt, um die zukünftigen Aktionen zu verbessern sowie Fehler zu verringern [24] [39]. Hierbei wählt der Algorithmus ausgehend von jedem der Datenpunkte eine Aktion aus (die Datenpunkte werden großteils über Sensoren, die die Umgebung analysieren, erfasst) und erhält im Anschluss die Rückmeldung, ob die jeweilige Aktion gut oder schlecht war. Der Algorithmus wird somit durch das Empfangen positiver bzw. negativer Rückmeldungen trainiert und passt seine Strategie zur Maximierung der Belohnungen an [7].
Diese Art des maschinellen Lernens wird u. a. in der Robotik, beispielsweise zum Erlernen der bestmöglichen Greifbewegungen für Objekte, bzw. zur Optimierung von Lieferketten bei Logistikunternehmen eingesetzt [24] [40].

Deep Learning
Zusätzlich zu den drei bereits angeführten Hauptkategorien existiert noch eine weitere Form des maschinellen Lernens, das sogenannte „Deep Learning“, das aufgrund seiner großen Bedeutung häufig als vierte Hauptkategorie angeführt wird [7]. Die Deep-Learning-Algorithmen, deren Strukturen auch als Künstliche Neuronale Netze (KNN) bezeichnet werden, können sowohl für überwachtes als auch für unbeaufsichtigtes oder bestärkendes Lernen verwendet werden [39].
Beim Deep Learning werden in mehreren Schichten arbeitende Algorithmen verwendet, die von der Funktion des menschlichen Gehirns inspiriert sind. Jede der Schichten setzt sich aus miteinander verknüpften Einheiten, den sogenannten künstlichen Neuronen, zusammen. Wie die Synapsen in einem menschlichen Gehirn kann jede der Verknüpfungen ein Signal von einem Neuron zu einem anderen übertragen. Sofern ein Neuron ein Signal empfängt, kann es dieses verarbeiten und weitere mit ihm verbundene Neuronen „aktivieren“. Dies erfolgt mittels einer Aktivierungsfunktion, die die gewichtete Summe der Eingangswerte (d. h. der Verbindungen der vorherigen Schicht) heranzieht und neue Gewichte berechnet, die der nächsten Schicht einer jeden Verbindung zugeordnet werden. Der Lernprozess basiert auf dem Backpropagation-Algorithmus, der die vorhergesagten Ausgaben iterativ gegen die richtige Antwort misst und die Fehlerwerte zur Abstimmung der Gewichte in jeder Schicht in das Netzwerk zurückgibt [7] [39]. „Nach mehreren Durchläufen konvergiert die Vorhersage zur richtigen Antwort und die Fehler werden reduziert“ [7, p. 15]. Die Funktionsweise von KNN wird in Abbildung 4 durch eine schematische Darstellung veranschaulicht.

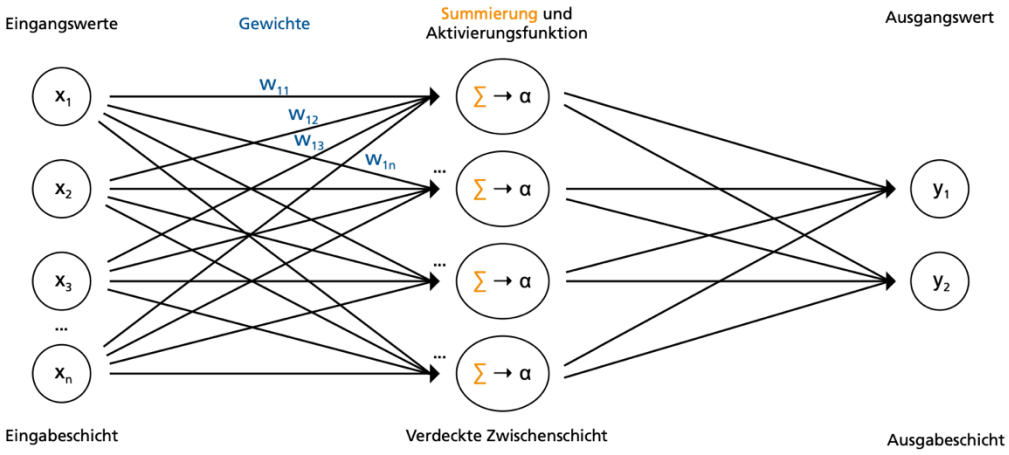

Abbildung 4: Schematische Darstellung eines Künstlichen Neuronalen Netzes (KNN) **[24, p. 12]**

In jüngster Zeit hat Deep Learning zu einer beachtlichen Weiterentwicklung in verschiedenen Bereichen, wie der Verarbeitung natürlicher Sprache (NLP – Natural Language Processing) sowie der Bilderkennung, beigetragen [19]. Dies ist auf die Fähigkeit von Deep-Learning-Algorithmen zurückzuführen, verallgemeinerbare Konzepte zu erkennen, wie beispielsweise das Konzept eines Autos in einer Reihe von Bildern. Bei NLP wird es Computern durch Deep-Learning-Algorithmen ermöglicht, die menschliche Sprache sowohl in schriftlicher als auch in gesprochener Form zu verstehen, zu analysieren und zu erzeugen (wie z. B. bei Text- und Sprachübersetzern, Text-zu-Sprache-Anwendungen oder Stimmungsanalysen) [7] [19] [41].
Besonders vielversprechend für die Erkennung von komplexen Mustern bzw. für die algorithmische Verarbeitung von natürlicher Sprache sind Künstliche Neuronale Netze, die über eine Vielzahl an Ebenen verfügen. So können KNN wie Residuale Neuronale Netze (ResNet) zum Teil in hunderten bis tausenden von Schichten organisiert sein und dabei bis zu hundert Millionen einzelner Parameter aufweisen [42].
Zur Veranschaulichung der Komplexität zeigt die nachfolgende Abbildung den Unterschied zwischen einer linearen Regressionsfunktion und einem tiefen Neuronalen Netzwerk.

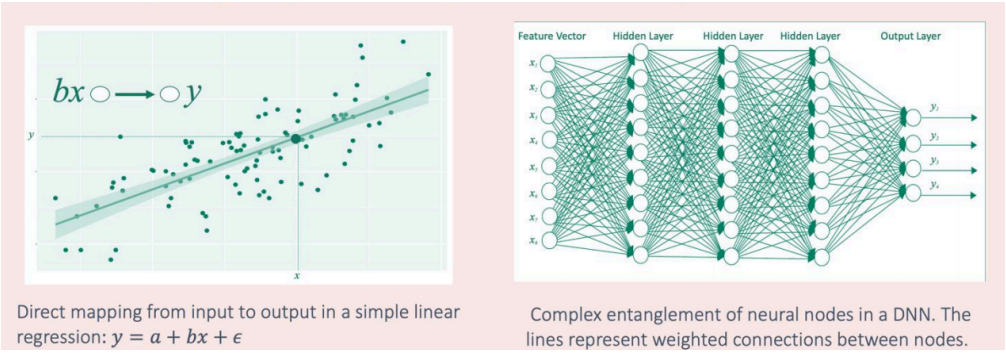

Abbildung 5: Vergleich einer einfachen linearen Regression mit einem tiefen Neuronalen Netzwerk **[15, p. 43]**

Mehrschichtige KNN liefern im Bereich der Bilderkennung und Sprachverarbeitung zwar hervorragende Ergebnisse, die Interpretierbarkeit der Parameter sowie die Erklärbarkeit des Zustandekommens der Ergebnisse sind hier jedoch nur noch eingeschränkt gegeben [42] [22]. Durch die zunehmende Entwicklung und Anwendung mehrschichtiger KNN können somit vermehrt sogenannte Black Boxes entstehen, also Systeme, deren „interner Mechanismus der Benutzerin bzw. dem Benutzer normalerweise

verborgen ist“ [18] [22]. Aufgrund dieser Entwicklung kann es zunehmend schwierig werden, mögliche Verzerrungen bzw. Probleme in den Trainingsdaten zu identifizieren und sicherzustellen, dass Algorithmen erwartungsgemäß funktionieren bzw. faire Entscheidungen treffen [42].

Um dem entgegenzuwirken, beschäftigen sich KI-Expert/-innen auf der ganzen Welt in den letzten Jahren vermehrt mit der Konzeption und Entwicklung spezieller „Explainable-AI“-Techniken. Durch diese sollen KI-Systeme ihre Entscheidungswege sowie ihre Schwächen und Stärken erläutern und ein Verständnis über ihr zukünftiges Verhalten vermitteln können [2]. Dieser Umstand wirft jedoch einige Fragen auf: Existieren bereits Tools, die es ermöglichen, die Entscheidungswege und das zukünftige Verhalten von komplexeren ML-Systemen für überwachtes Lernen, wie KNN, zu ermitteln? Ist das Verhalten von KNN tatsächlich schwieriger zu interpretieren und zu erklären als jenes von linearen Modellen? Und können auch Auditor/-innen ohne umfassende KI-Kenntnisse die aus der mangelnden Erklärbarkeit von KI-Systemen resultierenden Risiken und Problemfelder identifizieren und bewerten?
Diese Fragestellungen sollen zusätzlich zu der Hauptforschungsfrage in der nachfolgenden Arbeit mit Hilfe einer umfassenden Literaturrecherche und der Ausarbeitung des Prüfrahmenwerks beantwortet werden.

3. Erklärbare Künstliche Intelligenz

In den nachfolgenden Unterkapiteln werden sowohl die Begriffe Transparenz und Erklärbarkeit näher definiert als auch die Bedeutung und Notwendigkeit von erklärbarer Künstlicher Intelligenz im Detail erläutert. Ergänzend werden zur Beantwortung der Forschungsfrage die renommiertesten und am häufigsten in der Praxis angewendeten Erklärbarkeitsmethoden für Modelle des überwachten Lernens vorgestellt.

3.1. Transparenz und Erklärbarkeit

Das wichtigste Forschungsziel im KI-Bereich stellt aus Sicht der Fachexpert/-innen die Sicherstellung der Nachvollziehbarkeit von KI-Lösungen dar. Hierfür hat sich der Begriff „Explainable AI (XAI)“ bzw. „erklärbare KI“ etabliert, der sowohl die Transparenz als auch die Erklärbarkeit von KI-Anwendungen umfasst. Expert/-innen zufolge wird Transparenz dann ermöglicht, wenn Zugang zu den Regeln eines Systems, wie z. B. Zugriff auf dessen Quellcode, besteht [24] [43] [44]. Ein solcher Zugang zu den Regeln ermöglicht jedoch weder einen Einblick in die Datenbestände, die in unterschiedlichen Szenarien in ein KI-System eingepflegt werden, noch in die Art, wie Daten vom KI-System verarbeitet werden [44]. Da es aufgrund dieser fehlenden Informationen bei komplexeren Modellen (z. B. Deep Learning) nicht nachvollziehbar ist, warum KI-Systeme in bestimmten Situationen gewisse Entscheidungen treffen, wird der Ansatz der Transparenz oftmals als nicht ausreichend bzw. nicht wünschenswert angesehen [43] [16].

Erklärbarkeit bedeutet hingegen, dass die wesentlichen Einflussfaktoren und Gründe für konkrete Entscheidungen der KI-Systeme aufgezeigt werden können [42] [44] [45]. Erklärbarkeit wird häufig mit Interpretierbarkeit gleichgesetzt, wobei es Gilpin et al. zufolge allerdings wichtig ist, zwischen den beiden Begriffen zu differenzieren. So sind erklärbare Modelle standardmäßig auch interpretierbar, wohingegen interpretierbare Modelle nicht zwingend Auskunft über die konkrete Entscheidungsfindung liefern. Folglich reicht Interpretierbarkeit im Gegensatz zu Erklärbarkeit nicht aus, um das Vertrauen der KI-Anwender/-innen langfristig zu gewinnen [42].

Aus der zuvor angeführten Bedeutung von Erklärbarkeit kann folgende Definition für erklärbare KI (XAI) abgeleitet werden: „*unter XAI ist eine Reihe von Techniken zu verstehen, die es menschlichen Benutzer/-innen ermöglichen, die aufstrebende Generation künstlich intelligenter Systeme zu verstehen, ihnen angemessen zu vertrauen und sie effektiv zu verwalten*“ [46, pp. 2, 4]. Diese Definition vereint die beiden Konzepte Verständnis sowie Vertrauen. Darüber hinaus können jedoch auch noch andere Zwecke existieren, die den Einsatz von erklärbaren KI-Modellen begründen, wie Fairness, Informativität, Übertragbarkeit oder Kausalität [47]. Barredo Arrieta et al. schlagen daher vor, die Definition etwas weiter zu fassen und erklärbare KI als KI, „die Details oder Gründe hervorbringt, um ihre Funktionsweise klar oder leicht verständlich zu machen“, zu definieren [47, p. 6].
Es kann davon ausgegangen werden, dass alle Mittel zur Reduktion der Komplexität des KI-Modells oder zur Vereinfachung von dessen Ergebnissen als XAI-Ansatz betrachtet werden können [47].

3.2. Notwendigkeit von Erklärbarer Künstlicher Intelligenz

In der untersuchten Literatur werden unterschiedliche Beweggründe für die Entwicklung und den Einsatz von KI-Erklärungssystemen angeführt, darunter ethische, kommerzielle bzw. regulatorische Gründe [48]. Jene Gründe, aus denen die Notwendigkeit für XAI am meisten hervorgeht, werden in den nachfolgenden Unterkapiteln kurz zusammengefasst.

3.2.1. Erklärbarkeit zur Identifizierung von Verzerrungen

In den 1900er Jahren lebte „der kluge Hans" – ein Pferd, das angeblich zählen, rechnen und buchstabieren konnte und somit als wissenschaftliche Sensation galt. Später stellte sich heraus, dass Hans die Mathematik nicht wirklich beherrschte, sondern die richtige Antwort in 90 % aller Fälle aus der Reaktion des Fragestellers ableitete. Analoge Verhaltensweisen konnten von Expert/-innen auch bei modernen KI-Systemen beobachtet werden [1] [49]. Hier haben die Algorithmen gelernt, auf falsche Korrelate in den Test- und Trainingsdaten zurückzugreifen, um ähnlich wie Hans „das Richtige aus dem ‚falschen' Grund vorherzusagen" [1].
So hat beispielsweise das Gewinner-Modell eines renommierten KI-Wettbewerbs (PASCAL VOC-Wettbewerb) die Kontexte bzw. Korrelationen in den Daten verwendet, um die Objekte in Bildern zu klassifizieren. Das Modell erkannte somit Boote aufgrund des Vorhandenseins von Wasser und Züge aufgrund der Anwesenheit von Schienen. Außerdem erkannte es Pferde aufgrund von wiederholt in den Bildern auftretenden digitalen Wasserzeichen [49] [50].
Die amerikanischen KI-Experten Ribeiro et al. konnten ein weiteres relevantes Beispiel für einen solchen „Kluger-Hans-Prädiktor" nachweisen. So trainierten die Experten im Rahmen einer Studie einen logischen Regressionsklassifikator, der die Klasse „Wolf" von der Klasse „Husky" unterscheiden sollte. Die im Training verwendeten Bilder wurden dabei bewusst so ausgewählt, dass nur auf den Bildern der Wölfe im Hintergrund Schnee abgebildet war. Als Ergebnis der Studie zeigte sich, dass der Klassifikator gelernt hatte, den Schnee im Bild als relevantes Klassifizierungsmerkmal heranzuziehen, und folglich auf allen Bildern mit hellem Hintergrund einen Wolf vermutete (siehe Abbildung 6) [10].

Der Klassifikator zur Unterscheidung von Wölfen und Huskys erzielte, wie auch andere „Kluger-Hans-Prädiktoren", bei den Testsätzen gute Ergebnisse. Beim Einsatz in der realen Welt, wo sowohl Huskys als auch Wölfe in Regionen ohne Schnee leben bzw. Segelboote auch auf Bootsanhängern abgelegt werden, würden sie jedoch vollständig versagen [1] [49].
Für den Fall, dass es sich beim entsprechenden KI-System um eine Black Box handelt, sind solche Prädiktoren Samek et al. zufolge besonders schwierig zu entlarven. Erklärbarkeit könnte jedoch dabei helfen, diese Art der Verzerrungen im Modell bzw. in den Daten zu erkennen. So könnte das Fehlverhalten des Klassifikators (z. B. der Fokus auf das digitale Wasserzeichen) mit Hilfe von Erklärungen bereits anhand eines einzigen Testbilds identifiziert werden [1].

Abbildung 6: Erklärung zur Identifizierung von Verzerrungen **[10, p. 9]**

3.2.2. Erklärbarkeit zur kontinuierlichen Verbesserung

Erklärbarkeit kann somit dabei helfen, die Schwächen eines KI-Systems, wie z. B. potenzielle Verzerrungen, zu identifizieren und zu verstehen [1]. Es kann davon ausgegangen werden, dass Modelle, die ihre Schwächen umfassend darlegen sowie ihre Ergebnisse verständlich erklären können, auch einfacher zu verbessern sind [1] [48]. So liefert ein gutes Verständnis über das Systemverhalten oftmals

eine bessere Übersicht über potenzielle unbekannte Schwachstellen sowie Fehler und kann dabei helfen, Fehler mit geringer Kritikalität im Rahmen des Debugging schnell zu identifizieren und zu korrigieren [48]. Erklärungen könnten daher zukünftig in den Trainings- und Validierungsprozess von neuen KI-Modellen eingebunden werden [1].

3.2.3. Erklärungen als Voraussetzung für neue Erkenntnisse

Es ist unumstritten, dass KI-Systeme über das Potential verfügen, Muster in Daten zu erkennen, die für das menschliche Auge nicht zugänglich sind [1]. Bei diesen Mustern kann es sich z. B. um eine neue Spielstrategie, wie im Schach- oder Go-Spiel, aber auch um bis dato unbekannte Assoziationen im Bereich der Natur- oder Materialwissenschaften handeln [1] [3] [48] [51]. So wären KI-Systeme u. a. in der Lage, Assoziationen zwischen Krankheiten und Genen, kognitiven Zuständen und Gehirnaktivierungen oder chemischen Verbindungen und Materialeigenschaften zu erkennen [52] [53] [54]. Samek et al. führen in ihrer Publikation an, dass die Identifizierung solcher Muster häufig zu bedeutenderen wissenschaftlichen Erkenntnissen führen kann als die Vorhersage selbst. Damit es Forscher/-innen gelingt, neue Muster zu identifizieren, ist es jedoch essenziell zu erklären und zu interpretieren, welche Merkmale von einem KI-System zur Vorhersage herangezogen werden [1].

Die Anforderung erklärbarer und interpretierbarer Ergebnisse hat in der Vergangenheit zu einer Dominanz linearer Modelle in der Wissenschaft geführt. Bei linearen Modellen besteht die Option, diese intrinsisch zu interpretieren, wodurch eine einfache Extraktion der gelernten Muster ermöglicht wird. Durch stetige Verbesserungen bei der Entwicklung von Tools und Methoden zur Erklärbarkeit werden jedoch zukünftig auch leistungsfähigere Modelle, wie z. B. tiefe Neuronale Netze, im Bereich der wissenschaftlichen Forschung eine zentrale Rolle spielen [1].

