Erklärbare KI

Einführung, Motivation, Herausforderungen

KARL
Künstliche Intelligenz
für Arbeit und Lernen

Kompetenzzentren
Arbeitsforschung

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Erklärbare KI

Einführung, Motivation, Herausforderungen

– Veröffentlichung: Februar 2024

Inhaltsverzeichnis

Abkürzungsverzeichnis

DSGVO ..... Datenschutz-Grundverordnung
EU ..... Europäische Union
KI ..... künstliche Intelligenz
ML ..... maschinelles Lernen
SHAP ..... Shapley Additive Explanations
XAI ..... erklärbare künstliche Intelligenz
XUI ..... Benutzeroberfläche für Erklärungen

Autor:innen

Die Inhalte wurden in Zusammenarbeit des Fraunhofer-Instituts für Optronik, Systemtechnik und Bildauswertung (IOSB) und des Instituts für Lernen und Innovation in Netzwerken (ILIN) der Hochschule Karlsruhe erarbeitet.

Verantwortliche Autor:innen:

– Dr. Jutta Hild, Maximilian Becker

Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB, Fraunhoferstraße 1, 76131 Karlsruhe

jutta.hild@iosb.fraunhofer.de, maximilian.becker@iosb.fraunhofer.de

– Robin Weitemeyer

Hochschule Karlsruhe, Moltkestraße 30, 76133 Karlsruhe

robin.weitemeyer@h-ka.de

Projekt-Webseite:

https://kompetenzzentrum-karl.de/

1. Einführung

Um erklärbare künstliche Intelligenz (engl. explainable artificial intelligencen, XAI) verstehen und ihren Nutzen nachvollziehen zu können, werden hier zunächst grundlegende Begriffe, Konzepte und die Hauptproblematik der Black Box-Eigenschaft vieler KI-Modelle, die man mit dem Einsatz von XAI lösen möchte, eingeführt.

1.1. Künstliche Intelligenz vs. Maschinelles Lernen

Der Begriff künstliche Intelligenz (KI) ist nicht einheitlich definiert. Das liegt vor allem daran, dass sich dieses Gebiet interdisziplinär und über mehrere Jahrzehnte hinweg entwickelt hat.

Eine für die Praxis nützliche, kurze Definition, an der das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) mitgewirkt hat, ist folgende:

„Künstliche Intelligenz ist die Eigenschaft eines IT-Systems, „menschenähnliche“, intelligente Verhaltensweisen zu zeigen.“ (Weber und Buschbacher 2017)

Sinngemäß ähnlich, aber etwas ausführlicher definiert das Gabler-Wirtschaftslexikon Künstliche Intelligenz als:

„Erforschung „intelligenten” Problemlösungsverhaltens sowie die Erstellung „intelligenter” Computersysteme. Künstliche Intelligenz beschäftigt sich mit Methoden, die es einem Computer ermöglichen, solche Aufgaben zu lösen, die, wenn sie vom Menschen gelöst werden, Intelligenz erfordern.“ (Lackes und Markus 2018)

Was genau „intelligent“ bedeutet, wird je nach Anwendungsgebiet unterschiedlich beantwortet. Eine sehr frühe Definition aus dem Gebiet der Sprachinteraktion sieht ein System als intelligent an, wenn der Mensch nicht unterscheiden kann, ob sein Gegenüber ein anderer Mensch oder eine Maschine ist („Turing-Test“) (Turing 1950).

Gewöhnlich wird zwischen verschiedenen historischen Phasen bei der KI-Entwicklung unterschieden (Weber und Buschbacher 2017; Matzka 2021; Wahlster 2017). In der ersten Phase bis etwa zum Jahr 1970 standen heuristische Systeme (Heuristische Such- und Schlussfolgerungsverfahren) im Fokus. Ein Beispiel für ein System mit intelligentem Verhalten war etwa ein Taschenrechner (da er Aufgaben löst, für die ein Mensch Intelligenz aufwenden müsste).

In der zweiten Phase bis etwa zum Jahr 1990 standen wissensbasierte Systeme wie Expertensysteme im Fokus. Hier wurden von Menschen manuell Wissensbasen erstellt, auf denen dann maschinelle Wissensverarbeitung betrieben wurde. Heuristische und wissensbasierte Systeme zeigten intelligentes Verhalten. Solche Systeme basieren darauf, dass der Mensch je nach An-

wendungsfall geeignete Repräsentationen von Wissen sowie Regeln, nach denen aus dem Wissen intelligentes Verhalten als Systemausgabe entsteht, manuell erstellt, d.h. klassisch programmiert. In dieser Zeit wurden zudem im Fachgebiet der Mensch-Computer-Interaktion erste kognitive Modellierungsmethoden entwickelt, die intelligentes Verhalten des Menschen, das aus den Fähigkeiten Wahrnehmung, Denken/Kognition und Handeln zusammengesetzt ist, als Analogie zu Informationseingabe, Informationsverarbeitung und Informationsausgabe von Computersystemen betrachten und daraus Methoden zur menschlichen Leistungsprädiktion ableiten (Card 2018).