3.2.4. Erklärungen zur Schaffung von Vertrauen

In der Vergangenheit wurden diverse Schriften über KI-Systeme, die nachweislich voreingenommene bzw. diskriminierende Entscheidungen getroffen haben, veröffentlicht (z. B. [55] [56] [57]). Publikationen wie diese haben das ohnehin bereits bestehende Misstrauen der Bevölkerung gegenüber den von KI-Systemen getroffenen Entscheidungen erheblich bestärkt. So war z. B. im Rahmen einer Umfrage der Firma Pegasystems mehr als die Hälfte der befragten Personen der Meinung, dass KI-Systeme nicht in der Lage seien, unvoreingenommene Entscheidungen zu treffen [58].[2]
Um dem entgegenzuwirken und das Vertrauen in KI-Lösungen zu stärken, ist es essenziell, algorithmische Entscheidungen durch XAI zu überprüfen [1] [29] [48] [59]. Dies kann einerseits in Situationen von Relevanz sein, in denen praktische Entscheidungen durch eine KI-Lösung getroffen werden, wie dies z. B. beim autonomen Fahren oder in der autonomen Chirurgie der Fall ist. Andererseits ist Erklärbarkeit auch in Situationen relevant, in denen KI nur eine unterstützende Rolle spielt (wie z. B. bei der medizinischen Diagnose) [1] [29]. Für den Fall, dass ein KI-System autonom über ein Krankheitsbild bzw. eine Behandlungsmethode entscheidet, sollte es diese Entscheidungen detailliert erläutern, anstatt ausschließlich die medizinische Diagnose zu stellen. Eine detaillierte Erklärung hilft zu verstehen, dass sich das KI-System nicht wie der kluge Hans verhält und das Problem auf eine sichere und robuste Weise löst, und trägt somit nachhaltig zu Vertrauensbildung bei [1].
Es sollte darüber hinaus auch der soziale Aspekt von Erklärungen nicht außer Acht gelassen werden [1] [60]. So stellen Erklärungen der eigenen Entscheidungen einen wichtigen Bestandteil der menschlichen Interaktion dar, sei es im Bereich des menschlichen Lernens oder im Feld des menschlichen Erziehens [1] [3] [61]. Weiters tragen Erklärungen dazu bei, die Akzeptanz schwieriger Entscheidungen zu fördern, vor allem dann, wenn diese nicht verständlich sind oder suboptimal erscheinen [1] [62]. Selbst wenn Patient/-innen keine zusätzlichen Informationen zur Überprüfung der vom medizinischen Personal ausgewählten

[2] Pega befragte 5 000 Personen aus den USA, Großbritannien, Frankreich, Deutschland und Japan zu ihren Ansichten über Moral, ethisches Verhalten und Empathie im Zusammenhang mit Künstlicher Intelligenz.

Therapieentscheidung zur Verfügung stehen (z. B. weil die Patient/-innen über kein umfassendes medizinisches Wissen verfügen), wird er sich durch den Erhalt von Erklärungen in den Entscheidungsprozess des medizinischen Personals integriert und somit sicherer fühlen.
Da Erklärungen essenziell sind, um Vertrauen in einer zwischenmenschlichen Beziehung aufzubauen, kann davon ausgegangen werden, dass mit Menschen interagierende KI-Systeme auch erklärbar sein sollten [1].

3.2.5. Erklärungen als Teil der Gesetzgebung

Die zunehmende Einbettung von KI in das tägliche Leben führt unvermeidlich dazu, dass vermehrt rechtliche Fragen zu deren Entscheidungsfindung und Autonomie aufkommen [1]. Mit einigen dieser Fragestellungen hat sich auch die Gesetzgebung bereits auseinandergesetzt und es wurden in den letzten Jahren diverse Gesetze und Verordnungen im Zusammenhang mit XAI verabschiedet. Ein Beispiel hierfür ist die Allgemeine Datenschutzverordnung (DSGVO) der EU, die in den Artikeln 13, 14 und 22 ein Recht auf Erklärbarkeit fordert [1] [44] [62] [63] [64]. Diese Bestimmung kommt dann zur Anwendung, wenn eine Entscheidung ausschließlich auf einer vollautomatisierten Datenverarbeitung basiert und sich dadurch rechtliche Auswirkungen bzw. ähnliche erhebliche Beeinträchtigungen für die betroffene Person ergeben [44] [64], Art. 22 DSGVO. Dies trifft beispielsweise bei einer Person zu, der vom KI-System einer Bank aufgrund von Persönlichkeitsmerkmalen (Gesundheitszustand, Einkommen etc.) ein Kredit verweigert wird [1]. Das Recht auf Erklärbarkeit würde die Bank in einem solchen Fall dazu verpflichten, der betroffenen Person aussagekräftige Informationen über die Tragweite und die Auswirkungen der Datenverarbeitung sowie über die involvierte Logik zu liefern [64], Art. 14 DSGVO.
Ergänzend dazu soll auch die EU-Verordnung über In-vitro-Diagnosegeräte, die 2022 in Kraft tritt, eine Erklärbarkeit von KI-Gesundheitssystemen fordern [44].

Die Gesetzgebung in Großbritannien plant jedoch noch einen Schritt weiter zu gehen und das „Recht auf Erklärung“ in das nationale Recht aufzunehmen. So fordert ein britischer Gesetzesentwurf, dass die betroffene Person nach jeder Entscheidung durch ein KI-System über das Ergebnis zu informieren ist. Weiters soll die betroffene Person verlangen können, dass die Entscheidung von einem Menschen überdacht oder auch alternativ von einem Menschen getroffen wird [44] [62].

3.2.6. Zusammenfassende Beurteilung der Notwendigkeit

Trotz der erheblichen Vorteile, die XAI mit sich bringt, sind sich KI-Expert/-innen einig, dass nicht immer dringender Bedarf an einer besseren Erklärbarkeit von KI-Systemen besteht [48]. So stellte der Forschungsdirektor von Google, Peter Norvig, im Rahmen eines Podiumsgesprächs fest, dass auch Menschen oftmals nicht in der Lage seien, ihre Entscheidungen umfassend zu erklären. Weiters könne die Glaubwürdigkeit von KI-Systemen laut Norvig auch durch eine längere Beobachtung ihrer Ergebnisse gemessen werden [65].
KI-Expert/-innen sind sich darüber hinaus einig, dass die Erklärbarkeit von KI-Systemen oftmals nicht im Verhältnis zu den damit verbundenen Kosten und Aufwänden steht. So stiege sowohl der technische Aufwand als auch der Ressourcenaufwand erheblich an, wenn jedes KI-System alle seine Entscheidungen unabhängig vom jeweiligen Kontext erklären müsste [48] [62].
Diese Hindernisse könnten wiederum dazu beitragen, dass Unternehmen suboptimale und weniger effiziente, aber dafür leicht zu erklärende Modelle verwenden. Genau wie auch bei den Anforderungen an menschliche Erklärungen muss somit gründlich überdacht werden, wann und warum KI-Erklärungen nützlich genug sind, um die daraus resultierenden Kosten und Aufwände aufzuwiegen.
So kann Doshi-Velez et al. zufolge der Aufwand, einen Toaster zu Erklärungen zu zwingen (z. B. warum der Toast bereits fertig ist), ein Unternehmen an der Implementierung einer intelligenten Toastfunktion hindern. Auf der anderen Seite wäre das Unternehmen hingegen möglicherweise bereit, die Kosten eines

erklärbaren – und somit nicht diskriminierenden –, aber dafür weniger genauen Kreditgenehmigungssystems zu akzeptieren [62].
Die Notwendigkeit der Erklärbarkeit ist KI-Expert/-innen zufolge von mehreren Faktoren abhängig. So ist einerseits der Grad der funktionalen Transparenz bzw. Interpretierbarkeit von Relevanz, die durch die Komplexität von KI-Algorithmen verursacht wird. Sollte dieser ohnehin hoch ausfallen, müssen keine speziellen Methoden zur Erklärbarkeit zum Einsatz kommen [47] [48]. Aber auch der Grad der Widerstandsfähigkeit der KI-Anwendung gegen potenzielle Fehler spielt eine wesentliche Rolle. So würde bei einem KI-System für personalisierte Werbung ein geringes Maß an Erklärbarkeit ausreichen, da die Folgen eines Fehlers in diesem Fall vernachlässigbar sind. Sollte ein KI-System jedoch in einem sicherheitskritischen Anwendungsbereich zum Einsatz kommen, so kann eine einzelne falsche Entscheidung durch das System bereits zu einer Gefährdung von Leben und Gesundheit der Anwender/-innen führen. Bei einem KI-basierten Diagnosesystem oder einem autonomen Fahrzeug wäre die Anforderung an Erklärbarkeit somit signifikant höher, da Fehler nicht nur den Patient/-innen bzw. Fahrer/-innen schaden, sondern auch eine Einführung solcher KI-Systeme verhindern [1] [48].
Es kann somit davon ausgegangen werden, dass es umso wichtiger ist, eine KI-Entscheidung nachvollziehbar zu erklären, je mehr diese Entscheidung das Leben einer dadurch betroffenen Person beeinflusst (sei es gesundheitlich, finanziell, sozial oder politisch) [48] [66]. Für den Fall, dass KI-Entscheidungen einen großen Einfluss haben, werden auch Regulierungsbehörden zukünftig die Befugnis haben müssen, die Verwendung erklärbarer Formen von KI zu fordern – selbst wenn dies auf Kosten der Leistung bzw. der Genauigkeit geschieht [44].

3.3. Empfänger und Informationsgehalt der Erklärungen

Je nachdem um welche Empfänger es sich handelt – sei es Benutzer/-innen, Entwickler/-innen, Regulierungsbehörden oder Auditor/-innen –, können Erklärungen mit unterschiedlichem Detaillierungsgrad und Informationsgehalt benötigt werden [1] [66] [47]. KI-Expert/-innen zufolge kann es in vielen Fällen nicht hilfreich sein, „jedem alles zu sagen“ [44, p. 38]. So können grobe und leicht interpretierbare Erklärungen für Benutzer/-innen von KI-Systemen oftmals ausreichen, wohingegen KI-Forscher/-innen und Entwickler/-innen Erklärungen bevorzugen, die tiefere Einblicke in die Funktionsweise des KI-Modells ermöglichen [1] [59] [66] [47]. Bei der Bildklassifizierung könnten durch einfache Erklärungen jene Bildbereiche grob hervorgehoben werden, die für das KI-Modell am relevantesten sind. So könnten mehrere Vorverarbeitungsschritte, wie z. B. Filtern, Glätten oder Kontrastnormalisierung, eingesetzt werden, um die Visualisierungsqualität weiter zu optimieren. Obwohl dadurch einige Informationen verworfen werden, könnte diese Art der groben Erklärungen Benutzer/-innen dabei unterstützen, Vertrauen in die KI-Technologie aufzubauen. Im Gegensatz dazu benötigen KI-Forscher/-innen und -Entwickler/-innen zur Verbesserung des Modells und um Einblicke in die (Fehl)-Funktionsweise des Modells zu erhalten alle verfügbaren Informationen, einschließlich Beweise über die Entscheidungen des KI-Systems in der höchsten Auflösung (z. B. pixelweise Erklärungen) [1].
Es ist zumeist einfach, weitere Empfängergruppen zu identifizieren, die an unterschiedlichen Arten von Erklärungen interessiert sind. Für den Fall, dass ein System im medizinischen Bereich verwendet wird, könnten diese Empfängergruppen Patient/-innen, medizinisches Personal bzw. die medizinische Einrichtung darstellen. Ein KI-System zur Analyse von Patientendaten könnte den Patient/-innen in diesem Fall einfache Erklärungen liefern und z. B. einen zu hohen Blutzuckerspiegel anzeigen, wohingegen es dem medizinischen Personal detailliertere Erklärungen, wie z. B. ungewöhnliche Zusammenhänge zwischen verschiedenen Blutparametern, liefern könnte. Darüber hinaus ist es für Einrichtungen wie Krankenhäuser möglicherweise von größerer Bedeutung, globale oder aggregierte Erklärungen zu erhalten, d. h. Erklärungen über jene Muster, die das KI-System durch die Analyse einer Vielzahl an Patientendaten gelernt hat [1].

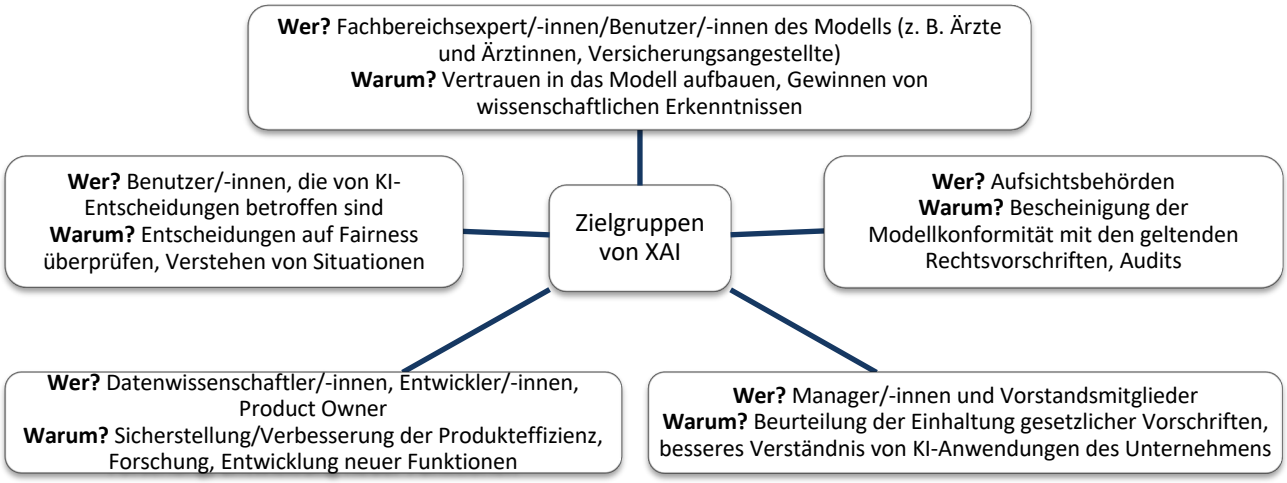

Abbildung 7: Unterschiedliche Erklärungszwecke und die zugehörigen Empfängergruppen
Abbildung inspiriert von den in Barredo Arrieta et al. **[47]** dargestellten Inhalten

3.4. Erklärungsmethoden für Künstliche Intelligenz

Es arbeiten derzeit zahlreiche Unternehmen und KI-Expert/-innen an der Entwicklung von Erklärungsmethoden, mit deren Hilfe die von KI-Systemen getroffenen Entscheidungen konsolidiert, übersetzt und in für menschliche Benutzer/-innen verständlicher Form dargestellt werden können. Darunter sind große Technologieunternehmen vertreten, wie z. B. Google mit dem Glassbox-Framework für interpretierbares maschinelles Lernen oder Microsoft mit den Best Practices für verständliche KI-Systeme [44].

Wie bereits eingangs erwähnt, liegt der Fokus der nachfolgenden Kapitel auf Erklärungsmethoden für überwachte Lernverfahren, da die meisten im Einsatz befindlichen KI-Lösungen auf überwachtem maschinellem Lernen basieren [16].
Ein Großteil der bisher publizierten Verfahren für erklärbare überwachte ML-Modelle kann durch die folgenden Erklärbarkeitsansätze kategorisiert werden:

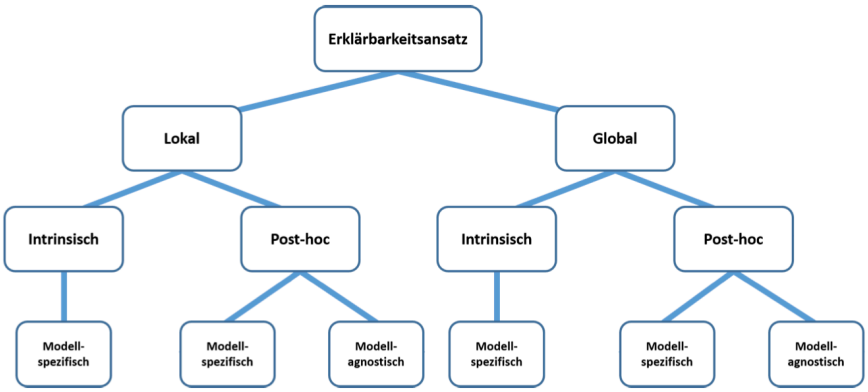

Abbildung 8: Taxonomie der Erklärbarkeitsansätze **[67, p. 24]**

So können die verschiedenen Ansätze zur Erklärbarkeit, abhängig vom jeweiligen Bereich, den das Verfahren erklärt, als globale bzw. lokale Erklärungsmethode betrachtet werden. Es handelt sich dabei um einen lokalen Ansatz, wenn das Verfahren eine einzelne Vorhersage, d. h. eine Vorhersage für einen

einzelnen Datenpunkt, erklärt [16] [47] [60] [67]. Ein globaler Ansatz betrachtet hingegen das Modell als Ganzes und erklärt dessen Verhalten somit in seiner Gesamtheit. Es ist jedoch auch möglich, Rückschlüsse auf das globale Verhalten eines Modells zu ziehen, indem ein lokaler Erklärungsansatz auf jede der Beobachtungen im Datensatz angewendet wird [60] [67].
Weiters können Modelle dann als intrinsisch erklärbar angesehen werden, wenn sie über eine einfache und wenig komplexe Struktur verfügen. Für Modelle, bei denen aufgrund ihrer komplexeren Struktur keine intrinsische Erklärbarkeit gegeben ist, müssen hingegen nach dem Training (post hoc) spezielle Methoden zur Erklärung des Modellverhaltens zum Einsatz kommen [67] [47]. In Abbildung 9 wird die aus der Modellkomplexität resultierende Erklärbarkeit diverser Modelltypen in Abhängigkeit mit ihrer Lernleistung dargestellt.

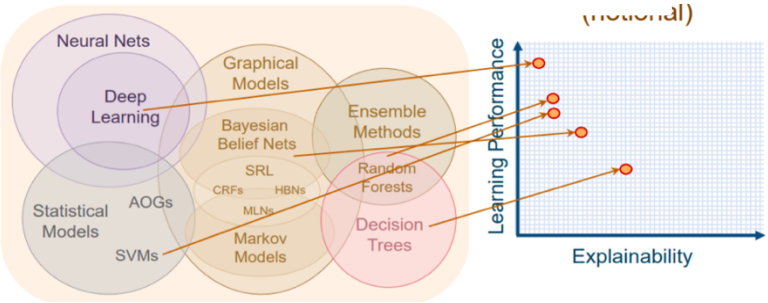

Abbildung 9: Modelltypen und ihre Lernleistung sowie Erklärbarkeit **[46, p. 5]**

Eine letzte Unterscheidung kann zwischen modellspezifischen und modellagnostischen Ansätzen getroffen werden, wobei sich spezifische Ansätze auf bestimmte Modellklassen beschränken [67] [47]. Dazu zählen z. B. jene Tools, die die Struktur eines Entscheidungsbaums zur Vorhersage von dessen Verhalten nutzen und folglich nicht für andere Modelltypen eingesetzt werden können [67]. Im Gegensatz dazu können Modellagnostische Ansätze post hoc für jedes maschinelle Lernverfahren angewendet werden [67] [47].

In den nachfolgenden Kapiteln wird ein Überblick über jene Modelle gegeben, die aufgrund ihres Designs leicht interpretiert werden können und für die keine spezifischen Post-hoc-Erklärungsmethoden eingesetzt werden müssen.