In der dritten Phase bis etwa zum Jahr 2010 standen lernende Systeme im Fokus. KI-Verfahren umfassten jetzt auch die menschliche Fähigkeit des Lernens. Dafür etablierte sich der Begriff maschinelles Lernen (ML). Dabei wird ein Computersystem mithilfe statistischer Methoden befähigt, aus Daten zu lernen, bspw. indem es die Lösung einer Aufgabe schrittweise verbessert, ohne dafür explizit programmiert worden zu sein. Voraussetzung dafür ist, dass sehr große Mengen an Daten (Massendaten) vorliegen. Maschinelles Lernen ist gegenwärtig die am häufigsten genutzte Methode im Forschungsgebiet der künstlichen Intelligenz. Während bei heuristischen und wissensbasierten Systemen der Mensch das Modell für die Problemlösung selbst erstellt, sind die Möglichkeiten zu modellieren beim maschinellen Lernen eingeschränkter. Der Mensch wählt jetzt nur eine ML-Methode, bspw. eine Support-Vektor-Maschine oder ein neuronales Netz, wählt zusätzlich die sogenannten Hyperparameter (i.e. Parameter, die vor dem Modelltraining festgelegt werden und den Lernprozess des Algorithmus steuern, bspw. die Lernrate oder die Netz-Architektur) und führt dieser Methode Massendaten zu. Die Methode lernt dann auf Basis der Daten das Modell, das das Problem löst.

Man unterscheidet beim maschinellen Lernen zwischen überwachtem, unüberwachtem und verstärkendem Lernen. Beim überwachten Lernen liegen Eingabedaten und Ausgabedaten des Systems vor. Ein Beispiel ist die Klassifikation von Tieren in Bildern. Eingabe für das maschinelle Lernverfahren sind hier Bilder mit verschiedenen Tieren. Jedes Bild ist mit dem korrekten Tiernamen gekennzeichnet (man sagt auch „gelabelt“). Die Tiernamen stellen die möglichen Ausgaben des Systems dar. Das Lernverfahren lernt (man sagt auch „trainiert“) die Zuordnung von Bildern und Tiernamen. Das Ziel ist, dass das Lernverfahren nach dem Training auch ein neues Tierbild korrekt klassifiziert und den passenden Tiernamen als Ausgabe liefert. Beim unüberwachten Lernen sind im Trainingsdatensatz nur Eingabedaten, jedoch keine Ausgabedaten gegeben (man sagt auch, die Eingabedaten sind „ungelabelt“). Das Lernverfahren hat hier das Ziel, Muster und Beziehungen in den Eingabedaten zu entdecken. Beim verstärkenden Lernen wird das Lernverfahren für erfolgreiches Vorgehen belohnt und für erfolgloses Vorgehen bestraft. Auf diese Weise „lernt ein Algorithmus, etwas zu tun“ (Wilmott 2020). Für überwachtes, unüberwachtes und verstärkendes Lernen existieren jeweils eine Vielzahl verschiedener Algorithmen.

Ein Problem des vollautomatischen maschinellen Lernens ist, dass die Verfahren wie eine Black Box agieren. Das heißt, die Systemausgaben sind unter Umständen nur schwierig nachvollzieh-

bar und noch schwieriger sind Systemausgaben „[…] zu korrigieren oder zukünftig zu unterbinden“ (Weber und Buschbacher 2017). Aus diesem Grund etablierte sich parallel der Forschungszweig der erklärbaren künstlichen Intelligenz, der sich mit Verfahren befasst, Modelle des maschinellen Lernens und ihre Systemausgaben nachvollziehbar zu machen (Burkart und Huber 2021).

In der vierten Phase ab etwa dem Jahr 2010 stehen Kognitive Systeme im Fokus. Jetzt werden Lernverfahren mit wissensbasierten Methoden kombiniert. Das Ziel ist hier ebenfalls, der Black Box-Eigenschaft maschineller Lernverfahren entgegenzuwirken. Die gleichzeitige Nutzung von Expertenwissen bringt Kontrolle ins Gesamtsystem mit dem Ziel, dass dieses „[…] dann auch bei unsicherer Faktenlage ähnlich gut wie ein Mensch handeln kann.“ (Weber und Buschbacher 2017).