3.4.1. Intrinsisch erklärbare/transparente Modelle

Einfache Klassifikatoren, wie z. B. lineare Modelle oder flache Entscheidungsbäume, weisen auch ohne den Einsatz von Erklärungsmethoden einen gewissen Grad an Transparenz und Erklärbarkeit auf (siehe Abbildung 7) [1] [47] [16]. KI-Modelle, auf die dies zutrifft, können Barredo Arrieta et al. und Lipton zufolge als simulierbar, zerlegbar bzw. algorithmisch transparent angesehen werden [47] [16]:

- Simulierbarkeit (Ebene des gesamten Modells): Simulierbare Modelle können von Menschen vollständig simuliert bzw. durchdacht werden [47] [16]. Ein Mensch sollte somit in der Lage sein, die Eingabedaten zusammen mit den Parametern des Modells zu verstehen und in einer angemessenen Zeitspanne alle Berechnungen ohne zusätzliche Hilfsmittel durchzuführen, die zur Erstellung einer Vorhersage erforderlich sind [16]. Folglich können einzelne neuronale Perzeptron-Netzwerke als simulierbar angesehen werden, wohingegen einfache, jedoch umfangreiche (d. h. mit zahlreichen Regeln ausgestattete) regelbasierte Systeme nicht in diese Kategorie fallen [47]. Um Simulierbarkeit auch bei einem zerlegbaren Modell zu ermöglichen, muss dieses so weit in sich geschlossen sein, dass es ein Mensch im Ganzen durchdenken und verstehen kann [47] [16].
- Zerlegbarkeit (Ebene der einzelnen Komponenten): Ein KI-Modell wird als zerlegbar angesehen, wenn jeder Teil des Modells (d. h. Input, Parameter sowie Berechnung) für einen Menschen ohne

den Einsatz zusätzlicher Hilfsmittel verständlich und erklärbar ist. Ein Modell ist folglich nicht zerlegbar, wenn z. B. nicht jede Eingabe interpretiert werden kann, wodurch Modelle mit hochtechnisierten bzw. anonymen Merkmalen disqualifiziert werden [47] [16].

- Algorithmische Transparenz (Ebene des Trainingsalgorithmus): Bei algorithmisch transparenten Modellen können Benutzer/-innen verstehen, wie sich das Modell in jeder möglichen Situation unabhängig von den Eingabedaten verhalten wird. Im Falle von linearen Modellen können die Benutzer/-innen beispielsweise die Fehleroberfläche verstehen und somit Sicherheit darüber erlangen, dass die tatsächlichen Trainingsergebnisse selbst bei bisher unbekannten Datensätzen zu den erwarteten Ergebnissen konvergieren [47] [16]. Bei tiefen Architekturen (z. B. moderne Modelle des Deep Learnings) ist eine solche algorithmische Transparenz aufgrund der Undurchsichtigkeit und der notwendigen Approximation der Ergebnisse durch heuristische Optimierung (z. B. durch stochastische Gradientenverfahren) nicht gegeben [47]. Barredo Arrieta et al. führen in ihrer Publikation darüber hinaus an, dass ein Modell durch mathematische Analysen und Methoden vollständig erforschbar sein muss, um als algorithmisch transparent zu gelten [47].

Es kann davon ausgegangen werden, dass ein simulierbares Modell gleichzeitig auch die beiden anderen Eigenschaften (Zerlegbarkeit und Transparenz) erfüllt [47].
In den nachfolgenden Unterkapiteln werden alle wesentlichen Modelle angeführt, die einer der drei Kategorien für einfache Modelle zugeordnet werden können.

Lineare/logistische Regression
Mit Hilfe der linearen/logistischen Regression (LR) können beobachtete abhängige Variablen (Kategorien) durch eine oder mehrere unabhängige Variablen vorhergesagt werden [68]. Das Modell geht dabei von der Annahme einer linearen Abhängigkeit zwischen den vorhergesagten Variablen und den Prädiktoren aus, wodurch eine flexible Anpassung an die Daten verhindert wird [68] [47]. Aus diesem spezifischen Grund (Steifheit des Modells) kann die LR als interpretierbares bzw. erklärbares Modell angesehen werden.
Wie zuvor angeführt, wäre ein Modell für LR jedoch ab einer gewissen Größe nicht mehr simulierbar und könnte, wenn es sich bei den Eingabedaten um komplexe bzw. schwer verständliche Merkmale handelt, auch nicht mehr als zerlegbar eingestuft werden [16] [47].

Es kann bei LR generell davon ausgegangen werden, dass es stark von der jeweiligen Benutzerin bzw. vom jeweiligen Benutzer abhängig ist, ob ein Modell als erklärbar angesehen wird. So müssen für logistische und lineare Regressionsverfahren, auch wenn diese die Merkmale für einfache Modelle (Simulierbarkeit, Zerlegbarkeit und algorithmische Transparenz) eindeutig erfüllen, vereinzelt Post-hoc-Erklärungstechniken zum Einsatz kommen. Dies ist insbesondere dann der Fall, wenn das Modell weniger fachkundigen Benutzer/-innen erklärt werden soll. Zur besseren Erklärung der Ergebnisse der Modelle wendeten Wissenschaftler in der Vergangenheit u. a. statistische Verfahren an, wie statistische Tests einzelner Prädiktoren, Verfahren zur Validierung der vorhergesagten Wahrscheinlichkeiten oder die Goodness-of-Fit-Statistik bzw. die Wald-Chi-Quadrat-Statistik. Darüber hinaus werden auch Visualisierungstechniken als wirkungsvoll angesehen, um weniger statistikaffinen Benutzer/-innen statistische Schlussfolgerungen anschaulich zu präsentieren [47].

Entscheidungsbäume
Entscheidungsbäume werden als weiteres Beispiel für ein Modell angesehen, das alle Anforderungen an Transparenz und Interpretierbarkeit problemlos erfüllen kann [47]. Bei Entscheidungsbäumen handelt es sich um hierarchische Strukturen zur Entscheidungsfindung, die bei Regressions- und Klassifikationsproblemen zum Einsatz kommen [68] [69].
Bei der Entwicklung von Entscheidungsbäumen wurde in der Vergangenheit zumeist darauf geachtet, die menschliche Entscheidungsfindung abzubilden und Modelle wenig komplex und verständlich zu gestalten [47]. In der einfachsten Ausprägung stellen Entscheidungsbäume somit simulierbare Modelle dar, sie

können aufgrund ihrer Eigenschaften jedoch auch zerlegt bzw. algorithmisch transparent dargestellt werden. Ein Entscheidungsbaum, der über einen kleinen Umfang bzw. eine geringe Tiefe verfügt und dessen Merkmale und Merkmalsbedeutung leicht verständlich sind, kann problemlos von menschlichen Benutzer/-innen verwaltet und bewertet, d. h. simuliert werden [16] [47] [70]. Durch einen Größenzuwachs würde die vollständige Bewertung bzw. Simulation durch einen Menschen zunehmend unmöglich und das Modell könnte nur noch zerlegt (z. B. sofern jeder Knoten in einem Entscheidungsbaum einer Klartextbeschreibung entspricht) bzw. mittels mathematischer Hilfsmittel erklärt werden [16] [47]. Teilweise kann ein einzelner Entscheidungsbaum auch so komplex sein (z. B. verfügt ein Baum der Tiefe 10 bereits über Tausende Knoten), dass zur Erklärung von dessen Verhalten Post-hoc-Erklärbarkeitstechniken angewendet werden müssen (siehe Kapitel 2.4.2.1) [70].
Um die Klassifikationsgüte von Entscheidungsbäumen zu verbessern, kommen oftmals sogenannte Ensemble-Techniken, wie Entscheidungswälder, zum Einsatz. Die Idee der Entscheidungswälder beruht darauf, dass Mehrheitsentscheidungen einer Menge geeigneter Bäume bessere Klassifikationen als einzelne Bäume liefern [7] [47] [69]. Durch den Einsatz solcher Baum-Ensembles gehen jedoch die transparenten Eigenschaften des Modells verloren und es müssen auch hier Post-hoc-Erklärungstechniken angewendet werden [47].

Nächste-Nachbarn-Klassifikation (kNN)
Eine weitere Methode, die zur Gruppe der transparenten bzw. interpretierbaren Modelle gezählt wird, ist die Nächste-Nachbarn-Klassifikation [47]. Dieses Klassifikationsverfahren befasst sich mit Klassifizierungsproblemen auf eine methodisch einfache Weise: So nimmt es die Klassenzuordnung einer Stichprobe unter Berücksichtigung ihrer k nächsten Nachbarn vor. Dies bedeutet, dass bei der Klassifikationsfragestellung gezählt wird, wie viele Nachbarn welcher Klasse angehören, und anschließend eine Mehrheitsentscheidung getroffen wird [60] [71]. Bei der Regressionsfragestellung wird der Mittelwert der Zielgröße der nächsten Nachbarn gebildet und dieser Wert dem neuen Objekt zugewiesen [71].
Im Hinblick auf die Interpretierbarkeit bzw. Erklärbarkeit des Modells ist hervorzuheben, dass kNN-Modelle, wie auch Menschen, bei ihrer Entscheidungsfindung die Faktoren Entfernung und Ähnlichkeit zwischen den Beispielen heranziehen. Da die Entscheidungen von kNN-Modellen folglich leicht von Menschen nachvollziehbar sind, werden diese Klassifikatoren häufig in Situationen, in denen die Erklärbarkeit eine wesentliche Voraussetzung darstellt, eingesetzt.
Die Interpretierbarkeit bzw. Transparenz von kNN hängt eng von den Merkmalen, der Anzahl der Nachbarn sowie von der Distanzfunktion ab, die zur Messung der Ähnlichkeiten zwischen Dateninstanzen verwendet wird. Durch einen hohen k-Wert wird die vollständige Simulation des Modells durch menschliche Benutzer/-innen deutlich erschwert. Weiters würde die Verwendung komplexer Distanzfunktionen und/oder das Vorhandensein komplexer Merkmalsbeziehungen die Zerlegbarkeit des Modells behindern und somit seine Verständlichkeit ausschließlich auf die Transparenz der algorithmischen Operationen beschränken [47].

Regelbasiertes Lernen
Regelbasiertes maschinelles Lernen (RBML) umfasst jede maschinelle Lernmethode, die Regeln zur Charakterisierung der Daten, aus denen gelernt werden soll, generiert. Regelbasierte Methoden umfassen meist eine Wissensbasis bzw. einen Satz von Regeln, wie einfache Wenn-dann-Regeln oder komplexere Kombinationen aus einfachen Regeln [47].
Ebenfalls zur Familie des regelbasierten Lernens gehören Fuzzy-Regel-basierte Systeme, die für einen breiteren Aktionsradius ausgelegt sind und die Definition von sprachlich formulierten Regeln erlauben. Fuzzy-Systeme bieten gegenüber klassischen Regelsystemen zwei Vorteile: So ermöglichen sie aufgrund der sprachlich formulierten Sätze und Regeln einerseits verständlichere Modelle und schneiden andererseits in Kontexten mit bestimmter Unsicherheit besser ab [47] [72].
Ein zentrales Problem solcher Regelgenerierungssätze stellen jedoch die Abdeckung (Menge) sowie die Spezifität (Länge) der generierten Regeln dar. So ist die Leistung des Modells bei einer großen Menge an Regeln besonders hoch, wohingegen die Erklärbarkeit bei einer hohen Abdeckung bzw. Spezifität gering ausfällt.

In der Literatur wird darüber hinaus häufig darauf hingewiesen, dass regelbasierte Modelle aufgrund ihrer Fähigkeit, für menschliche Benutzer/-innen verständliche Regeln zu generieren, auch zur Erklärung von komplexeren Modellen eingesetzt werden können [47].

Bayes'sches Modell

Ein Bayes'sches Modell ist ein gerichteter azyklischer Graph, der effizient gemeinsame Wahrscheinlichkeitsverteilungen sowie Aussagen zur bedingten Unabhängigkeit der Zufallsvariablen beschreibt. Es bildet den Zusammenhang der verschiedenen Eigenschaften nicht 1:1 ab, sondern stellt eine vereinfachte Form zur effizienten Berechnung der Wahrscheinlichkeiten dar [60] [73]. Mittels Bayes'scher Modelle könnten beispielsweise die Beziehungen zwischen Krankheiten und Symptomen dargestellt werden. So kann bei gegebenen Symptomen die Wahrscheinlichkeit für das Vorhandensein diverser Krankheiten berechnet werden. Bayes'sche Modelle erfüllen ebenfalls alle Anforderungen, um in die Kategorie der simulierbaren, zerlegbaren sowie algorithmisch transparenten Modelle eingeordnet zu werden. Es muss jedoch darauf hingewiesen werden, dass ein Modell unter Umständen (z. B. bei übermäßig komplexen Variablen) die ersten beiden Eigenschaften verlieren kann [47].

Tabelle 2: Einstufung von ML-Modellen in Bezug auf ihre Erklärbarkeit **[47]**

Modell	Simulierbarkeit	Zerlegbarkeit	Algorithmische Transparenz	Post-hoc-Analyse
Lineare/logistische Regression	Die Prädiktoren können von Menschen verstanden werden und die Wechselwirkungen zwischen den Prädiktoren sind auf ein Minimum beschränkt.	Die Variablen sind zwar immer noch lesbar, jedoch ist die Zahl der Interaktionen sowie der an ihnen beteiligten Prädiktoren so weit gewachsen, dass das Modell nur durch Zerlegung interpretiert werden kann.	Variablen sowie Interaktionen sind zu komplex, um sie ohne mathematische Hilfsmittel zu analysieren.	nicht notwendig
Entscheidungs-bäume	Benutzer/-innen können die Vorhersage eines Entscheidungsbaums auch ohne mathematische Hilfsmittel selbst simulieren.	Das Modell umfasst Regeln, die Daten nicht verändern und ihre Lesbarkeit bewahren.	Das Modell weist eine große Tiefe und hohe Anzahl von Merkmalen auf, so dass das Modellverhalten nur mit Hilfe mathematischer Tools erklärt werden kann.	nicht notwendig
Nächste-Nachbarn-Klassifikation	Das Modell kann aufgrund seiner geringen Komplexität (Anzahl und Verständlichkeit der Variablen sowie verwendetes Ähnlichkeits-maß) von den Benutzer/-innen simuliert werden.	Die Anzahl der Variablen ist zu groß und/oder das Ähnlichkeitsmaß ist zu komplex, um das Modell vollständig simulieren zu können. Das Ähnlichkeitsmaß sowie die Menge der Variablen können jedoch zerlegt und getrennt analysiert werden.	Die Anzahl der Variablen ist so hoch, dass die Benutzer/-innen auf mathematische und statistische Tools zurückgreifen müssen, um das Modell zu analysieren.	nicht notwendig
Regelbasiertes Lernen	Die in den Regeln enthaltenen Variablen sind für menschliche Benutzer/-innen lesbar und die Größe des Regelsatzes ist auch ohne externe Hilfsmittel überschaubar.	Der Regelsatz ist zu groß, so dass dieser zur Analyse in kleinere Teile zerlegt werden muss.	Es müssen mathematische Tools zur Überprüfung des Modellverhaltens eingesetzt werden, da die Regeln kompliziert sind und die Größe des Regelsatzes stark angewachsen ist.	nicht notwendig
Bayes'sche Modelle	Die statistischen Zusammenhänge, die zwischen Variablen modelliert werden, sowie die Variablen selbst sind für Benutzer/-innen verständlich.	Statistische Zusammenhänge beinhalten so viele Variablen, dass sie zur Erleichterung der Analyse zerlegt werden müssen.	Statistische Zusammenhänge können auch nach der Zerlegung nicht interpretiert werden und die Prädiktoren weisen eine so hohe Komplexität auf, dass das Modell nur mit mathematischen Tools analysiert werden kann.	nicht notwendig

Baum-Ensembles	X	X	X	Normalerweise Modellvereinfachungs- oder Merkmalsrelevanz-Techniken
Support-Vektor-Maschinen	X	X	X	Normalerweise Modellvereinfachungs- oder lokale Erklärungstechniken
Tiefe Neuronale Netze	X	X	X	Normalerweise Modellvereinfachungs-, Merkmalsrelevanz- oder Visualisierungs-techniken

In den nachfolgenden Kapiteln werden komplexe Modelle sowie gängige Erklärungsmethoden vorgestellt, mit Hilfe derer die Entscheidungsfindung dieser Modelle interpretierbar bzw. erklärbar dargestellt werden kann. Hierbei werden auch modellspezifische Erklärungsmethoden für die in Tabelle 2 angeführten Baum-Ensembles, Support-Vektor Maschinen und Tiefen Neuronale Netze behandelt (siehe Kapitel 2.4.2.2).

3.4.2. Post-hoc-Erklärbarkeitstechniken

Komplexe Modelle des maschinellen Lernens, wie beispielsweise tiefe Neuronale Netze, bieten zwar eine weitaus bessere Vorhersagekraft, aber enthalten im Gegenzug mehrere Schichten nichtlinearer Transformationen (siehe Abbildung 7). Dadurch wird die Aufgabe, die zugrundeliegende Argumentation des Modells zu extrahieren, weitgehend erschwert [1] [74].

Wie eingangs beschrieben, müssen zur Verbesserung des Verständnisses, wie diese komplexen Modelle ihre Vorhersagen für einen gegebenen Input produzieren, spezielle Post-hoc-Erklärbarkeitstechniken zum Einsatz kommen [47]. Die wesentlichen dieser Techniken werden im nachfolgenden Absatz zusammengefasst:

- Texterklärungen: Hierbei wird dem Modell beigebracht, Texte zu generieren, die zur Erklärung der Ergebnisse beitragen sollen [16] [47]. Zu den Texterklärungen kann jede Methode hinzugezählt werden, bei der Symbole zur Darstellung der Funktionsweise des Modells generiert werden [75].
- Visuelle Erklärungstechniken: Diese Techniken für die Post-hoc-Erklärbarkeit zielen darauf ab, das Verhalten eines Modells zu visualisieren. Um das Verständnis der Visualisierungen zu verbessern, können diese auch problemlos mit anderen Techniken gekoppelt werden [16] [47]. Barredo Arrieta et al. führen in ihrer Publikation an, dass Visualisierungen die geeignetste Methode sind, um auch jene Benutzer/-innen, die mit der KI-Modellierung nicht vertraut sind, in komplexe Interaktionen der am Modell beteiligten Variablen einzuführen [47].
- Vereinfachungserklärungen: Darunter sind jene Techniken zu verstehen, bei denen ein neues Modell auf Basis des zu erklärenden trainierten Modells aufgebaut wird. Dieses neue Modell versucht meist seine Ähnlichkeit mit dem zu erklärenden Modell zu optimieren, seine Komplexität zu verringern und dabei eine ähnliche Leistung beizubehalten [47].
- Erklärungen anhand von Beispielen: Bei dieser Form der Erklärungstechniken werden Datenbeispiele herangezogen, um ein besseres Verständnis über das Modell selbst zu erhalten. Dies bedeutet, dass repräsentative Beispiele extrahiert werden, die die inneren Beziehungen und Korrelationen des Modells erfassen [16] [47].
- Merkmalsrelevanz-Techniken: Schließlich erklären Post-hoc-Merkmalsrelevanz-Erklärungsmethoden die innere Funktionsweise eines Modells durch die Berechnung eines Relevanz-Scores für dessen verwaltete Variablen. Durch diese Scores wird die Auswirkung (Sensitivität), die ein Merkmal auf den Output des Modells hat, quantifiziert. Ein Vergleich der von

den verschiedenen Variablen erzielten Scores hilft im Anschluss dabei, die Wichtigkeit bzw. Relevanz jeder dieser Variablen für die Erzeugung des Outputs zu erkennen [47].

3.4.2.1 Modellagnostische Techniken

Es existieren Methoden der Post-hoc-Erklärbarkeit, die auf jedes Modell, unabhängig von dessen Interna (z. B. Architektur oder Gewichte eines Klassifikators), angewendet werden können [1] [47]. Barredo Arrieta et al. führen in ihrer Publikation an, dass sich modellagnostische Techniken im Wesentlichen auf Modellvereinfachung, Merkmalsrelevanzschätzung bzw. Visualisierungstechniken stützen [47]. Da die Vielzahl der bereits publizierten Techniken den Rahmen dieser Arbeit sprengen würde, beschränkt sich die nachfolgende Aufstellung auf die bekanntesten und am häufigsten angewendeten Methoden.

- Techniken zur Modellvereinfachung: Die meisten modellagnostischen Verfahren zur Post-hoc-Erklärbarkeit fallen in die Kategorie der Modellvereinfachung, bzw. konkreter in jene der Regelextraktionstechniken [47]. Zu den bekanntesten Vertretern dieses Ansatzes gehören die Local Interpretable Model-Agnostic Explanations (LIME), die 2016 von Ribeiro et al. vorgestellt wurden [10]. Diese Methode ermöglicht es, die Einflüsse von Merkmalsänderungen auf die Vorhersagen des Modells zu untersuchen. Dafür wird das zu erklärende Klassifikationsmodell zunächst wie gewohnt trainiert, wobei jeder beliebige Algorithmus, wie Random Forests bzw. Neuronale Netze, verwendet werden kann. Da es sich bei Black-Box-Modellen zumeist um hochgradig nichtlineare sowie mehrdimensionale Funktionen handelt, ist das globale Verhalten dieser Modelle nur schwer zu erfassen. Eine einzelne Instanz enthält hingegen nur einen geringen Teil dieser Komplexität, weshalb LIME – wie auch das Akronym bereits beschreibt – versucht, Erklärungen lokal und unabhängig für jede Instanz zu finden. Folglich wird nach dem Training des Modells eine einzelne Instanz permutiert, d. h., die entsprechenden Daten werden multipliziert bzw. leicht verändert. Im Anschluss wird das trainierte komplexe Modell auf jede einzelne der Permutationen angewendet und es werden Vorhersagewerte in Form von Wahrscheinlichkeitsverteilungen für die Permutationen sowie die Distanzen und Ähnlichkeiten zu den Original-Instanzen berechnet. Abschließend wird ein interpretierbares Modell (Surrogat), wie z. B. Entscheidungsbäume oder eine lineare Regression, lokal an die Vorhersagen des zugrundeliegenden Modells (d. h. an die Wahrscheinlichkeitsverteilung) approximiert. Die Ähnlichkeiten zu den Original-Instanzen fließen hierbei als Gewichte in das Modell ein [10] [67] [60].

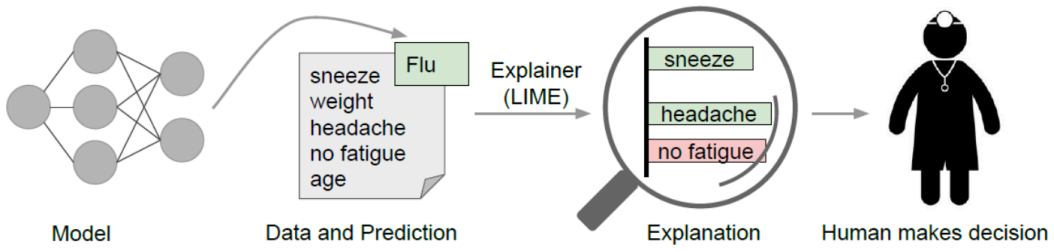

Abbildung 10: Erklärung individueller Vorhersagen durch Verwendung des LIME-Modells **[10, p. 2]**

In dem in Abbildung 10 dargestellten Beispiel wird durch das zu erklärende Modell vorausgesagt, dass der Patient oder die Patientin an einer Grippe erkrankt ist. Durch das LIME-Modell werden sowohl jene Symptome (Eingangsmerkmale), die zur Vorhersage des Modells beigetragen haben (Niesen sowie Kopfschmerzen), als auch jene Symptome, die Indizien gegen die Schlussfolgerung darstellen (keine Müdigkeit), hervorgehoben. Dabei ist zu beachten, dass die

vom LIME-Modell hervorgehobenen Symptome, wie zuvor erläutert, eine lokale und für den jeweiligen Patienten oder die jeweilige Patientin gültige Erklärung darstellen und somit möglicherweise nicht die Funktionsweise des Modells für alle Patient/-innen repräsentieren [7] [10].

Der Hauptnachteil von LIME ist die hohe Rechenkomplexität, da es bei Modellen, die dem aktuellen Stand der Technik entsprechen (wie z. B. GoogleNet), mehrere Minuten für die Berechnung einer einzelnen Vorhersage benötigt [1].

Auch das 2008 von König et al. vorgestellte Framework G-REX stellt ein bekanntes Verfahren zur Regelextraktion dar. Somit verfolgt G-REX den gleichen Ansatz wie LIME und dient dazu, ein transparentes Modell auf Grundlage eines genauen, aber undurchsichtigen Modells zu erstellen [76] [77].

- Merkmalsrelevanz-Erklärungstechniken: Auch im Bereich der modellagnostischen Merkmalsrelevanz-Erklärungstechniken wurden in den vergangenen Jahren zahlreiche unterschiedliche algorithmische Ansätze publiziert [47]. Einer der bekanntesten Ansätze ist der von Shapley 1953 geprägte Shapley-Wert, der ursprünglich aus der kooperativen Spieltheorie stammt und zur fairen Verteilung eines Gewinns zwischen Spieler/-innen eingesetzt wird [47] [60]. Mittels Shapley-Werten kann die Vorhersage eines Modells so aufgeschlüsselt werden, dass ersichtlich wird, wie viel jedes einzelne der Merkmale zu dieser Vorhersage beiträgt. Als Beispiel hierfür kann ein Modell herangezogen werden, das zur Vorhersage von Immobilienpreisen trainiert wird. Das Modell prognostiziert einen Immobilienpreis von 100.000 €, die durchschnittliche Vorhersage beträgt hingegen 85.000 €. Als Merkmalswerte werden vom Modell `size`, `area`, `floor`, `cat-banned` und `park-nearby` herangezogen. Zur Beantwortung der Fragestellungen, wie jedes der Merkmale die Vorhersage beeinflusst und wieso die tatsächliche und die durchschnittliche Vorhersage um 15.000 € voneinander abweichen, kann eine mögliche Interpretation nach Shapley lauten: Aufgrund der Nähe zum Park erhöht sich der Preis um zusätzlich 25.000 €, wohingegen das Verbot von Haustierhaltung sowie die Lage im ersten Stock den Preis um je 5.000 € reduzieren.
  Diese Berechnung wird im Anschluss für alle möglichen Konstellationen (z. B. `size` und keine weiteren Merkmalswerte, `size` + `area`, `size` + `floor`, …, `size` + `area` + `floor` etc.) durchgeführt. Der Shapley-Wert bildet dabei den durchschnittlich erwarteten marginalen Beitrag eines Merkmalswerts, nachdem alle möglichen Merkmalskombinationen berücksichtigt worden sind. Somit wird durch den Shapley-Wert eine perfekte Verteilung des marginalen Effekts eines gegebenen Merkmals über die Merkmalswerte der Instanz garantiert.

  Mit Hilfe von Shapley-Werten kann somit die Merkmalsrelevanz für eine Beobachtung berechnet und zunächst eine lokale Erklärung geliefert werden. Da die Vorhersage durch die Reihenfolge der Merkmale beeinflusst werden kann, sind alle möglichen Reihenfolgen sowie Teilmengen der Variablen zu betrachten. Es können auch globale Eindrücke des Modellverhaltens durch Anwenden der Berechnungsformel für jede der Beobachtungen gewonnen werden [15] [60] [78].

  SHAP (SHapley Additive exPlanations) ist ein von Lundberg und Lee im Jahr 2016 vorgestellter lokaler Surrogatmodell-Ansatz, der zur Ermittlung der Merkmalsbedeutung dient. Um die Merkmalsbedeutungen optimal zuzuordnen, greift SHAP auf das spieltheoretische Konzept der Shapley-Werte zurück. Eine Neuerung von SHAP ist jedoch, dass die Erklärung des Shapley-Werts als lineares Modell (wie bei LIME) dargestellt wird [15] [60] [11].

  Einen weiteren beliebten und einfachen Ansatz stellen Sensitivitätsanalysen (SA) dar, bei denen ein einzelnes Eingangsmerkmal geringfügig gestört und die Änderung des Ausgangs gemessen wird. Dies ergibt eine lokale, merkmalspezifische Annäherung an die Reaktion des KI-Modells.

Durch iterative Wiederholung des Vorgangs für zahlreiche Eingangswerte sowie für jedes der Eingangsmerkmale kann das Verhalten des Modells im Anschluss gesamthaft dargestellt werden. Dieser Ansatz wird oftmals mit Partial Dependence Plots (PDP) oder Individual Conditional Expectation (ICE) Plots erweitert, um eine grafische globale Darstellung der Bedeutung der Merkmale zu erhalten [7] [79].

Auch die sogenannte permutationsbasierte Merkmalsrelevanz wird häufig als modellagnostische Post-hoc-Erklärungstechnik eingesetzt. Dieser Ansatz basiert auf der Tatsache, dass Inputmerkmale eines Modells unterschiedliche Relevanz aufweisen und oftmals nur wenige dieser Merkmale einen wesentlichen Einfluss auf die Vorhersage haben [80]. Folglich wird bei der permutationsbasierten Merkmalsrelevanz ein Merkmal aus dem Datensatz entfernt und im Anschluss die daraus resultierende Performanceänderung (Genauigkeit, F1-Score etc.) des Modells analysiert. Ein Merkmal kann als relevant angesehen werden, wenn die Manipulierung seiner Ausprägung zu einer signifikanten Veränderung des Modellfehlers führt. In diesem Fall ist die Prognose des zu erklärenden Modells stark vom entsprechenden Merkmal abhängig. Im Gegensatz dazu kann das Merkmal als unwichtig betrachtet werden, wenn der Modellfehler nach dessen Entfernung nicht signifikant ansteigt [60] [67].

Ergänzend dazu stellten Henelius et al. 2017 in ihrer Publikation die sogenannte Automatic STRucture IDentification-Methode (ASTRID) vor. Mit Hilfe dieser Methode können jene Attribute identifiziert werden, die von einem Klassifikator zur Erstellung von Vorhersagen genutzt werden [47]. Dies erfolgt durch Identifizierung einer Teilmenge an Merkmalen, die so groß ist, dass sich die Ergebnisse eines nur mit dieser Teilmenge trainierten Klassifikators nicht von jenen Ergebnissen des mit dem ursprünglichen Merkmalssatz trainierten Klassifikators unterscheiden [47] [81].

- Visuelle Erklärungstechniken: Visuelle Erklärungen sind im Bereich der modellagnostischen Verfahren zur Post-hoc-Erklärbarkeit nur wenig verbreitet. Dies ist in erster Linie auf die Komplexität der Konzeption solcher Techniken zurückzuführen. So muss das Design der visuellen Erklärungsmethode eine nahtlose Anwendung auf jedes Modell, ohne Rücksicht auf dessen innere Struktur, ermöglichen. Aus diesem Grund werden nahezu alle modellagnostischen Visualisierungsmethoden mit Merkmalsrelevanz-Techniken kombiniert. Eine weitere Ursache für die geringe Verbreitung von modellspezifischen visuellen Erklärungstechniken besteht darin, dass visuelle Darstellungen z. B. hinsichtlich Struktur und Operationen oftmals eng mit dem zu erklärenden spezifischen Modell verknüpft sind [47].

  Dennoch wurden in der Vergangenheit vereinzelt Schriften über modellagnostische Erklärungstechniken veröffentlicht, darunter u. a. die Publikation von Friedman aus dem Jahr 2001. Hierbei wurden die sogenannten Partial Dependence Plots (PDP), oder partielle Abhängigkeitsdiagramme, vorgestellt, die für diverse Black-Box-Modelle eingesetzt werden können. PDP sind in der Lage, grafisch zu ermitteln, inwiefern sich die Vorhersage von Modellen basierend auf den Ausprägungen einer Reihe von Variablen (Prädiktoren) verändert, während die Auswirkung der restlichen Variablen herausgemittelt wird [67] [12]. PDPs stellen somit eine Methode zur Interpretation potenzieller Auswirkungen der Variablen auf Ensemble-Vorhersagen dar und können für Klassifizierungs- und Regressions-Ensembles verwendet werden [12].

  Goldstein et al. stellten im Jahr 2015 eine Erweiterung von PDPs vor, die als ICE-Plots (Individual Conditional Expectation) bezeichnet wird und zur Visualisierung jedes beliebigen Algorithmus des überwachten Lernens eingesetzt werden kann. Mit ICE-Plots kann eine Zeile pro Instanz dargestellt werden, die zeigt, wie sich die Vorhersage der Instanz bei Manipulation der Merkmalsausprägung verändert. ICE-Plots können somit als lokale Methode bezeichnet werden, während es sich bei PDPs – die sich nicht auf bestimmte Instanzen, sondern auf einen Gesamtdurchschnitt konzentrieren – um eine globale Methode handelt [60] [82].

Auch Cortez et al. beschrieben 2011 ein Portfolio an visuellen Erklärungstechniken (wie z. B. die Variable Effect Characteristic (VEC)-Kurve), die für die bessere Erklärbarkeit von überwachten ML-Modellen zum Einsatz kommen können. Hier wurden die visuellen Techniken mit einer Merkmalsrelevanz-Technik, dem Global Sensitivity Analysis (GSA)-Algorithmus, kombiniert, um eine bessere Erklärbarkeit zu erzielen [74].

Ein weiteres interessantes Beispiel für eine visuelle Erklärungstechnik ist die 2017 von Smilkov et al. vorgestellte SmoothGrad-Methode, die bei Bildklassifikatoren angewendet wird. Diese Methode baut darauf auf, dass eine Verbesserung der Erklärbarkeit bei Bildklassifikatoren durch die Identifizierung jener Pixel, die die endgültige Entscheidung besonders stark beeinflussen, erzielt werden kann. Dies kann mittels gradientenbasierter „Sensitivity Maps“ erfolgen, wobei die SmoothGrad-Methode zur visuellen Schärfung dieser Maps beitragen kann [83].

Tabelle 3: Vor- und Nachteile modellagnostischer Erklärbarkeitsmethoden

Modellagnostische Methode	Vorteile	Nachteile
LIME	• LIME kann als eine der wenigen Methoden für tabellarische Daten, Texte und Bilder eingesetzt werden. • Bei der Verwendung von kurzen Entscheidungsbäumen oder linearen Regressionen als interpretierbares LIME-Modell sind die Erklärungen kurz (d. h. selektiv) und können kontrastiv dargestellt werden. LIME kann daher auch für weniger fachkundige Benutzer/-innen verständliche Erklärungen liefern [60]. • Die durch die lokalen Surrogate erstellten Erklärungen können auch andere (interpretierbare) Merkmale verwenden als das zu erklärende Modell. Dies kann einen erheblichen Vorteil gegenüber anderen Methoden darstellen, vor allem wenn es nicht möglich ist, die ursprünglichen Merkmale zu interpretieren [60] [67].	• Ein großes Problem stellt die Instabilität des Verfahrens dar. So besteht die Möglichkeit, dass Erklärungen von zwei sehr ähnlichen Beobachtungen stark voneinander abweichen [60] [84]. • Das Sampling neuer Beobachtungen erfolgt durch Abtasten der Datenpunkte aus einer Normalverteilung, wobei Korrelationen zwischen den Merkmalen ignoriert werden. Dies kann zu unrealistischen Datenpunkten führen. • Bei einer Verwendung von LIME mit tabellarischen Daten stellt die korrekte Definition der Nachbarschaft ein großes Problem dar, da die Erklärungen stark hiervon abhängen [60].
Shapley-Wert	• Der Shapley-Wert ermöglicht sowohl die Generierung von lokalen als auch von globalen Erklärungen des Modells. • Der Shapley-Wert stellt nach dem aktuellen Stand die einzige Erklärungsmethode dar, die auf einer fundierten mathematischen Theorie basiert. Bei anderen Methoden wie z. B. LIME wird heuristisch von einem lokal linearen Verhalten eines Modells ausgegangen, ohne dies entsprechend mathematisch untermauern zu können [60]. Molnar stellt in seiner Publikation die Vermutung auf, dass Shapley-Werte	• Eine exakte Berechnung des Shapley-Wertes ist rechenintensiv, da hierfür Monte-Carlo-Simulationen eingesetzt werden [60] [67]. • Erklärungen, die mit Shapley-Werten erstellt wurden, verwenden immer alle Merkmale. Benutzer/-innen bevorzugen jedoch selektive Erklärungen, wie sie z. B. von LIME erzeugt werden. Mit SHAP können hingegen auch Erklärungen mit weniger Merkmalen geliefert werden.