1.2. Das Black Box-Problem

Viele Erfolge in den Bereichen Bilderkennung und Sprachverarbeitung sind neuronalen Netzen zu verdanken. Diese KI-Modelle besitzen bei all ihrem Potenzial jedoch auch einen gravierenden Nachteil: ihre Black Box-Eigenschaft (Adadi und Berrada 2018). Neuronale Netze sind mathematische Konstrukte. Die Netzstruktur dieser Modelle stellt eine Verknüpfung von numerischen Funktionen dar, in denen sich die Logik der Entscheidungsfindung verbirgt. Aufgrund der für Menschen nicht direkt verständlichen Repräsentation dieser Logik, ist es ihnen nicht möglich, die Abbildung der Eingabedaten auf die KI-Ausgabe nachzuvollziehen (London 2019). Insbesondere darin enthaltene kausale Zusammenhänge bleiben dem Menschen deshalb verborgen.

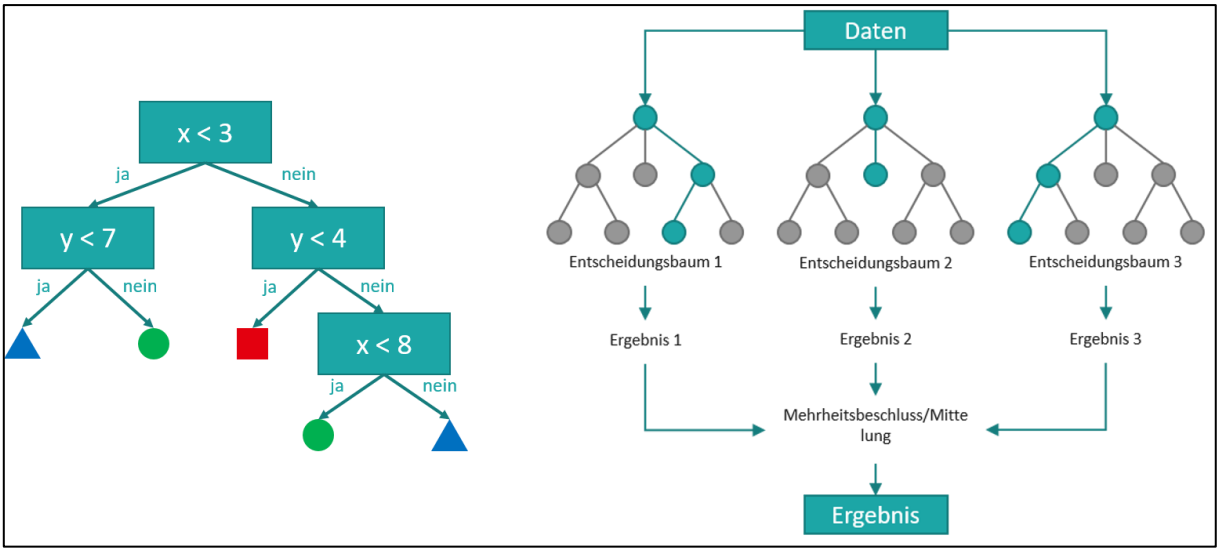


Abbildung 1: Beispielarchitektur eines Entscheidungsbaumes (links) und dessen Erweiterung Random Forest (rechts).

Im Bereich des maschinellen Lernens gibt es jedoch auch Verfahren, die Menschen direkt interpretieren können. Ein populäres Beispiel dafür sind Entscheidungsbäume (Kotsiantis 2013). Im Gegensatz zu neuronalen Netzen sind Entscheidungsbäume eine Verknüpfung von logischen Operationen. Jeder Knoten entspricht einer Regel, nach der Werte miteinander verglichen werden. Durch diese für Menschen verständliche symbolische Darstellung lässt sich zum einen der Pfad einer Entscheidungsfindung durch die Baumstruktur zurückverfolgen, zum anderen aber auch die gesamtheitliche Logik des Modells nachvollziehen. Aber auch KI-Modelle wie Entscheidungsbäume können die Black Box-Eigenschaft aufweisen, falls seine Struktur zu groß und somit das darin enthaltene Regelwerk zu komplex wird, um für Menschen nachvollziehbar zu sein. Dies wird am Beispiel des Random Forest-Algorithmus deutlich (siehe Abbildung 1: Beispielarchitektur eines Entscheidungsbaumes (links) und dessen Erweiterung Random Forest (rechts).Abbildung 1). Hier werden mehrere Entscheidungsbäume im Verbund Eingesetzt, um die KI-Ausgabe zu erzeugen. Um die Entscheidungsfindung dieser KI zu verstehen, muss daher nicht nur die Logik eines einzelnen Baumes nachvollzogen werden, sondern von vielen verschiedenen gleichzeitig. Bei ML-Verfahren gibt es jedoch keine harte Grenze, ab der ein Modell als Black Box erachtet wird. Stattdessen ist der Übergang zu dieser Eigenschaft unscharf (Bathaee 2017). Dies wird auch daran deutlich, dass die Interpretierbarkeit der Logik eines Entscheidungsbaums von der Fähigkeit des einzelnen Menschen abhängt, komplexe Regelwerke nachzuvollziehen.

1.3. Komplexität vs. Interpretierbarkeit

Ein zunehmendes Dilemma in der Welt des maschinellen Lernens liegt im Konflikt zwischen Komplexität und Interpretierbarkeit von Modellen. Moderne ML-basierte KI-Modelle haben aufgrund ihrer komplexen Strukturen und der Nutzung tiefer neuronaler Netze eine enorme Leistungsfähigkeit erlangt. Diese Modelle können komplexe Muster erkennen, präzise Vorhersagen treffen und komplexe Aufgaben wie Bilderkennung oder Textgenerierung bewältigen. Allerdings sind sie oft für Menschen schwer nachvollziehbar, da sie als "Black Box" agieren, in der die Entscheidungsfindung und die zugrundeliegende Logik schwer zu verstehen sind. Die Suche nach einem Gleichgewicht zwischen Komplexität und Interpretierbarkeit stellt eine fortlaufende Herausforderung dar. Komplexe Modelle bieten oft eine höhere Leistungsfähigkeit, sind aber schwieriger zu interpretieren. Einfachere Modelle hingegen sind leichter verständlich, können jedoch möglicherweise nicht die gleiche Leistung erbringen. In Abbildung 2 werden unterschiedliche Modelle in Bezug auf ihre Leistung und Interpretierbarkeit verglichen. Modelle wie Lineare Regression oder Entscheidungsbäume sind in kleiner Form leichter verständlich, können dann aber nur einfache Probleme lösen. Neuronale Netze liegen am anderen Ende des Spektrums und können sehr komplexe Probleme lösen, worunter jedoch ihre Interpretierbarkeit leidet (Molnar 2020).

Diese mangelnde Interpretierbarkeit kann zu ernsthaften Herausforderungen führen. In sicherheitskritischen Bereichen wie in der Medizin, im Finanzsektor oder im Rechtswesen ist es von entscheidender Bedeutung, dass Entscheidungen nachvollziehbar sind und erklärt werden können. Wenn ein Kreditantrag abgelehnt wird oder eine medizinische Diagnose gestellt wird, sollten die Gründe für diese Entscheidung klar sein, um Fehler auszuschließen und mögliche Vorurteile oder Diskriminierung zu vermeiden. Hier kommt erklärbare künstliche Intelligenz ins Spiel. Erklärbarkeit zielt darauf ab, das Verständnis für die Funktionsweise von KI-Modellen zu verbessern, die nicht direkt interpretierbar sind und die Entscheidungsfindung dadurch transparenter zu machen. Dadurch können Modelle so konstruiert werden, dass sie sowohl hohe Leistung als auch Nachvollziehbarkeit bieten. Allerdings ist es wichtig anzumerken, dass nicht alle Modelle in gleicher Weise verständlich sein müssen. In einigen Fällen kann es ausreichend sein, die Entscheidungsprozesse auf einer abstrakteren Ebene zu verstehen, während in