	aus diesem Grund im Sinne der DSGVO als einzig konforme Methode für Erklärbarkeit angesehen werden. • Ein weiterer Vorteil ist, dass Shapley-Werte kontrastive Erklärungen ermöglichen. Anstatt die Vorhersagen einzelner Datenpunkte mit den durchschnittlichen Vorhersagen des gesamten Datensatzes zu vergleichen, können sie auch mit einer Gruppe bzw. Teilmenge an Beobachtungen bzw. mit einem anderen einzelnen Datensatz verglichen werden [60].	• Mit Shapley-Werten kann nur eine Zahl pro Merkmal und kein Vorhersagemodell, (wie z. B. bei LIME) zurückgegeben werden. Somit können keine Aussagen über Prognoseveränderungen durch Anpassung der Merkmale getroffen werden. • Wie auch bei anderen Ansätzen, die mit Permutationen von Daten arbeiten, können beim Einsatz von Shapley-Werten im Falle von korrelierten Merkmalen unrealistische Merkmals-ausprägungen entstehen [60].
SHAP	• Alle Vorteile von Shapley-Werten lassen sich auch auf SHAP übertragen [60]. • Im Gegensatz zum normalen Shapley-Wert können mit SHAP auch selektive Erklärungen mit wenigen Merkmalen generiert werden [60] [11]. • Durch die effiziente Berechnung können auch Interaktionen zwischen Merkmalen betrachtet werden. Hierdurch wird ein noch tieferer Einblick in das Modellverhalten ermöglicht. • SHAP ermöglicht aufgrund der effizienteren Algorithmen für baumbasierte Modelle eine schnellere Berechnung vieler Beobachtungen [60].	• Wie auch bei Shapley-Werten besteht ein Nachteil von SHAP darin, dass unrealistische Datenpunkte aufgrund von permutierten Daten entstehen können [60]. • Ein weiterer Nachteil, der auch bei Shapley-Werten auftritt, ist, dass der Rechenaufwand für die Untersuchung aller möglichen Merkmalskombinationen exponentiell mit der Anzahl der eingegebenen Merkmale ansteigt. Folglich ist dieser Ansatz für die überwiegende Mehrheit der Probleme nicht geeignet und Näherungen müssen ausreichen. • SHAP liefert, wie auch Shapley-Werte, kein lokal gültiges Vorhersagemodell. • Da die Schätzung durch LIME erfolgt, lassen sich auch einige Nachteile dieses Verfahrens auf SHAP übertragen [15] [60].
Sensitivitäts-analyse	• Sehr einfacher Ansatz und somit leicht zu implementieren und auch für Laien zu verstehen [1]. • Es existieren zahlreiche Erweiterungen der klassischen SA-Methoden, wie die globale SA (GSA). Folglich können diese Methoden für ein breites Anwendungsfeld eingesetzt werden [79] [1].	• Durch SA ist es aufgrund der Einfachheit der Methode nicht möglich, die diversen Korrelationen zwischen den Merkmalen zu erfassen [7]. • Mit dem Verfahren kann technisch gesehen nicht die Vorhersage des Modells selbst, sondern nur die Veränderung der Vorhersage erklärt werden [1]. • SA leiden unter fundamentalen Problemen, wie beispielsweise Gradient Shattering oder Erklärungsdiskontinuitäten, und werden daher von einigen KI-Expert/-innen als

		suboptimal für die Erklärung heutiger KI-Modelle angesehen [85].
Permutations-basierte Merkmals-relevanz	• Durch die permutationsbasierte Merkmalsrelevanz wird ein hochkomprimierter sowie globaler Einblick in das Verhalten eines Modells ermöglicht. • Die Ergebnisse können von den Benutzer/-innen leicht interpretiert werden. So ist ein Merkmal relevant, wenn die Modellfehler nach dessen Entfernung zunehmen. • Das Retraining des Modells nach Entfernen eines jeden Merkmals ist rechenintensiv, insbesondere dann, wenn viele Merkmale verwendet werden. Bei diesem Verfahren kann das Merkmal jedoch auch nur aus dem Testdatensatz entfernt und somit ein aufwendiges Retraining vermieden werden. • Bei der permutationsbasierten Merkmalsrelevanz wird automatisch jede Interaktion mit anderen Merkmalen berücksichtigt. Dies bedeutet, dass durch Permutieren eines Merkmals auch die Interaktionseffekte mit anderen Merkmalen zerstört werden. Dies ist Vor- wie auch Nachteil zugleich (siehe Nachteile) [60].	• Es ist unklar, ob zur Berechnung der Relevanz eines Merkmals Trainings- oder Testdaten verwendet werden sollen. • Falls Merkmale miteinander korrelieren, können durch die zufällige Permutation unrealistische Beobachtungen entstehen (wie dies auch bei PDPs der Fall ist). • Die Ergebnisse können willkürlich ausfallen, was auf deren Abhängigkeit von der zufälligen Permutation der Merkmale zurückzuführen ist. • Wie bereits bei den Vorteilen angeführt, berücksichtigt die permutationsbasierte Merkmalsrelevanz automatisch alle Interaktionen mit anderen Merkmalen. Somit wird die Bedeutung der Wechselwirkung zwischen zwei Merkmalen in die Wichtigkeitsmessung der beiden Merkmale einbezogen. Dies bedeutet, dass sich die Merkmalsrelevanz nicht zum Gesamtperformanceverlust addiert, sondern die Summe größer ist. Die Wichtigkeiten summieren sich nur dann entsprechend, wenn keine Interaktionen zwischen den Merkmalen bestehen (wie dies bei linearen Modellen der Fall ist) [60].
PDP	• Die Idee hinter PDP ist intuitiv, wodurch diese einfach zu implementieren und auch für Laien verständlich sind [60]. • Durch die Manipulation eines Merkmals und die anschließende Messung der Vorhersageänderung können Kausalzusammenhänge analysiert werden [86]. Es handelt sich dabei jedoch lediglich um Kausalität aus der Sicht des Modells, wodurch diese nicht den Abläufen in der realen Welt entsprechen muss. • Der mittlere Einfluss eines Merkmals auf die Vorhersage kann exakt dargestellt werden. Dies ist aber nur dann möglich, wenn das entsprechende Merkmal nicht mit anderen Merkmalen	• Es können nur ein- oder zweidimensionale Beziehungen zwischen den Modellen dargestellt werden. Mehrdimensionale Modelle können folglich nicht mehr verständlich visualisiert werden [60]. • Ohne Berücksichtigung der Merkmalsverteilung kann es bei diesem Verfahren zu falschen Interpretationen kommen. So könnten Bereiche des Plots zu große Aufmerksamkeit erhalten, obwohl sie wenige Beobachtungen enthalten und somit für die Erklärung des Modellverhaltens weniger relevant sind [60] [67]. • Heterogene Effekte können unter Umständen nicht erfasst werden, da ausschließlich die durchschnittlichen

	korreliert [60].	Randverteilungen betrachtet werden. Dies kann im schlimmsten Fall zu einer falschen Darstellung des Merkmalseinflusses führen, vor allem wenn der Einfluss der Merkmale stark streut bzw. die Merkmale stark miteinander korrelieren [60].
ICE	• ICE-Kurven sind zumeist noch intuitiver zu verstehen als PDPs. • Im Gegensatz zu PDP können durch ICE-Kurven auch heterogene Beziehungen durch Interaktionen aufgedeckt werden [60].	• Mit ICE-Kurven kann nicht mehr als ein Merkmal dargestellt werden. Für die Darstellung mehrerer Merkmale wäre das Zeichnen von überlagerten Flächen erforderlich. • Wie auch bei PDPs besteht bei ICE-Kurven das Problem, dass ungültige Beobachtungen entstehen können, falls das betrachtete Merkmal mit etwaigen anderen Merkmalen korreliert. • Aufgrund zu vieler ICE-Kurven kann die Darstellung unübersichtlich werden und relevante Informationen können überdeckt werden [60].

3.4.2.2 Modellspezifische Techniken

Der nachfolgende Abschnitt konzentriert sich auf zwei flache ML-Modelle (Baum-Ensembles und Support-Vektor-Maschinen) sowie auf Deep-Learning-Modelle und stellt populäre Post-hoc-Erklärungstechniken vor, die spezifisch für diese Modelltypen zum Einsatz kommen können.

Baum-Ensembles
Ähnlich wie bei den modellagnostischen Verfahren handelt es sich bei den bisher publizierten Erklärungstechniken für Baum-Ensembles zumeist um Merkmalsrelevanz-Techniken. Es wurden in der Vergangenheit auch einige modellspezifische Techniken zur Modellvereinfachung für Baum-Ensembles veröffentlicht, die sich jedoch hinsichtlich ihrer Funktionalität kaum gegen modellagnostische Verfahren, wie z. B. LIME, durchsetzen können [47].
Baum-Ensembles zählen zu den genauesten Modellen, die heutzutage im Einsatz sind, und können, wie bereits angeführt, dabei helfen, die Klassifikationsgüte bzw. Generalisierungsfähigkeit von Entscheidungsbäumen zu verbessern und somit Overfitting[3] zu vermeiden [7] [47] [69]. Ein großer Vorteil der baumbasierten Ensemble-Methoden liegt darin, dass diese oft eine implizite Merkmalsselektion durchführen und somit nur eine kleine Teilmenge an besonders trennscharfen Merkmalen für eine Vorhersage herangezogen wird [87]. Das Ergebnis dieser impliziten Merkmalsselektion kann durch die Gini-Wichtigkeit bzw. den Gini-Koeffizienten, einen 1984 von Breiman et al. vorgestellten Indikator für die Relevanz eines Merkmals, dargestellt werden. Mittels Gini-Koeffizienten wird an den inneren Knoten eines binären Entscheidungsbaumes der optimale Split bestimmt. Hierbei wird immer jener Split ausgewählt, der die größte Reduzierung der Unreinheiten der Daten in den jeweiligen Knoten ermöglicht [67] [80].

[3] Rauschen bzw. zufällige Schwankungen in den Trainingsdaten werden vom Modell als Konzepte erfasst und gelernt, was sich in weiterer Folge negativ auf die Leistung des Modells auswirkt.

Falls der Gini-Koeffizient erfolgreich für einen einzelnen Entscheidungsbaum berechnet werden kann, lässt sich das Vorgehen durch Aggregation der Werte aller Bäume des Waldes auch auf ein ganzes Ensemble erweitern [67] [80].
Ein weiteres bekanntes modellagnostisches Verfahren für Entscheidungsbäume bzw. Baum-Ensembles stellen Bauminterpreter dar. Solche Interpreter können selbst in Fällen, in denen Entscheidungsbäume tief sind, d. h. viele Ebenen enthalten, eingesetzt werden. Der Zweck von Interpretern ist es, die Entscheidungsbäume zu analysieren und jene Entscheidungsschritte innerhalb der Bäume zu zeichnen, die vorwiegend zur endgültigen Entscheidung bzw. Vorhersage beitragen. Das Ergebnis kann im Anschluss auch aggregiert werden, um jenen Entscheidungspfad zu zeichnen, der zur Vorhersage von Baum-Ensembles führt. Die Erklärungen von Bauminterpretern gelten sowohl lokal als auch global [7] [70].

Tabelle 4: Vor- und Nachteile einer modellspezifischen Erklärbarkeitsmethode

Modellspezifische Methode	Vorteile	Nachteile
Gini-Wichtigkeit	• Einen großen Vorteil dieser Merkmalsrelevanz-Technik stellt ihre einfache Berechnung dar. So lässt sich die Wichtigkeit eines Merkmals bereits im Training berechnen. • Es wird durch dieses Verfahren ein hochkomprimierter und globaler Einblick in das Verhalten eines Modells ermöglicht [60]. • Interaktionen zwischen Merkmalen werden automatisch durch das Verfahren berücksichtigt, da diese bereits beim Training der Bäume Eingang finden [80].	• Dieses Verfahren wird als grob und statisch angesehen, da es wenig Aufschluss über die tatsächlichen Daten und die daraus resultierenden individuellen Entscheidungen liefert [70]. • Ein weiterer Nachteil des Verfahrens ist, dass es stetige bzw. kategorielle Merkmale, die über viele verschiedene Ausprägungen verfügen, zumeist bevorzugt [88].

Support Vector Machine (SVM)
SVM-Modelle verfügen über hervorragende Vorhersage- und Generalisierungsfähigkeiten und zählen somit zu den am häufigsten verwendeten Modellen für maschinelles Lernen [47] [79]. SVM sind jedoch zumeist komplexer als Baum-Ensembles und weisen eine undurchsichtigere Struktur auf. Die Ausgangsbasis für den Bau einer SVM bildet eine Menge an Trainingsobjekten, deren Klassenzugehörigkeit bekannt ist. Die Trainingsobjekte werden dabei durch Vektoren in einem Vektorraum repräsentiert. Technisch gesehen ist es die Aufgabe der SVM, eine Trennfläche in diesen Vektorraum einzupassen, die als Hyperebene fungiert und die Trainingsobjekte in zwei Klassen teilt. Somit wird eine Hyperebene oder eine Menge solcher Hyperebenen in einem hoch- bzw. unendlich dimensionalen Raum konstruiert, die für Klassifikationen, Regressionen oder auch andere Aufgaben wie beispielsweise Ausreißererkennung eingesetzt werden kann [15] [34] [47] [89].

In der Literatur finden sich zahlreiche unterschiedliche Ansätze, um das Verhalten von SVM für Benutzer/-innen verständlich darstellen zu können, darunter Erklärungen durch Vereinfachung, lokale Erklärungen, Visualisierungen sowie Erklärungen durch Beispiele [47].
Bei der Erklärung durch Modellvereinfachung kann zwischen verschiedenen Arten unterschieden werden, die unterschiedlich tief in die innere Struktur des Algorithmus eindringen. Bei der ersten Art handelt es sich um Techniken, die nur die Support-Vektoren eines trainierten Modells heranziehen, um erklärbare regelbasierte Modelle zu erstellen [47]. So beschreiben z. B. Barakat und Bradley in ihrer Publikation aus dem Jahr 2007 eine Methode, bei der Regeln durch modifizierte sequenzielle Abdeckungsalgorithmen

direkt aus den Support-Vektoren eines trainierten SVM-Modells extrahiert werden können [90]. In [91] schlägt Barakat eine weitere Erklärungsmethode durch Regelextraktion vor, wobei auch hier nur die Support-Vektoren des trainierten Modells berücksichtigt werden. Ein weiterer interessanter Ansatz zur ersten Art der Modellvereinfachung wird in [92] von Da Costa Chaves et al. vorgestellt. Die KI-Expert/-innen generieren hierfür Fuzzy-Regeln anstatt klassischer Regeln, um ein linguistisch verständlicheres Ergebnis zu erhalten.
Die zweite Art der Modellvereinfachungen kann u. a. durch die Publikation von Fu et al. veranschaulicht werden, in der vorgeschlagen wird, die Hyperebene der SVM zusätzlich zu den Support-Vektoren zur Erstellung der Regeln heranzuziehen [93]. Im dritten Ansatz zur Modellvereinfachung werden auch die eigentlichen Trainingsdaten als Komponente zur Erstellung der Regeln hinzugefügt [47].
Aufbauend darauf werden auch diverse Clustering-Methoden vorgestellt, die zur Bestimmung von Prototyp-Vektoren für jede Klasse eingesetzt werden können. Indem die Autoren diese Punkte mit Hilfe geometrischer Methoden mit den Support-Vektoren kombinieren, können Ellipsoide bzw. Hyper-Rechtecke im Eingaberaum definiert werden, die anschließend in Wenn-dann-Regeln überführt werden (z. B. [94] [95] [96]).
Abgesehen von Methoden zur Regelextraktion wurden in den vergangenen Jahren auch einige Visualisierungstechniken zum besseren Verständnis des Verhaltens von SVM publiziert. So wird beispielsweise in [97] ein innovativer Ansatz für Support-Vektor-Regressionsmodelle vorgestellt, mit dessen Hilfe die tatsächlichen Zusammenhänge zwischen den Eingangsvariablen und den zugehörigen Ausgangsdaten visualisiert werden können. Ein weiteres Beispiel stellt die von Rosenbaum et al. publizierte Visualisierungsmethode dar, bei der die Ausgabe linearer SVM durch eine Heatmap-Molekül-färbungstechnik erklärt wird. Basierend auf den Gewichten des linearen Modells werden bei dieser Technik jedes Atom und jede der Bindungen einer Verbindung entsprechend ihrer Relevanz eingefärbt [98].
Weiters argumentieren die Autoren von [99], dass viele der bisher durchgeführten Studien zur besseren Erklärbarkeit von SVM nur die Gewichtsvektoren heranziehen, wodurch das sogenannte Margin[4] außer Acht gelassen wird. Im Rahmen ihrer Studie zeigen die Autoren die Relevanz des SVM-Margin auf und erstellen eine Statistik, in der dieses explizit berücksichtigt wird. Die Statistik ist dabei spezifisch genug, um die multivariaten Muster im Bereich des Neuroimaging zu erklären [99].

Nach näherer Betrachtung der bereits publizierten Post-hoc-Erklärungstechniken kann ein Unterschied zwischen Techniken für SVM und Techniken für zuvor vorgestellte KI-Modelle ausgemacht werden. So zählen bei letzteren die Modellvereinfachung, aber auch die Merkmalsrelevanz zu den mit Abstand am häufigsten angewendeten Methoden zur Post-hoc-Erklärung. Im Gegensatz dazu sind für SVM auch Ansätze zur lokalen Erklärung von hoher Relevanz.
Als abschließende Bemerkung kann festgehalten werden, dass keine der untersuchten Erklärbarkeitstechniken für SVM nach 2017 publiziert wurde. Dies kann unter anderem auf die gute Funktionalität bereits publizierter Ansätze zurückzuführen sein, durch die das Verhalten von SVM bereits umfassend erklärt werden kann [47].

Künstliche Neuronale Netze (KNN)
KNN wurden seit ihren Anfängen von KI-Expert/-innen geschätzt, vor allem aufgrund ihrer guten Fähigkeit, komplexe Beziehungen zwischen den Variablen abzuleiten. Wie bereits angeführt, werden Neuronale Netze jedoch als Black-Box-Modelle betrachtet, deren internes Verhalten für Benutzer/-innen nicht erklärbar bzw. nicht transparent ist [42] [22] [47].
Aufgrund der Tatsache, dass Erklärbarkeit oftmals zwingend notwendig ist, um KI-Modelle erfolgreich in der Praxis einzusetzen, wurden in den vergangenen Jahren unterschiedliche Erklärbarkeitstechniken für KNN entwickelt, einschließlich Modellvereinfachungs- und Merkmalsrelevanz-Techniken, Texterklärungen, lokaler Erklärungen sowie Modellvisualisierungen [47].

[4] Kleinster Abstand der Trainingspunkte zur Hyperebene.

Im Bereich der Modellvereinfachung wurden zwar diverse Ansätze für die Erklärbarkeit von KNN vorgeschlagen, jedoch sind diese primär nur für Neuronale Netze mit einer einzelnen versteckten Schicht anwendbar. Einer der wenigen Ansätze für mehrschichtige KNN ist der sogenannte DeepRED-Algorithmus [13], der auf dem von Sato und Tsukimoto [100] vorgestellten Ansatz zur Regelextraktion aufbaut. Der ursprüngliche Algorithmus [100] dient dazu, ein einschichtiges Neuronales Netz mit Hilfe von Entscheidungsbäumen zu zerlegen und die aus jedem Baum extrahierten Regeln im Anschluss zusammenzuführen. Bei DeepRED wird dieser Ansatz für einschichtige Neuronale Netze durch Hinzufügen weiterer Entscheidungsbäume und Regeln für mehrschichtige Netze erweitert [13].

Aufgrund der Tatsache, dass die Vereinfachung von KNN mit steigender Anzahl an Schichten zunehmend komplex wird, ist die Erklärung dieser Modelle durch Merkmalsrelevanz-Techniken immer beliebter geworden. Eine der repräsentativsten Arbeiten in diesem Bereich ist [101], in der eine Methode zur effizienten Bewertung der Wichtigkeit einzelner Pixel in Bildklassifizierungsanwendungen vorgestellt wird, die auf einer sogenannten tiefen Taylor-Zerlegung basiert. Die Funktionsweise der tiefen Taylor-Zerlegung wird grafisch in Abbildung 11 dargestellt. Konkret wird durch Vorwärtspropagation der Pixelwerte {xp} ins Neuronale Netz der Funktionswert f(x) erhalten, wobei dem Ausgangsneuron die Relevanz Rf = xf zugewiesen wird. Die Relevanzen werden daraufhin von der obersten Schicht bis hin zum Input zurückgespeist (backpropagiert), wodurch {Rp} als Relevanzwerte aller Pixel angesehen werden kann. Neuronen der untersten versteckten Schicht, die von den höheren Schichten als relevant angesehen werden, verteilen im Anschluss die ihnen zugewiesenen Relevanzen auf rote Pixel um. Dadurch können Heatmaps dargestellt werden, die es ermöglichen, die Relevanz der Eingabepixel bei der Klassifizierung eines ungesehenen Datenpunktes klar und intuitiv zu verstehen (siehe Abbildung 11) [101].