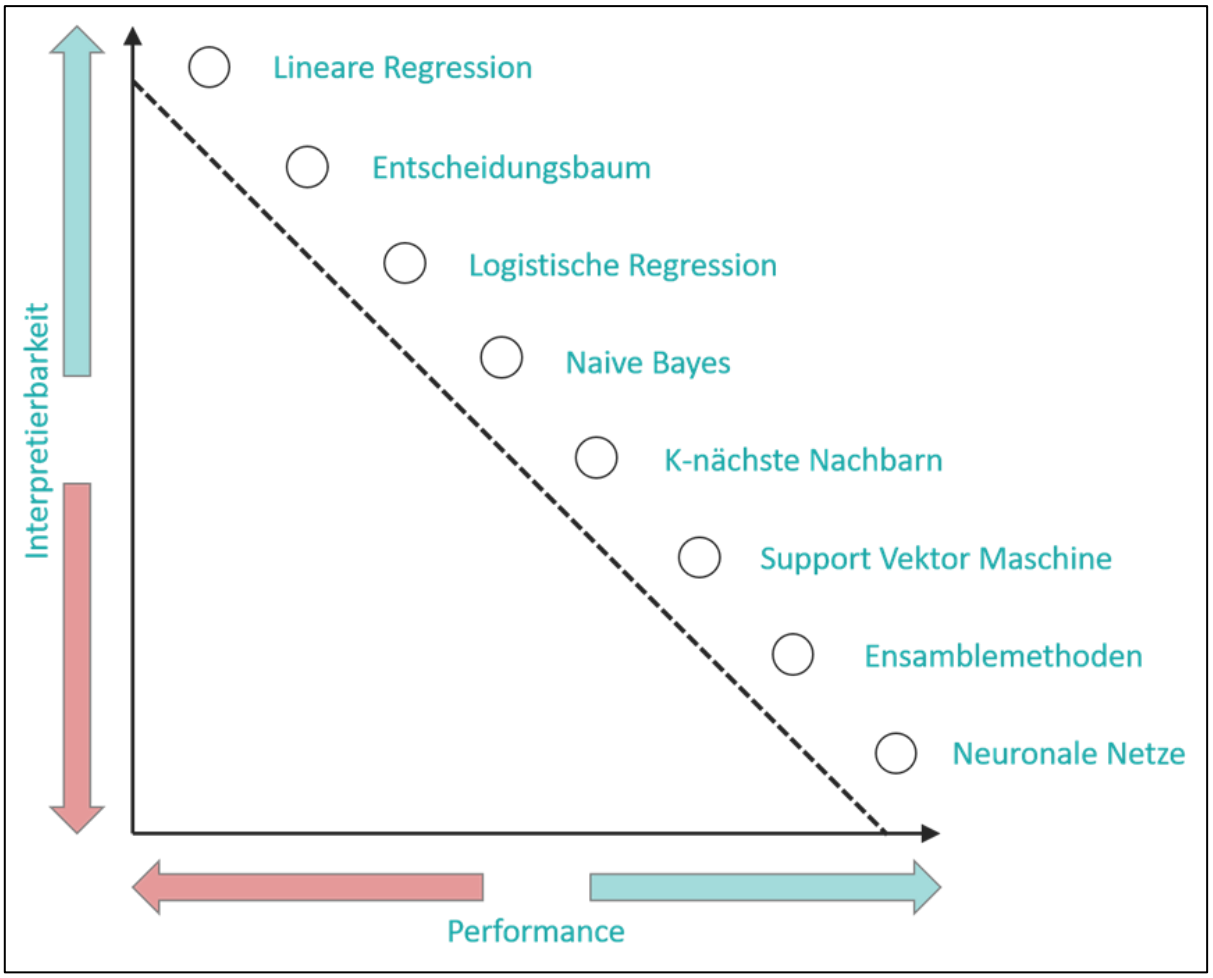


Abbildung 2: Vergleich unterschiedlicher KI-Modelle abhängig ihrer Leistungsfähigkeit und Interpretierbarkeit (Amazon AWS Whitepaper 2021).

anderen Fällen eine detaillierte Erklärung erforderlich ist. Die Anforderungen an die Interpretierbarkeit variieren je nach Anwendungsfall und den damit verbundenen Risiken (Burkart und Huber 2021).

Der Konflikt zwischen Komplexität und Interpretierbarkeit stellt bei ML-Modellen eine Herausforderung dar, die durch den Einsatz von XAI angegangen werden kann. Durch die Verbesserung der Interpretierbarkeit kann Vertrauen aufgebaut, Vorurteile aufgedeckt und Entscheidungen nachvollziehbar gemacht werden.

2. Motivation und Gründe für den Einsatz erklärbarer KI

Ein erfolgreicher Einsatz von erklärbarer KI kann einen erheblichen Zeit- und Kostenaufwand bedeuten. Dies gilt insbesondere, wenn die Erklärungen den Endnutzenden in einer geeigneten Benutzungsoberfläche bereitgestellt werden soll. Neben rechtlichen Verpflichtungen, die eine Erklärung der KI-Entscheidungen fordern, bietet der XAI-Einsatz aber den Anbietenden und Endnutzenden der KI-Systeme Mehrwerte, die den Implementierungsaufwand der XAI-Komponenten rechtfertigen können.

2.1. Nutzen des XAI-Einsatzes

Mit dem Einsatz von erklärbarer KI lässt sich nicht nur Nutzen beispielweise in Form einer höheren Akzeptanz und Zufriedenheit für die Endnutzenden generieren. Die Anbietenden der KI-Systeme können Nutzen gewinnen, wenn mit XAI die Modellevaluation und -entwicklung unterstützt wird oder wenn durch die Erklärungen etwas über das trainierte KI-Modell gelernt und somit neues Wissen über das von der KI zu lösende Problem gewonnen wird.

2.1.1. Modellevaluation

Erklärbare künstliche Intelligenz spielt eine wichtige Rolle in der Modellevaluation, da sie dazu dient, verschiedene Aspekte eines Modells zu bewerten und zu verbessern. XAI ermöglicht ein besseres Modellverständnis, die Erkennung und Behebung von Fehlern sowie die Betrachtung des Modells unter Aspekten der Sicherheit, Robustheit und Zuverlässigkeit. Durch das Offenlegen und Darstellen der internen Prozesse und der Entscheidungsfindung des Modells können Entwickler:innen Einblicke gewinnen, wie es zu bestimmten Vorhersagen kommt. Außerdem fördern Erklärungen das Vertrauen in ein Modell, da sie den Anwenderinnen und Anwendern ermöglichen die Ergebnisse nachzuvollziehen und zu überprüfen (Dwivedi et al. 2023).

Weiterhin spielt XAI eine wichtige Rolle bei der Fehlererkennung und -behebung. Durch die Überwachung und Analyse der Modellleistung können potenzielle Fehler und Unregelmäßigkeiten identifiziert werden. XAI-Techniken wie beispielsweise die Erkennung von Datenverzerrungen helfen, problematische Muster oder Vorurteile aufzudecken. Dies ermöglicht es, das Modell zu optimieren, um danach präzisere und zuverlässigere Vorhersagen zu erzielen. Auch in den Bereichen Sicherheit, Robustheit und Zuverlässigkeit von Modellen spielt XAI eine entscheidende Rolle. Durch Transparenz und Interpretierbarkeit des Modells können potenzielle Schwachstellen oder Angriffsvektoren erkannt und behoben werden. XAI ermöglicht die Überprüfung der Modellentscheidungen auf mögliche Verletzungen ethischer oder rechtlicher

Richtlinien, zum Beispiel ob ein Modell diskriminiert. Zudem können Abhängigkeiten von bestimmten Daten oder Merkmalen identifiziert und alternative Ansätze entwickelt werden, um die Robustheit des Modells zu verbessern (Burkart und Huber 2021).

Die Integration von XAI in die Modellevaluation bietet viele Vorteile. Sie fördert das Verständnis des Modells, ermöglicht die Fehlererkennung und -anpassung zur Verbesserung der Vorhersagegenauigkeit und trägt zur Gewährleistung von Sicherheit, Robustheit und Zuverlässigkeit bei.

2.1.2. Vertrauen und Akzeptanz

Indem man dem Menschen das Verhalten eines KI-Systems erklärt, kann Vertrauen gegenüber dieser neuen Technologie geschaffen werden (Gerlings et al. 2021). Das KI-Verhalten zu verstehen ermöglicht es, potenziellen Kund:innen die Sinnhaftigkeit der Entscheidungsfindung sowie die Zuverlässigkeit und Robustheit des KI-Systems zu beweisen. Ein Unternehmen, das solche Eigenschaften eines KI-Systems beurteilen kann statt in eine Black Box vertrauen zu müssen, ist gewillter dieses System in Betrieb zu nehmen. Doch auch bei Endnutzenden führt ein besseres Verständnis des KI-Verhaltens und die bessere Kontrolle, die sie durch die Erklärungen über das System erhalten, zu einem Vertrauensgewinn und dadurch zu einer Steigerung der Akzeptanz gegenüber der neuen Technologie.

2.1.3. Wissensextraktion

Algorithmen des maschinellen Lernens, insbesondere künstliche neuronale Netze, sind in der Lage, versteckte Strukturen und Zusammenhänge in großen Datenmengen zu erkennen. Dieses „Wissen“ der trainierten KI-Modelle kann durch XAI-Methoden extrahiert und beispielsweise als Regelsystem dargestellt werden (Gerlings et al. 2021). Gerade im wissenschaftlichen Kontext lassen sich dadurch latente Zusammenhänge für den Menschen offenlegen und somit neue Erkenntnisse über eine Problemstellung gewinnen.

2.2. Rechtliche Vorgaben auf Ebene der Europäischen Union (EU)

Auf EU-Ebene existieren verschiedene Verordnungen bzw. Verordnungsentwürfe, die Aspekte von Erklärbarkeit adressieren. Während die Datenschutz-Grundverordnung (DSGVO) die Regelung der Verarbeitung personenbezogener Daten adressiert, befasst sich die KI-Verordnung mit der Regelung des Umgangs mit Systemen der künstlichen Intelligenz. Auf beide wird im Folgenden kurz eingegangen. Tiefergehende Erläuterungen zum Thema liefert der Technische Report „Datenschutz bei künstlicher Intelligenz“ (Bao et al. 2023).