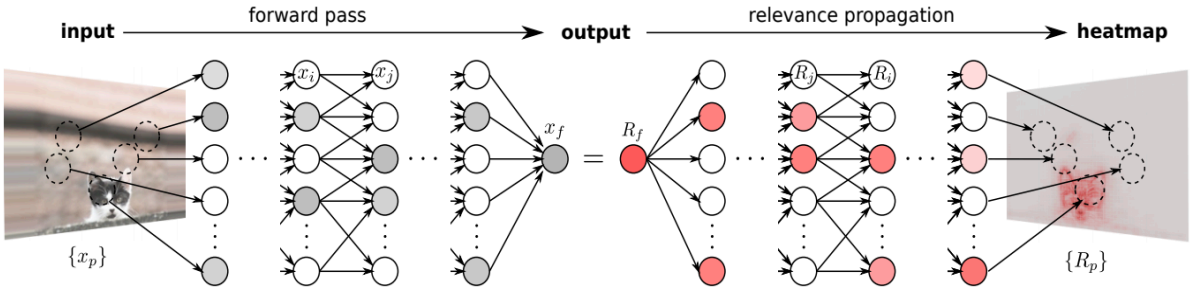

Abbildung 11: „Grafische Darstellung des Rechenflusses der tiefen Taylor-Zerlegung“ **[101, p. 5]**

Den gleichen Ansatz wählen die Autor/-innen in [102] mit DeepLIFT, einem Verfahren zur Berechnung von Wichtigkeitsscores in mehrschichtigen Neuronalen Netzen. Hierbei wird die Aktivierung von Neuronen mit der Referenzaktivierung verglichen und ein Score entsprechend der Differenz zugeordnet.
Ein weitere repräsentative Merkmalsrelevanz-Technik zur Erklärbarkeit von KNN wurde im Jahr 2018 von Montavon et al. vorgestellt [85]. Diese Technik kann als Gegenteil der Sensitivitätsanalyse angesehen werden, in dem Sinne, dass ausgehend von der Ausgabe des Modells auf jeder der Schichten die Relevanz der Eingabe der vorherigen Schicht analysiert wird, bis die Eingabeebene erreicht ist.
Auch hier stellt das Ergebnis eine Heatmap dar, in der jene Eingabemerkmale visuell hervorgehoben werden, die hauptsächlich zur Ausgabe beigetragen haben. In dem in der nachfolgenden Abbildung dargestellten Beispiel wird ein Bild x vom KNN als „Boot“ klassifiziert, wobei die Pixel mit hoher Relevanz rot eingefärbt sind [7] [85].

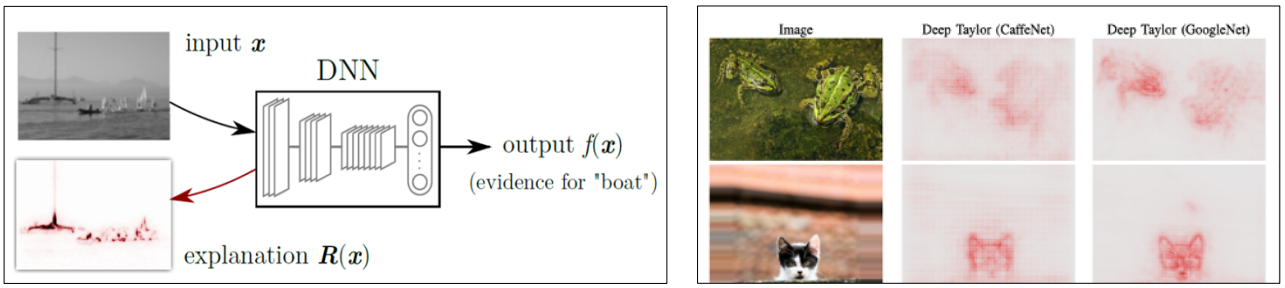

Abbildung 12: Erklärung der Klassifizierungsentscheidung von KNN, links durch DeepLIFT **[85, p. 4]** und rechts durch die tiefe Taylor-Zerlegung **[101, p. 12]**

3.4.3. Evaluierung von XAI-Techniken

Damit sich das Feld der XAI in Zukunft weiterentwickeln kann, wird es zwingend erforderlich sein, eine gemeinsame Basis zu schaffen, auf der neue Techniken und Methoden durch die Gemeinschaft entwickelt werden können. So sollten eine standardisierte Terminologie im Bereich der XAI sowie eine gemeinsame Struktur für jedes XAI-System entwickelt werden.
Ein weiteres Schlüsselmerkmal, das zukünftig entwickelt werden sollte, ist eine quantifizierbare Metrik bzw. eine Gruppe solcher Metriken. Diese sollten es ermöglichen zu messen, inwiefern KI-Modelle die Definition der Erklärbarkeit erfüllen [47].
In den letzten Jahren wurden bereits diverse Versuche zur Messung von XAI unternommen, wie unter anderem in den Publikationen [103] und [104] beschrieben. Diese XAI-Messungen dienen im Allgemeinen dazu, die Güte von Erklärungen, die Auswirkungen der Erklärungen auf die Modellleistung und das Vertrauen der Benutzer/-innen sowie die Nützlichkeit der Erklärung und die daraus resultierende Zufriedenheit der Benutzer/-innen zu bewerten [103] [104] [47]. Die im Rahmen von [103] und [104] vorgestellten Messmethoden (z. B. Erklärungszufriedenheitsskala, rechnerische Maße für Erklärtreue, Modellzuverlässigkeit und Erklärungsvertrauenswürdigkeit) liefern gute erste Ansätze zur Evaluierung von XAI-Systemen. Für eine umfassende Evaluierung ist jedoch eine Unterstützung dieser bereits bestehenden Ansätze durch allgemeine und quantifizierbare XAI-Metriken notwendig [47].

3.4.4. Zusammenfassende Betrachtung Erklärbarer KI-Modelle

Es kann abschließend festgehalten werden, dass lineare Modelle trotz ihrer Einfachheit nicht immer besser interpretiert bzw. erklärt werden können als tiefe Neuronale Netze. So arbeiten KNN häufig mit leicht verarbeiteten oder rohen Merkmalen, wohingegen lineare KI-Modelle zur Erzielung vergleichbarer Leistungen oftmals mit stark handgefertigten Merkmalen arbeiten müssen. Die Behauptung, dass lineare Modelle einfacher interpretiert bzw. erklärt werden könnten, mag somit in Bezug auf die algorithmische Transparenz unumstritten sein. Bei stark konstruierten oder hochdimensionalen Merkmalen können lineare Modelle jedoch ihre Simulierbarkeit bzw. Zerlegbarkeit verlieren [16].
So zeigten auch Lipton et al. in ihrer Publikation aus dem Jahr 2016 auf, dass sich lineare Modelle nur dann der Leistung rückgekoppelter Neuronaler Netze annähern können, wenn sie ihre Eigenschaft der Zerlegbarkeit aufgeben [105].
Wenn sowohl für lineare Modelle als auch für KNN Post-hoc-Erklärbarkeitstechniken eingesetzt werden müssen, bringen KNN einen klaren Vorteil mit sich. So sind KNN in der Lage, Repräsentationen zu lernen, die im Anschluss verständlich verbalisiert, visualisiert oder für Clustering angewendet werden können [16]. Folglich kann der oftmals aufgestellten Behauptung, dass lineare Modelle aufgrund ihrer einfacheren Interpretierbarkeit und Erklärbarkeit KNN immer vorzuziehen seien [106], nur bedingt zugestimmt werden. So ist die Erklärbarkeit nicht nur von der Komplexität des Modells, sondern auch von der Aufbereitung der

Eingabemerkmale, den verfügbaren Datenressourcen, dem Anwendungsszenario und der angewendeten Post-hoc-Erklärbarkeitstechnik abhängig.
Lineare Modelle sollten daher immer dann ausgewählt werden, wenn Daten in der Praxis sorgfältig vorverarbeitet werden können. Hierdurch kann die Genauigkeit der interpretierbaren Modelle in vielen Fällen so weit verfeinert werden, dass die aus der Kombination von Transparenz und Leistung resultierenden Vorteile jene Vorteile von semantisch intransparenteren Modellen überwiegen [15] [107].

4. Prüfrahmenwerk für Erklärbare Künstliche Intelligenz

Der rasante Aufstieg von KI ging mit einer Verzögerung der Festlegung und Einführung von KI-spezifischen Regulierungs- und Compliance-Rahmenbedingungen einher. So liegen derzeit weder ein ausgereifter Prüfungsrahmen, in dem KI-Prozesse detailliert beschrieben werden, noch AI-spezifische Vorschriften oder Standards vor. Auditor/-innen stehen somit aktuell vor der Frage, wie Audits erfolgreich durchgeführt werden sollen, wenn keine weit verbreiteten Präzedenzfälle für die Auditierung von KI-Systemen und KI-Anwendungsfällen vorhanden sind [108].
Diese Problemstellung wird im nächsten Abschnitt der vorliegenden Arbeit durch die Vorstellung eines Prüfrahmenwerk für XAI aufgegriffen. Dieses Prüfrahmenwerk soll es den Auditor/-innen ermöglichen, zu bewerten, ob transparente KI-Design- und Implementierungs- und Betriebsprozesse eingerichtet sind, die eine kontinuierliche Erklärbarkeit des Verhaltens der im Einsatz befindlichen KI-Systeme gewährleisten. Dadurch sind die Auditor/-innen in der Lage, potenzielle mit mangelhafter Erklärbarkeit im Zusammenhang stehende Risiken zu bewerten, zu verstehen und an relevante Stakeholder zu kommunizieren [14] [15] [10].

KI-Governance
Ist ein unternehmensweites Governance-Rahmenwerk etabliert, das die Erklärbarkeit von KI-Systemen über deren gesamten Lebenszyklus hinweg sicherstellt?

In der Organisation sollten angemessene Strukturen, Verfahren und Prozesse implementiert sein, um die KI-Aktivitäten zu steuern, zu verwalten sowie zu überwachen. Dadurch sollte sichergestellt werden, dass KI-bezogene Aktivitäten, Entscheidungen und Erklärungen im Einklang mit den Werten der Organisation sowie den sozialen, ethischen und rechtlichen Verantwortlichkeiten stehen.
Es sollten dabei u. a. Richtlinien und Verfahren für Leistungsmessung der KI-Systeme, KI-spezifische Schulungen, Berichterstattung und XAI-Tools festgelegt werden. Die KI-Richtlinien und -Verfahren sollten zur Gewährleistung einer gleichbleibenden Qualität für den gesamten KI-Lebenszyklus festgelegt und regelmäßig auf deren Aktualität und Angemessenheit überprüft werden [108] [109] [110].

KI-Strategie
Liegt eine angemessene Erklärbarkeitsstrategie für KI-Systeme vor?

In der Organisation sollte eine Strategie für die Entwicklung sowie den Einsatz von KI-Systemen vorliegen. Diese KI-Strategie sollte die beabsichtigen Ergebnisse der KI-Aktivitäten formulieren sowie definieren wie die KI-Ziele erreicht werden können. Somit sollte festgelegt werden, wie KI-Systeme zur Erreichung eines geschäftlichen Mehrwerts eingesetzt werden können und wie dieser Mehrwert in Bezug auf Governance, Technologie, Ressourcen und Prozesse erreicht werden kann [15] [108] [110] [109].
Die KI-Strategie sollte auch Erklärbarkeitsstrategien umfassen, die im Laufe der Design- sowie Implementierungsphase des KI-Systems zu verfolgen sind. Hierbei sollte definiert werden, wie Modelle und deren Ergebnisse erklärt werden, d. h. welche Erklärbarkeitstools bzw. Arten von Erklärungen in der Organisation zum Einsatz kommen können. Weiters sollten Umfang und Reichweite der Erklärungen festgelegt werden. Aus dem Erklärungsumfang sollte ersichtlich sein, ob einzelne Instanzen des Modells oder die zugrundeliegende Modelllogik erklärbar dargestellt werden bzw. ob KI-Systeme lokal und/oder global interpretierbar bzw. erklärbar sein sollen [15].
Die KI-Strategien sollten von den Leitungsorganen der Organisation entwickelt werden, die sowohl das beabsichtigte Ergebnis der KI-Aktivitäten artikulieren können als auch den Zusammenhang dieser Ergebnisse mit den Unternehmenszielen sowie die KI-Technologiefähigkeiten, -bestrebungen und -einschränkungen der Organisation verstehen.
Die KI-Systeme und XAI-Werkzeuge sollten regelmäßig mit der Strategie der Organisation abgeglichen werden, um deren Einhaltung durchgängig gewährleisten zu können [108].

Aktionsplan für Erklärbarkeit
Liegt ein angemessener Aktionsplan für Erklärbarkeit vor?

Es sollte ein Aktionsplan für Erklärbarkeit erstellt werden, in dem detailliert festgehalten wird, wie die Ergebnisse der Entscheidungen sowie die Erklärungen des Verhaltens von KI-Systemen den Benutzer/-innen, Entscheidungsträger/-innen bzw. anderen betroffenen Parteien bestmöglich zur Verfügung gestellt werden können.
Der Aktionsplan sollte dabei Folgendes beinhalten:

- Klar formulierte Erklärungsstrategien und einen detaillierten Plan, der die Phasen im Projektablauf festlegt, in denen der Entwurf sowie die Entwicklung dieser Strategien stattzufinden haben. Für den Fall, dass Post-hoc-Erklärbarkeitstechniken zum Einsatz kommen, sind die daraus resultierenden Anforderungen (z. B. Zurverfügungstellen einfacher und benutzerzentrierter Aufbereitung der Erklärungen) in den Strategien festzulegen.
- Einen detaillierten Zeitrahmen für die Bewertung des Fortschritts des Erklärbarkeits-Aktionsplans und eine Liste aller Verantwortlichkeiten, die für die Ausführung des Aktionsplans zu erfüllen sind [15].

Verantwortungen und Rechenschaftspflichten
Sind klare Verantwortlichkeiten und Rechenschaftspflichten für die von KI-Systemen getroffenen Entscheidungen definiert?

Die Verantwortung und Rechenschaftspflicht für die von den KI-Systemen getroffenen Aktionen und Entscheidungen sollten klar zugewiesen werden, wobei die Letztverantwortung bei der Geschäftsleitung liegen sollte. Für den Fall, dass Standardpakete mit KI-Technologie erworben werden, sind weiters Haftungsbestimmungen auf vertraglicher Ebene festzulegen.
Darüber hinaus sollten Rollen und Verantwortlichkeiten entlang des gesamten KI-Lebenszyklus, einschließlich der KI-Entwicklungs- und Betriebsaktivitäten, formal festgelegt und zugewiesen werden [7]. Hierbei sind die Anforderungen an eine angemessene Funktionstrennung zu beachten. Die entsprechenden Rollen und Verantwortlichkeiten sollten darüber hinaus in Stellenbeschreibungen und Organigrammen festgehalten werden [110] [109] [111].

KI-Kenntnisse und Fähigkeiten
Sind ausreichende KI-Kenntnisse und Fähigkeiten vorhanden bzw. werden diese regelmäßig durch angemessene Weiterbildungsmaßnahmen, wie z. B. Schulungen, gefördert?

Ein KI-Projekt erfordert neue Arten von Profilen (wie z. B. Datenwissenschaftler/-innen), die u. a. über Kompetenzen in den Bereichen Mathematik, Statistik und Programmierung verfügen. Ein Mangel an solchen Kompetenzen kann zu Schwierigkeiten bei der Aktualisierung des Modells und der Erklärung des Modellverhaltens oder auch zu einer übermäßigen Abhängigkeit von Schlüsselpersonen bzw. externen Parteien führen.
Es sollte daher sichergestellt werden, dass angemessene HR-Prozesse vorhanden sind, um Mitarbeiter zu rekrutieren, zu entwickeln bzw. langfristig in der Organisation zu halten. Diese Prozesse sollten u. a. sicherstellen, dass innerhalb der Organisation ein ausreichendes Maß an KI-Kenntnissen und Fähigkeiten vorhanden ist, um KI-Lösungen zu entwickeln und zu überwachen sowie um Erklärungstechniken anzuwenden und Erklärungen verständlich aufzubereiten.
Für die interne Umsetzung von KI-Projekten sollten multidisziplinäre Teams eingerichtet werden, an denen Entwickler/-innen, Datenwissenschaftler/-innen, Mitarbeiter/-innen aus dem Bereich der IT-Infrastruktur und Datenbankadministration, Unternehmensvertreter/-innen sowie Risiko- und Compliance-Teams (zur frühzeitigen Analyse etwaiger Risiken und Compliance-Aspekte, z. B. im Zusammenhang mit Datenschutz) beteiligt sind. Darüber hinaus sollten auch interne Auditor/-innen frühzeitig in das Projekt eingebunden

werden, um den Wissenstransfer zu erleichtern, Kontrollpunkte effizienter in das Projekt einbetten zu können und eine Verbesserung der Auditierbarkeit der finalen KI-Lösung zu erzielen [7] [15].
Alle Mitglieder dieser Teams sollten regelmäßig an Weiterbildungsmaßnahmen teilnehmen, bei denen sie vollumfänglich und entsprechend ihren Aufgaben und Verantwortlichkeiten im KI-Projekt geschult werden. Die Weiterbildungsmaßnahmen sollten unterschiedliche Formen der Aus- und Weiterbildung beinhalten, wie z. B. Präsenzschulungen oder Selbststudium [111] [109].

Für den Fall, dass Standardpakete mit KI-Technologie (wie z. B. RPA/IPA63) implementiert werden, sollte im Detail analysiert werden, welche Kenntnisse und Fähigkeiten für Wartung und Erklärung der KI-Lösung notwendig sind, sobald diese in Produktion ist [7].

Modellauswahl und Risikoanalysen
Wurden bei der Auswahl des Modells umfassende Analysen durchgeführt und wurde das von der ausgewählten KI-Lösung ausgehende Risiko evaluiert und transparent dargestellt?

Die Auswahl des am besten geeigneten KI-Ansatzes kann ein komplexes Unterfangen darstellen, weshalb dies im Rahmen einer umfassenden formalen Analyse durchgeführt werden sollte. Für diese Analyse sollten mehrere Faktoren herangezogen werden, darunter die domänenspezifischen Risiken und Bedürfnisse, die verfügbaren Datenressourcen, die Aufbereitung der Merkmale sowie die Eignung des Modells zur Lösung der Rechenaufgabe [15].
Vor allem bei sorgfältiger Datenvorverarbeitung bzw. iterativer Modellentwicklung sollte die Wahl primär auf standardisierte und interpretierbare Modelle (z. B. Entscheidungsbäume, lineare oder logistische Regression) anstelle von ausgefeilten, aber undurchsichtigen Modellen fallen [15] [107]. Black-Box-Modelle, wie die in der vorliegenden Arbeit betrachteten (nämlich SVM, Ensemble-Methoden und Neuronale Netze), sollten im Gegensatz dazu nur ausgewählt werden, „wenn ihre überlegenen Modellierungsfähigkeiten am besten zu den Eigenschaften des vorliegenden Problems passen“ [47, p. 35].

Es sollte darüber hinaus ein Prozess zur Identifizierung, Analyse, Behandlung und Überwachung von Risiken, die aufgrund des Einsatzes von KI-Systemen auftreten können, implementiert sein. Die Risikobewertung hat bei der Initiierung von bedeutenden Projektphasen und bei größeren Änderungsanforderungen zu erfolgen und sollte einen speziellen Fokus auf jene Risiken richten, die aus mangelnder Erklärbarkeit resultieren können. Die im Rahmen der Risikoanalyse identifizierten Risiken sollten zentral in einem Register zusammengefasst und kontinuierlich überwacht oder neu bewertet werden [112] [113].

Machbarkeitsstudie
Wurde für die ausgewählte KI-Lösung eine Machbarkeitsstudie durchgeführt, in deren Rahmen eine Erhebung der Anforderungen an Erklärbarkeit erfolgt?

Organisationen sollten vor dem Erwerb oder der Entwicklung von KI-Systemen eine Machbarkeitsstudie durchführen, bei welcher die Umsetzung der Anforderungen geprüft wird. Im Zuge der Erhebung sollten neben funktionalen Anforderungen auch nicht funktionale Anforderungen wie z. B. Performance, Ressourcenverbrauch und Erklärbarkeit erhoben werden. Es sollte dabei auch analysiert werden, ob die Anforderungen an Erklärbarkeit durch die Ergänzung des Systems mit XAI-Werkzeugen erfüllt werden können [113]. Im Rahmen dieser Analyse sollten potenzielle XAI-Werkzeugen im Hinblick auf ihre Fähigkeit, die Beweggründe für die Entscheidung sowie das Verhalten des KI-Systems für die Benutzer/-innen und die betroffenen Stakeholder im konkreten Anwendungsfall verständlich zu machen, bewertet werden.
Sollte es bei der Umsetzung der Anforderungen zu Spannungen kommen, können Kompromisse in Betracht gezogen werden, aber nur, wenn diese ethisch bzw. hinsichtlich ihres Risikos vertretbar sind. Solche Kompromisse sollten begründet, explizit anerkannt und dokumentiert und im Hinblick auf ihr Risiko

bewertet werden. Der Entscheidungsträger muss für die Art und Weise der angemessenen Abwägung rechenschaftspflichtig sein, und die getroffene Abwägung sollte kontinuierlich überprüft werden, um die Angemessenheit der Entscheidung sicherzustellen. Wenn es keinen akzeptablen Kompromiss gibt, sollten die Entwicklung, der Einsatz und die Nutzung des KI-Systems nicht in dieser Form fortgesetzt werden [15].

Datenqualität und Data-Governance
Ist ein Rahmenwerk zum ganzheitlichen Management von Daten, die für KI-Projekte herangezogen werden, implementiert?

Die Leistungsfähigkeit von KI und insbesondere maschinellem Lernen ist in erster Linie von den Daten abhängig, die zum Trainieren der Systeme verwendet werden. Zusätzlich wirken sich Datenressourcen, die sich für aussagekräftige und strukturierte Darstellungen eignen, positiv auf die Erklärbarkeit eines KI-Modells aus [7] [15].

Organisationen sollten daher ein Data-Governance-Framework zur Sicherstellung und Aufrechterhaltung der Qualität von KI-relevanten Daten über deren gesamten Lebenszyklus implementieren.
Dies beinhaltet u. a. die Festlegung und Zuweisung klarer Rollen und Verantwortlichkeiten für die Datenbestände (Data-Owner). Weiters sollte ein Prozess zur Identifizierung von Datenqualitätsproblemen implementiert werden, in dessen Rahmen eine Überprüfung auf fehlende oder inkonsistente Daten, Duplikate oder Daten im falschen Format erfolgen sollte. Der Prozess sollte dabei vorsehen, dass betroffene Datensätze entweder vollständig aus dem Datenumfang ausgeschlossen oder erst nach einer erfolgten Korrektur erneut aufgenommen werden. Alle während der Datenvorbereitungsphase des KI-Projekts identifizierten Datenqualitätsprobleme sollten an die zuständige Fachabteilung sowie den Data-Owner eskaliert werden, um sie direkt an der Quelle beheben zu können.
Es sollten zur Sicherstellung der Erklärbarkeit der Modellergebnisse auch jene Datentransformations-schritte formal dokumentiert und begründet werden, die von der Bearbeitung der Rohdaten bis hin zur Erstellung der Merkmale angewendet werden [7] [111] [114] [113].
Darüber hinaus sollte ein Data-Dictionary eingesetzt werden, um den Data-Owner, die Vertraulichkeitsstufe, den Speicherort, das Format, die Validierungsregeln sowie die Beziehung zu anderen Daten zu dokumentieren.
Für den Fall, dass die Daten aus externen Datenquellen übernommen werden, sollte vor deren Verwendung eine Due-Diligence-Analyse zur Überprüfung der Datenqualität, Vertrauenswürdigkeit des Anbieters sowie der Relevanz der Daten durchgeführt werden [7] [114].

Datensicherheit
Sind angemessene Maßnahmen zur Sicherstellung der Datensicherheit implementiert?

Die beabsichtigte bzw. auch unbeabsichtigte Manipulation von Daten kann dazu führen, dass KI-Systeme das Attribut der Zerlegbarkeit verlieren bzw. Erklärungen weniger informativ dargestellt werden können.
Um dem entgegenzuwirken, sollte vor Beginn des KI-Projekts eine Evaluierung relevanter Sicherheitsaspekte durchgeführt werden, um sowohl die Umsetzung von „Security by Design“ als auch des „Need-to-know-Prinzips“ sicherstellen zu können.
Ausgehend von dieser Evaluierung sollten diverse Sicherheitsmaßnahmen abgeleitet und während des gesamten KI-Entwicklungsprozesses umgesetzt werden. Diese Maßnahmen sollten neben organisatorischen Regelungen z. B. im Bereich des Änderungsmanagement-, des Entwicklungs- und des Berechtigungsmanagementprozesses auch technische Schutzmaßnahmen zur Verhinderung von unautorisierten Änderungen umfassen. Dazu zählen beispielsweise die Implementierung eines Softwareversionierungssystems, ein aktives Log- und Eventmanagementsystem bzw. die Einrichtung einer mehrstufigen Systemlandschaft. Letztere ermöglicht eine Trennung zwischen der Produktionsumgebung und jenen Umgebungen, die für Entwicklung und Testen des KI-Modells verwendet werden [7] [111] [109] [113].

Im Falle von maschinellem Lernen reichen die zu Testzwecken erstellten synthetischen bzw. anoymisierten Daten in den meisten Fällen nicht aus, um die Algorithmen korrekt zu trainieren. Folglich stellen die Daten in den Entwicklungs- bzw. Testumgebungen oft eine Kopie der Produktivdaten dar, die in einer bestimmten Frequenz, z. B. täglich, aktualisiert werden. Es sollten daher umfassende Maßnahmen zum Schutz dieser Produktivdaten implementiert werden, wie beispielsweise homomorphe Verschlüsselungstechniken bzw. Beschränkungen des Berechtigungszugriffs. So sollte die Vergabe der Zutrittsberechtigungen nach dem Prinzip der geringsten Privilegien erfolgen, d. h., Datenwissenschaftler/-innen bzw. Programmierer/-innen sollten nur Zugriff auf jene Daten erhalten, die für die Entwicklung bzw. Testung der Modelle erforderlich sind. Der Zugriff sollte dabei weitgehend lesend und nicht schreibend erfolgen, wobei jede Ausnahme hiervon begründet und angemessen genehmigt werden sollte [7] [111] [109] [113].

Wie zuvor angeführt, sollte weiters ein angemessener Änderungsmanagement-Prozess zur Überwachung und Dokumentation aller Änderungen an Datenquellen, die mit dem KI-System interagieren, implementiert sein. Der Prozess sollte sicherstellen, dass kritischere Änderungen einer vorherigen Genehmigung bedürfen, die auch eine Risikobewertung sowie eine Validierung der Datenqualität umfasst [110] [109] [113].

Erklärbarkeit durch Design

Werden im Rahmen der Designphase entsprechende Maßnahmen umgesetzt, um eine spätere Erklärbarkeit der KI-Systeme sicherstellen zu können?

Es sollten Maßnahmen ergriffen werden, um die Erklärbarkeit der KI-Systeme bereits in der Design- bzw. Entwurfsphase sicherzustellen. Diese Maßnahmen sollten Folgendes umfassen:

- Den formalen Entwurf des zur Erstellung der Eingabefunktionen verwendeten Datenaufbereitungsflusses, einschließlich jener Transformationen, die auf Rohdaten angewendet werden (z. B. Normalisierung, Dimensionalitätsreduktion, Ausschluss korrelierter Merkmale, Aggregation usw.).
- Gegebenenfalls eine Einbindung von Interpretern in den Modellentwurf, um die Nachvollziehbarkeit der relevanten internen Schritte, die zur finalen Vorhersage führen, zu gewährleisten.
- Genauigkeitsmetriken (KPI), die zur Überwachung der Modellleistung sowie zum sofortigen Identifizieren von Abweichungen eingesetzt werden können.
- Die Sicherstellung, dass die in der Produktion implementierte Lösung auditierbar ist (siehe nächsten Punkt zu Auditierbarkeit), um ein erstes Verständnis darüber zu erhalten, wie die Daten verarbeitet werden [7].

Auditierbarkeit

Wird eine durchgängige Auditierbarkeit gewährleistet, d. h., werden die wesentlichen mit der KI-Lösung im Zusammenhang stehenden Aktionen aufgezeichnet?

Da KI-Modelle in relevante Geschäftsprozesse integriert werden, müssen jene Informationen verwaltet und konsolidiert werden, die für die Gewährleistung einer durchgängigen Auditierbarkeit notwendig sind. Diese Informationen sollten sowohl Aufzeichnungen und Aktivitätsüberwachungsergebnisse als auch Modellentwicklungsdaten umfassen, die während der Modellierungs-, Test-, Schulungs-, Verifizierungs- und Implementierungsphasen gesammelt werden. Die Aufzeichnungen sollten dabei über einen ausreichenden Detailliertheitsgrad verfügen, um den Datenfluss durch den KI-Prozess (z. B. von den rohen Eingabedaten über die Berechnung bis hin zur KI-Ausgabe) verfolgen zu können bzw. um den betroffenen Parteien und Entscheidungsträgern die Vertretbarkeit der Ergebnisse des KI-Verarbeitungsverhaltens zu demonstrieren. Weiters sollte im Bedarfsfall eine erneute Simulierung der Eingabedaten möglich sein [7] [15] [110].

System- und Benutzerdokumentation
Liegen angemessene System- und Benutzerdokumentation für die Anwendung und den Betrieb von KI-Systemen und Erklärbarkeitstools vor?

Um einen reibungslosen Betrieb und eine ordnungsgemäße Anwendung von KI-Systemen, Infrastruktur und Erklärbarkeitstools gewährleisten zu können, sollten sowohl Benutzerhandbücher als auch technische Systemdokumentationen erstellt werden. Das Benutzerhandbuch sollte dabei jene Informationen enthalten, die für eine sachgerechte Bedienung des KI-Systems und der Erklärbarkeitstools erforderlich sind. Die Systemdokumentation sollte neben der technischen Systemkonfiguration auch Darstellungen der Datenströme, Regelungen für Wartungsarbeiten sowie Ansprechpartner und Verantwortlichkeiten bei Sicherheitsvorfällen bzw. Vorfällen, die aus mangelnder Erklärbarkeit resultieren, enthalten.
Weiters sollten auch der Quellcode von selbsterstellten KI-Systemen sowie Schnittstellen zu anderen Anwendungen angemessen dokumentiert sein [113] [110].

Einsatz in der Produktion
Wird die KI-Lösung vor der Implementierung in der Produktionsumgebung angemessen hinsichtlich ihrer Erklärbarkeit getestet?

Aufgrund der Komplexität der mathematischen Konzepte, die in KI-Modelle eingebettet sind, können die tatsächlichen Auswirkungen des Codes bzw. die tatsächliche Interpretierbarkeit/Erklärbarkeit vor dessen Implementierung in der Produktionsumgebung zumeist nicht abgeschätzt werden. Folglich sollten vor dem Go-live technische Tests durchgeführt werden, um z. B. die Schnittstellen zu anderen Systemen zu validieren. Zusätzlich sollte auch umfangreich getestet werden, ob das Modellverhalten mittels der eingeplanten XAI-Methoden umfassend erklärt werden kann, wobei alle durchgeführten Tests angemessen und nachvollziehbar dokumentiert werden sollten [7] [111] [109].
Ergänzend dazu sollte die Leistung des KI-Modells, sobald es sich in Produktion befindet, zur Erkennung von Verschlechterungen der Modellleistungen kontinuierlich überwacht werden [7]. Dies könnte im Falle von Black-Box-Modellen durch einen regelmäßigen Einsatz von Post-hoc-Erklärungstechniken erfolgen.

Unabhängiges Audit
Werden Algorithmen, Daten und Entwurfsprozesse in angemessenen Abständen (abhängig von ihrem Risikogehalt) durch die Interne Revision bzw. externe Auditor/-innen bewertet?

Die Einhaltung der mit KI-Systemen im Zusammenhang stehenden Prozesse und Kontrollen sollte in regelmäßigen Abständen d. h. abhängig vom Risiko des jeweiligen Systems durch die Interne Revision bewertet werden. Es sollten dabei formale Risikoanalysen durch die Interne Revision durchgeführt werden, aus denen die Prüffrequenz sowie der Umfang der Prüfungshandlungen nachvollziehbar abgeleitet werden. Der Umfang der Prüfungshandlungen sollte ausreichend detailliert sein, damit Auditor/-innen in der Lage sind, die Vertrauenswürdigkeit des KI-Systems einzuschätzen [115] [116].
Es sollte zusätzlich eine externe dritte Partei mit der Auditierung des KI-Systems betraut werden, falls es sich dabei um eine sicherheitskritische Anwendung handelt bzw. die Grundrechte der betroffenen Personen durch die KI-Entscheidungen berührt werden [115].

Internes Kontrollsystem (IKS)

Ist das interne Kontrollsystem ausreichend ausgestaltet und formalisiert, um die Erklärbarkeit des Verhaltens von KI-Systemen über deren gesamten Lebenszyklus sicherstellen zu können?

Organisationen sollten die Integration der KI in den Geschäftsprozess sorgfältig prüfen und angemessene IKS-Kontrollen zur Sicherstellung der kontinuierlichen Erklärbarkeit des Verhaltens der KI-Systeme implementieren [7] [15]. So sollten u. a. formelle Kontrollen im Bereich des Änderungsmanagements (z. B. Kontrolle zur Identifizierung von unautorisierten Änderungen oder zum Review der Logfiles), im Bereich des Berechtigungsmanagements (z. B. Kontrolle zur Sicherstellung eines regelmäßigen und vollständigen Reviews der Zugriffsberechtigungen für die jeweiligen Umgebungen), im Bereich des Datenmanagements (z. B. Kontrolle zur Überprüfung der Genauigkeit und Vollständigkeit der vom KI-System verwendeten Datensätze) oder im Zusammenhang mit der Revision (z. B. Kontrolle zum Review des Prüfplans) im IKS definiert werden [111] [109] [113].
Die Durchführung der Kontrollhandlungen sollte angemessen und nachvollziehbar dokumentiert und es sollte eine institutionalisierte IKS-Berichterstattung implementiert werden.
Weiters sollte eine zentrale und institutionalisierte Überprüfung erfolgen, die sicherstellt, dass die implementierten Kontrollhandlungen auch tatsächlich durchgeführt werden.

5. Schlussbetrachtungen

Ziel dieser Arbeit war es, einen Überblick über mögliche Methoden und Tools zur Sicherstellung der Erklärbarkeit von KI-Modellen vorzustellen. Um dieses Ziel zu erreichen, wurde eine umfassende Literaturrecherche durchgeführt, in deren Rahmen die Ergebnisse der renommiertesten Forschungsstudien zu XAI-Tools für überwachtes Lernen bzw. Berichte des Einsatzes dieser Tools in der Praxis zusammengetragen wurden. Ergänzend wurden diverse IT-Standards, wie das IT-Grundschutz-Kompendium und ISO2700X, herangezogen, um ein Prüfrahmenwerk zur Bewertung der Umsetzung von XAI-Maßnahmen und -Praktiken in Unternehmen auszuarbeiten.

In Bezug auf die Hauptforschungsfrage kann festgehalten werden, dass unterschiedliche Methoden existieren, um die Erklärbarkeit von ML-Modellen des überwachten Lernens sowohl während deren Entwurf und Entwicklung als auch post hoc, während des Betriebs, sicherzustellen. So können für die Post-hoc-Erklärbarkeit modellspezifische oder auch modellagnostische, d. h. für jedes Modell des überwachten Lernens anwendbare Techniken, zum Einsatz kommen. Auch wenn viele dieser Techniken in der Theorie Nachteile mit sich bringen, können sie bei geeigneter Auswahl und Anpassung an das Anwendungsszenario gute Ergebnisse erzielen. Besonders bei KNN, die mit wenig bearbeiteten Merkmalen arbeiten, können Repräsentationen mit Hilfe von Post-hoc-Techniken anschaulich verbalisiert bzw. visualisiert werden.
Zusätzlich können Überprüfungen durchgeführt werden, ob angemessene Verfahren, Prozesse und Kontrollen im Unternehmen implementiert sind, um die Erklärbarkeit des KI-Systems über dessen gesamten Lebenszyklus hinweg sicherzustellen. Diese Überprüfungen können mit Hilfe des im Rahmen der vorliegenden Arbeit vorgestellten Prüfrahmenwerk auch von Auditor/-innen ohne umfassende KI-Kenntnisse durchgeführt werden.

Auch wenn in dieser Arbeit eine Vielzahl an Post-hoc-Erklärbarkeitstechniken für KI-Modelle vorgestellt wird, liegt der Fokus ausschließlich auf Modellen des überwachten Lernens. Im Rahmen einer auf den Erkenntnisgewinnen dieser Arbeit aufbauenden Analyse könnte folglich eine Aufstellung von Erklärbarkeitstechniken für ML-Modelle des unüberwachten bzw. des bestärkenden Lernens angefertigt werden.
Aufgrund der methodischen Beschränkung der vorliegenden Arbeit auf eine Literaturanalyse kann derzeit weiters keine Auskunft über die Anwendbarkeit des Prüfrahmenwerks in der Praxis getroffen werden. Um dies zu gewährleisten, könnten weitere methodische Ansätze zum Einsatz kommen. So könnten beispielsweise Fallstudien zur Anwendung des vorgestellten Prüfrahmenwerks in der Praxis bzw. qualitative Experteninterviews durchgeführt werden, um die Vollständigkeit und Angemessenheit des Rahmenwerks zu überprüfen und sicherzustellen.
Aufbauend auf diese Fallstudien und Interviews könnten die im Rahmen des Prüfrahmenwerks beschriebenen Prozesse zur Sicherstellung von XAI in der Design-, Implementierungs- und Betriebsphase zur besseren Veranschaulichung im Detail dargestellt und formell abgebildet werden. Auch die notwendigen Kontrollen könnten ausformuliert werden, das heißt u. a. um eine angemessene Kontrollbeschreibung, Frequenz und die Verantwortlichkeiten (Durchführung sowie Kontrolle der Durchführung) ergänzt werden. Hierdurch könnte ein umfassendes Internes Kontrollsystem für XAI aufgebaut werden, das sowohl eine vereinfachte Optimierung der mit XAI im Zusammenhang stehenden Prozesse als auch eine Steuerung der aus mangelnder Erklärbarkeit von KI-Systemen resultierenden Risiken ermöglicht.

Literaturverzeichnis

[1] W. Samek und K.-R. Müller, „Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Cham, Springer, 2019, pp. 5-22.