2.2.1. Datenschutz-Grundverordnung (DSGVO)

Die DSGVO (Europäisches Parlament und Rat der Europäischen Union 2016) ist eine Verordnung der EU. Sie ist gültig im Europäischen Wirtschaftsraum (EWR), dem die Europäische Union sowie Norwegen, Island und Liechtenstein angehören. Sie regelt die Verarbeitung personenbezogener Daten[1]. Dadurch stärkt sie einerseits die Bürger:innen in ihren Rechten auf Selbstbestimmung und Kontrolle über ihre personenbezogenen Daten und regelt andererseits den freien Datenverkehr.

Dabei gelten für die Erhebung personenbezogener Daten verschiedene Grundsätze ((Europäisches Parlament und Rat der Europäischen Union 2016), Art. 5):

- Rechtmäßigkeit und Transparenz: Verarbeitung in einer für die betroffene Person nachvollziehbaren Weise (Art. 5 Abs. 1a)
- Zweckbindung: Erhebung nur für festgelegte, eindeutige und legitime Zwecke (Art. 5 Abs. 1b)
- Datenminimierung: Erhebung dem Zweck angemessen und auf das notwendige Maß beschränkt (Art. 5 Abs. 1c)
- Richtigkeit: Daten sind sachlich richtig und auf dem neuesten Stand (Art. 5 Abs. 1d)
- Speicherbegrenzung: Speicherung in einer Form, die die Identifizierung der betroffenen Person nur so lange ermöglicht wie für die Verarbeitungszwecke erforderlich (Art. 5 Abs. 5e)
- Integrität und Vertraulichkeit: Verarbeitung, die angemessene Sicherheit der personenbezogenen Daten durch geeignete technische und organisatorische Maßnahmen gewährleistet (Art. 5 Abs. 1f)

Um dies zu gewährleisten, sollen technische und organisatorische Maßnahmen getroffen werden wie „Grundsätze des Datenschutzes durch Technik (data protection by design) und […]

[1] Begriffsbestimmungen in der DSGVO, Artikel 4:

- Personenbezogene Daten: alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person“) beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann.
- Verarbeitung: Jeder mit oder ohne Hilfe automatisierter Verfahren ausgeführte Vorgang oder jede solche Vorgangsreihe im Zusammenhang mit personenbezogenen Daten wie das Erheben, das Erfassen, die Organisation, das Ordnen, die Speicherung, die Anpassung oder Veränderung, das Auslesen, das Abfragen, die Verwendung, die Offenlegung durch Übermittlung, Verbreitung oder eine andere Form der Bereitstellung, den Abgleich oder die Verknüpfung, die Einschränkung, das Löschen oder die Vernichtung.

datenschutzfreundliche Voreinstellungen (data protection by default)“ ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 78). Dazu gehören insbesondere die Grundsätze der Datenminimierung und der Transparenz, aber auch, dass personenbezogene Daten so schnell wie möglich pseudonymisiert[2] werden. Neue Technologien und Dienste sollen so gestaltet werden, dass das Systemdesign von Beginn an den Datenschutz berücksichtigt, so dass die Bürger sich nicht darum kümmern müssen. Artikel 25 adressiert den „Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen“, Artikel 32 die „Sicherheit der Verarbeitung“, zu der u.a. die Pseudonymisierung und Verschlüsselung personenbezogener Daten sowie Fähigkeiten bzgl. der Vertraulichkeit, Integrität, Verfügbarkeit und Belastbarkeit der verarbeitenden Systeme gehören.

Transparenz wird in der DSGVO an verschiedenen Stellen adressiert. Transparenz für natürliche Personen beinhaltet dabei, dass Klarheit darüber herrscht, wie personenbezogene Daten „erhoben, verwendet, eingesehen oder anderweitig verarbeitet werden […]“, aktuell und zukünftig ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 39). Voraussetzung dafür ist, dass Information bezüglich dieser Datenverarbeitung „präzise, leicht zugänglich und verständlich sowie in klarer und einfacher Sprache abgefasst“ sind ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 58). Besondere Sorgfalt wird gefordert, wenn die involvierten technischen Systeme komplex sind oder wenn Kinder betroffen sind. Transparenz in Bezug auf die Funktionen und die Verarbeitung personenbezogener Daten sollen von den betroffenen Personen einfach überwacht werden können ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 78). Es wird zudem angeregt, Zertifizierungsverfahren, Datenschutzsiegel und Datenschutzprüfzeichen zu etablieren, die den betroffenen Personen erlauben, das Datenschutzniveau rasch einschätzen zu können ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 100). Artikel 12 greift diese Aspekte von Transparenz bzw., durch welche Maßnahmen sie gewährleistet werden sollten, auf und formuliert detaillierte Forderungen. Artikel 88 adressiert die Datenverarbeitung im Beschäftigungskontext und nennt Transparenz der Verarbeitung personenbezogener Daten als berechtigtes Interesse und Grundrecht betroffener Personen ((Europäisches Parlament und Rat der Europäischen Union 2016), Art. 88 Abs. 2).