[2] M. Turek, „Explainable Artificial Intelligence (XAI)," [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence. [Zugriff am 30 10 2020].

[3] W. Samek, K.-R. Müller und T. Wiegand, „Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *ITU Journal: ICT Discoveries,* Bd. 1, Nr. Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services, pp. 1-10, 2017.

[4] E. Lindholm , J. Nickolls , S. Oberman und J. Montrym, „NVIDIA Tesla: A Unified Graphics and Computing Architecture," *IEEE Micro,* Bd. 28, Nr. 2, pp. 39-55, 2008.

[5] A. Karpathy, G. Toderici, S. Sanketh, T. Leung, R. Sukthankar und L. Fei-Fei, „Large-Scale Video Classification with Convolutional Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.

[6] C. Wasner und J. Traxler, „KI und Sicherheit im Automobil," *ATZ Elektronik,* Nr. 14, pp. 50-53, 2019.

[7] A. Curridori, „Artificial Intelligence: Opportunities, Risks and Recommendations for the Financial Sector," Commission de Surveillance du Secteur Financier, Luxemburg, 2018.

[8] Verband der TÜV, „Sicherheit und Künstliche Intelligenz: Erwartungen, Hoffnungen, Emotionen," 2020.

[9] ISACA, „Auditing Artificial Intelligence," 2018.

[10] M. T. Ribeiro, S. Singh und C. Guestrin, „Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[11] S. Lundberg und S.-I. Lee, „A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems,* pp. 4765-4774, 2017.

[12] J. H. Friedman, „Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics,* Bd. 29, Nr. 5, pp. 1189-1232, 2001.

[13] J. R. Zilke, E. L. Mencía und F. Janssen, „DeepRED–Rule Extraction from Deep Neural Networks," in *International Conference on Discovery Science*, 2016.

[14] R. C. LaBrie und G. H. Steinke, „Towards a Framework for Ethical Audits of AI Algorithms," in *Twenty-fifth Americas Conference on Information Systems*, Cancun, 2019.

[15] D. Leslie, „Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector," The Alan Turing Institute, 2019.

[16] Z. C. Lipton, „The Mythos of Model Interpretability," *Queue,* Nr. 16(3), pp. 31-57, 2018.

[17] J. N. Kok, E. J. W. Boers, W. A. Kosters und P. van der Putten, „Artificial Intelligence: Definition, Trends, Techniques and Cases," *Artificial Intelligence,* Nr. 1, pp. 1-20, 2009.

[18] The Merriam-Webster Dictionary, Merriam-Webster Inc., 2016.

[19] F. S. B. (FSB), „Artificial Intelligence and Machine Learning in Financial Services - Market Developments and Financial Stability Implications," 2017.

[20] M. Nadimpalli, „Artificial Intelligence Risks and Benefits," *International Journal of Innovative Research in Science, Engineering and Technology,* Bd. 6, Nr. 6, 2017.

[21] A. Hintze, „Understanding the Four Types of AI, from Reactive Robots to Self Artificial Intelligence," *Government and Technology Online,* 2016.

[22] The Institute of Internal Auditors, „Artificial Intelligence – Considerations for the Profession of Internal Auditing," 2017.

[23] T. Gonsalves, „The Summers and Winters of Artificial Intelligence," in *Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction*, USA, IGI Global, 2019, pp. 168-179.

[24] I. Döbel, M. Leis, M. M. Vogelsang, D. Neustroev, H. Petzka, A. Riemer, S. Rüping, A. Voss, M. Wegele und J. Welz, „Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung," Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., München, 2018.

[25] J. Lighthill, „Artificial Intelligence: A General Survey," *Artificial Intelligence: a Paper Symposium, Science Research Council,* 1973.

[26] M. Minsky und S. Papert, Perceptrons: An Introduction to Computational Geometry, USA: The Science Press, Inc., 1969.

[27] L. Dormehl, Thinking Machines: The Quest for Artificial Intelligence - and Where It's Taking Us Next, London: Penguin Books, 2016.

[28] K. Kelly, „The Three Breakthroughs That Have Finally Unleashed AI on the World," 27 10 2014. [Online]. Available: https://www.wired.com/2014/10/future-of-artificial-intelligence/. [Zugriff am 11 2020].

[29] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg und A. Holzinger, „Explainable AI: The New 42?," *CD-MAKE 2018: Machine Learning and Knowledge Extraction: International Cross-Domain Conference for Machine Learning and Knowledge Extraction,* pp. 295-303, 2018.

[30] Statista Research Department, „Umsatz mit Business-Anwendungen im Bereich künstliche Intelligenz weltweit bis 2025," 12 09 2016. [Online]. Available: https://de.statista.com/statistik/daten/studie/620443/umfrage/umsatz-mit-unternehmensanwendungen-im-bereich-kuenstliche-intelligenz-weltweit/. [Zugriff am 02 11 2020].

[31] European Commission, „Mitgliedstaaten und Kommission arbeiten gemeinsam an Förderung künstlicher Intelligenz „Made in Europe"," 07 12 2018. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/de/IP_18_6689. [Zugriff am 06 11 2020].

[32] P. Langley und J. E. Laird, „Artificial Intelligence and Intelligent Systems," 2006.

[33] DIN e.V. (Hrsg.), „ISO/IEC 38505-1:2017-04, Informationstechnik - IT-Betriebsführung - Teil 1: Die Anwendung von ISO/IEC 38500 zur Betriebsführung von Daten," Beuth-Verlag, Berlin, 2017.

[34] K. P. Murphy, Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning), Cambridge: MIT Press, 2012.

[35] I. H. Witten, E. Frank und M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, USA: Morgan Kaufmann, 2016.

[36] A. C. Müller, S. Guido und K. Rother, Einführung in Machine Learning mit Python Praxiswissen Data Science, Heidelberg: O'Reilly, 2017.

[37] S. Marsland, Machine Learning: An Algorithmic Perspective, CRC Press, 2015.

[38] A. Welsch, V. Eitle und P. Buxmann, „Maschinelles Lernen: Grundlagen und betriebswirtschaftliche Anwendungspotenziale am Beispiel von Kundenbindungsprozessen," *HMD Praxis der Wirtschaftsinformatik,* Nr. 55, p. 366–382, 2018.

[39] S. Russell und P. Norvig, Artificial Intelligence: A Modern Approach, USA: Prentice Hall, 2010.

[40] A. Mellouk, Advances in Reinforcement Learning, BoD--Books on Demand, 2011.

[41] K. R. Chowdhary, „Natural Language Processing," in *Fundamentals of Artificial Intelligence*, New Delhi, Springer India, 2020, pp. 603-649.

[42] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter und L. Kagal, „Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Italy, 2018.

[43] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson und H. Yu, „Accountable Algorithms," *U. Pa. L. Rev.,* Nr. 165, pp. 633, 656, 2016.

[44] House of Lords, Select Committee on Artificial Int, „Report of Session 2017-19, AI in the UK: Ready, Willing, and Able?," 16 04 2018. [Online]. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf. [Zugriff am 14 11 2020].

[45] V. Fink, „Kontrolle ist besser – Transparenz werteorientiert gestalten," in *Quick Guide KI-Projekte – einfach machen*, Wiesbaden, Springer Gabler, 2020, pp. 123-134.

[46] D. Gunning, „Explainable Artificial Intelligence (XAI)," Tech. rep., Defense Advanced Research Projects Agency (DARPA), 2017.

[47] A. Barredo Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila und F. Herrera, „Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion,* Nr. 58, pp. 82-115, 2020.

[48] A. Adadi und M. Berrada, „Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access,* Nr. 6, pp. 52138-52160, 2018.

[49] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek und K.-R. Müller, „Unmasking Clever Hans Predictors and Assessing what Machines Really Learn," *Nature communications,* Bd. 1, Nr. 10, pp. 1-8, 2019.

[50] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller und W. Samek, „Analyzing Classifiers: Fisher Vectors and Deep Neural Networks," in *2016 IEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016.

[51] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche und Graepe, „Mastering the Game of Go Without Human Knowledge," *Nature,* Nr. 550(7676), pp. 354-359, 2017.

[52] M. W. Libbrecht und W. S. Noble, „Machine Learning Applications in Genetics and Genomics," *Nature Reviews Genetics,* Nr. 16, p. 321–332, 2015.

[53] S. Lemm, B. Blankertz, T. Dickhaus und K.-R. Müller , „Introduction to Machine Learning for Brain Imaging," *Neuroimage,* Nr. 56(2), pp. 387-399, 2011.

[54] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran und R. Ramprasad, „Accelerating Materials Property Predictions Using Machine Learning," *Scientific Reports,* Nr. 3, p. Article number: 2810, 2013.

[55] J. Angwin, J. Larson, S. Mattu und L. Kirchner, „Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.," 23 05 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Zugriff am 30 11 2020].

[56] A. Howard, C. Zhang und E. Horvitz, „Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, Austin, Texas, USA, 2017.

[57] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm und N. Elhadad, „Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australien, 2015.

[58] Pegasystems, „Consumers Failing to Embrace AI Benefits, Says Research: Pega Study Highlights the Need for Greater Empathy in Artificial Intelligence Systems," 04 06 2019. [Online]. Available:

https://www.pega.com/about/news/press-releases/consumers-failing-embrace-ai-benefits-says-research. [Zugriff am 30 11 2020].

[59] A. Weller, „Transparency: Motivations and Challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Cham, Springer, 2019.

[60] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2020.

[61] R. L. Heath und J. Bryant, Human Communication Theory and Research: Concepts, Contexts, and Challenges, USA: Routledge, 2013.

[62] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Shieber, J. Waldo, D. Weinberger, A. Weller und A. Wood, „Accountability of AI Under the Law: The Role of Explanation," Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society Working Paper, 2017.

[63] B. Goodman und S. Flaxman, „European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation," *AI Magazine,* Bd. 38, Nr. 3, pp. 50-57, 2017.

[64] *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG.*

[65] P. Norvig, „Google's Approach to Artificial Intelligence and Machine Learning," 22 06 2017. [Online]. Available: https://www.engineering.unsw.edu.au/file/googles-approach-to-artificial-intelligence-and-machine-learning-a-conversation-with-peter. [Zugriff am 29 11 2020].

[66] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti und D. Pedreschi, „A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys (CSUR),* Nr. 51, pp. 1-42, 2018.

[67] C. Krech, „Erklärbarkeit maschineller Lernverfahren: Feature Engineering mit Black-Box-Modellen (Masterarbeit, Data Science)," Hochschule Darmstadt, 2019.

[68] E. Alpaydin, Maschinelles Lernen, Bd. 27, USA: Walter de Gruyter, 2019, pp. 221-234.

[69] E. Bauer und R. Kohavi, „An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants, Machine Learning," *Machine Learning,* Bd. 36, Nr. 1-2, pp. 105-139, 1999.

[70] A. Saabas, „Interpreting Random Forests," 19 10 2014. [Online]. Available: http://blog.datadive.net/interpreting-random-forests/. [Zugriff am 25 12 2020].

[71] M. Von der Hude, „K-Nächste Nachbarn (K Nearest Neighbours)," in *Predictive Analytics und Data Mining*, Wiesbaden, Springer Vieweg, 2020, pp. 99-106.

[72] L. Magdalena, „Fuzzy Rule-Based Systems," in *Springer Handbook of Computational Intelligence*, Berlin, Heidelberg, Springer, 2015, pp. 203-218.

[73] S. Otto, „Bayessche Netzwerke: Proseminar: Machine Learning," 27 06 2006. [Online]. Available: http://www.cogsys.cs.uni-tuebingen.de/lehre/ss06/pro_learning/Bayesnetze_SteffenOtto.pdf. [Zugriff am 12 12 2020].

[74] P. Cortez und M. J. Embrechts, „Using Sensitivity Analysis and Visualization Techniques to open Black Box Data Mining Models," *Information Sciences,* Bd. 225, pp. 1-17, 2013.

[75] A. Bennetot, J.-L. Laurent, R. Chatila und N. Díaz-Rodríguez, „Towards Explainable Neural-Symbolic Visual Reasoning," in *IJCAI19 Neural-Symbolic Learning and Reasoning Workshop*, Macau, China, 2019.

[76] R. König, U. Johansson, T. Löfström und L. Niklasson, „Improving GP Classification Performance by Injection of Decision Trees," in *2010 IEEE Congress on Evolutionary Computation (CEC)*, 2010.

[77] R. Konig, U. Johansson und L. Niklasson, „G-REX: A Versatile Framework for Evolutionary Data Mining," in *2008 IEEE International Conference on Data Mining Workshops*, 2008.

[78] C. Ambi, „SHapley Additive Erklärungen," 20 11 2020. [Online]. Available: https://ichi.pro/de/so-erklaren-sie-ihre-vorhersagen-zum-maschinellen-lernen-mit-shap-werten-184937739082257. [Zugriff am 2020 12 22].

[79] P. Cortez und M. J. Embrechts, „Opening Black Box Data Mining Models Using Sensitivity Analysis," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2011.

[80] L. Breiman, J. Friedman, C. J. Stone und R. A. Olshen, Classification and Regression Trees, CRC press, 1984.

[81] A. Henelius, K. Puolamäki und A. Ukkonen, „Interpreting Classifiers through Attribute Interactions in Datasets," *arXiv:1707.07576,* 2017.

[82] A. Goldstein, A. Kapelner, J. Bleich und E. Pitkin, „Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics,* Bd. 24, Nr. 1, pp. 44-65, 2015.

[83] D. Smilkov, N. Thorat, B. Kim, F. Viégas und M. Wattenberg, „SmoothGrad: Removing Noise by Adding Noise," in *Workshop on Visualization for Deep Learning, ICML*, 2017.

[84] D. Alvarez-Melis und T. S. Jaakkola, „On the Robustness of Interpretability Methods," *Workshop on Human Interpretability in Machine Learning,* 2018.

[85] G. Montavon, W. Samek und K.-R. Müller, „Methods for Interpreting and Understanding Deep Neural Networks," *Digit. Signal Process,* Nr. 73, p. 1–15, 2018.

[86] Q. Zhao und T. Hastie, „Causal Interpretations of Black-Box Models," *Journal of Business and Economic Statistics,* Nr. 1, pp. 1-19, 2019.

[87] L. Breiman, „Consistency for a Simple Model of Random Forests," Statistics Department University of California, Berkeley, 9 09 2004. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf. [Zugriff am 25 12 2020].

[88] T. Hastie, R. Tibshirani und J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2017.

[89] W. S. Noble, „What is a Support Vector Machine?," *Nature Biotechnology,* Nr. 24, p. 1565–1567, 2006.

[90] N. H. Barakat und A. P. Bradley, „Rule Extraction from Support Vector Machines: A Sequential Covering Approach," *IEEE Transactions on Knowledge and Data Engineering,* Nr. 19(6), pp. 729-741, 2007.

[91] N. H. Barakat und J. Diederich, „Eclectic Rule-Extraction from Support Vector Machines," *International Journal of Computer and Information Engineering,* Nr. 2(5), pp. 1672-1675, 2008.

[92] A. Da Costa F. Chaves, M. M. B. R. Vellasco und R. Tanscheit, „Fuzzy Rule Extraction from Support Vector Machines," in *International Conference on Hybrid Intelligent Systems, IEEE*, 2005.

[93] X. Fu, C. Ong, S. Keerthi, G. G. Hung und L. Goh, „Extracting the Knowledge Embedded in Support Vector Machines," in *IEEE International Joint Conference on Neural Networks*, Budapest, 2004.

[94] H. Núñez, C. Angulo und A. Català, „Rule Extraction from Support Vector Machines," *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (EASANN),* pp. 107-112, 2002.

[95] H. Núñez, C. Angulo und A. Català, „Rule-based Learning Systems for Support Vector Machines," *Neural Processing Letters,* Nr. 24(1), pp. 1-18, 2006.

[96] H. Nuñez, C. Angulo und A. Català, „Support Vector Machines with Symbolic Interpretation," in *VII Brazilian Symposium on Neural Networks*, 2002.

[97] B. Üstün , W. J. Melssen und L. Buydens, „Visualisation and Interpretation of Support Vector Regression Models," *Analytica Chimica Acta,* Nr. 595(1-2), pp. 299-309, 2007.

[98] L. Rosenbaum, G. Hinselmann, A. Jahn und A. Zell, „Interpreting Linear Support Vector Machine Models with Heat Map Molecule Coloring," *Journal of Cheminformatics,* Nr. 3(1), p. 11, 2011.

[99] B. Gaonkar, R. T. Shinohara, C. Davatzikos und Alzheimers Disease Neuroimaging Initiative, „Interpreting Support Vector Machine Models for Multivariate Group Wise Analysis in Neuroimaging," *Medical Image Analysis,* Nr. 24(1), pp. 190-204, 2015.

[100] M. Sato und H. Tsukimoto, „Rule Extraction from Neural Networks via Decision Tree Induction," in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222), Vol. 3, IEEE*, 2001.

[101] G. Montavon, S. Bach, A. Binder, W. Samek und K.-R. Müller, „Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition," *Pattern Recognition,* Nr. 65, pp. 211-222, 2017.

[102] A. Shrikumar, P. Greenside, A. Shcherbina und A. Kundaje, „Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," 2016.

[103] R. R. Hoffman, S. T. Mueller, G. Klein und J. Litman, „Metrics for Explainable AI: Challenges and Prospects," 2018.

[104] S. Mohseni, N. Zarei und E. D. Ragan, „A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," 2018.

[105] Z. C. Lipton, D. C. Kale und R. Wetzel, „Modeling Missing Data in Clinical Time Series with RNNs: Improved Classification of Clinical Time Series," in *Machine Learning for Healthcare Conference*, 2016.

[106] Y. Lou, R. Caruana und J. Gehrke, „Intelligible Models for Classification and Regression," in *KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.

[107] C. Rudin, „Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models instead," *Nature Machine Intelligence,* Nr. 1, p. 206–215, 2019.

[108] The Institute of Internal Auditors, „The IIA's Artificial Intelligence Auditing Framework – Practical Applications, Part A," 2017.

[109] DIN e.V. (Hrsg.), „ISO/IEC 27001:2017, Informationstechnik - Sicherheitsverfahren - Informationssicherheitsmanagementsysteme - Anforderungen (ISO/IEC 27001:2013 einschließlich Cor 1:2014 und Cor 2:2015)," Beuth-Verlag, Berlin, 2017.

[110] ISACA, „COBIT 5: a Business Framework for the Governance and Management of Enterprise IT," Illinois, 2012.

[111] Bundesamt für Sicherheit in der Informationstechnik (BSI), „IT-Grundschutz Kompendium," Bonn, 2020.

[112] Bundesamt für Sicherheit in der Informationstechnik (BSI), „BSI-Standard 200-3: Risikoanalyse auf der Basis von IT-Grundschutz," Bonn, 2017.

[113] European Banking Authority (EBA), „Leitlinien für das Management von IKT und Sicherheitsrisiken," 2019.

[114] ISACA, „Rethinking Data Governance and Management: A Practical Approach for Data-Driven Enterprises," 2020.

[115] High Level Expert Group on Artificial Intelligence, „Ethics Guidelines for Trustworthy AI," Tech. rep., European Commission, 2019.

[116] Bundesamt für Sicherheit in der Informationstechnik (BSI), „Informationssicherheitsrevision: Ein Leitfaden für die IS-Revision auf Basis von IT-Grundschutz," Bonn, 2018.