[2] Begriffsbestimmungen in der DSGVO, Artikel 4:

- Pseudonymisierung: Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden.

Systeme Künstlicher Intelligenz fallen in der die DSGVO unter den Begriff der „Automatisierten Entscheidungsfindung“. Diese soll nur auf der Grundlage besonderer Kategorien von personenbezogenen Daten und nur unter bestimmten Bedingungen erlaubt sein ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 71). Artikel 22 regelt die Details. Art. 22 Abs. 1 garantiert der betroffenen Person das Recht, „nicht einer ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt“. Art. 22 Abs. 2 formuliert Fälle, in denen Art. 22 Abs. 1 nicht gilt, bspw. wenn die Entscheidung „mit ausdrücklicher Einwilligung der betroffenen Person erfolgt“. Weiter wird verlangt, die „berechtigten Interessen der betroffenen Person zu wahren, wozu mindestens das Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen, auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung gehört“ ((Europäisches Parlament und Rat der Europäischen Union 2016), Art. 22 Abs. 3).

Die Artikel 13 bis 15 adressieren die Informationspflicht, die der oder die Verantwortliche gegenüber der betroffenen Person hat, wenn er oder sie deren personenbezogene Daten erhebt. Dies umfasst u.a. Informationen wie die Kontaktdaten des oder der Verantwortlichen, die Datenverarbeitungs-Zwecke und ggf. Angaben zu Datenschutzbeauftragten oder Empfänger:innen von personenbezogenen Daten ((Europäisches Parlament und Rat der Europäischen Union 2016), Art. 13 Abs.1). Um eine faire und transparente Verarbeitung zu gewährleisten, sollen zusätzlich Informationen bzgl. der Dauer der Datenspeicherung sowie bzgl. der Rechte der betroffenen Person (u.a. Auskunftsrecht, Recht, die Einwilligung zu widerrufen) gegeben werden; insbesondere ist die betroffene Person darüber zu informieren, ob automatisierte Entscheidungsfindung besteht ((Europäisches Parlament und Rat der Europäischen Union 2016), Art. 13 Abs.2). Sollen personenbezogene Daten für einen anderen Zweck weiterverarbeitet werden, hat der oder die Verantwortliche die Pflicht, der betroffenen Person vor dieser Weiterverarbeitung alle maßgeblichen Informationen zur Verfügung zu stellen ((Europäisches Parlament und Rat der Europäischen Union 2016), Art. 13 Abs.3). Artikel 14 legt die Regeln zur Informationspflicht fest, wenn die personenbezogenen Daten nicht bei der betroffenen Person erhoben wurden; Artikel 15 beinhaltet das Auskunftsrecht der betroffenen Person.

Ein weiterer wichtiger Aspekt ist die Einwilligung[3] in die Datenverarbeitung. Diese „sollte durch eine eindeutige bestätigende Handlung erfolgen“ ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 32) und zudem freiwillig, informiert und unmissverständlich sowie für genau einen Zweck. Bei mehreren Verarbeitungszwecken sollte für jeden

[3] Begriffsbestimmungen in der DSGVO, Artikel 4:

- Einwilligung (der betroffenen Person): Jede freiwillig für den bestimmten Fall, in informierter Weise und unmissverständlich abgegebene Willensbekundung in Form einer Erklärung oder einer sonstigen eindeutigen bestätigenden Handlung, mit der die betroffene Person zu verstehen gibt, dass sie mit der Verarbeitung der sie betreffenden personenbezogenen Daten einverstanden ist.

eine separate Einwilligung gegeben werden. Unter Gründe Abs. 42 wird formuliert, es „sollte eine vom Verantwortlichen vorformulierte Einwilligungserklärung in verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zur Verfügung gestellt werden […]“. Es muss zudem gewährleistet sein, dass bei Widerruf einer Einwilligung die entsprechenden Daten gelöscht und nicht mehr verarbeitet werden ((Europäisches Parlament und Rat der Europäischen Union 2016), Gründe Abs. 65).

Das Bundesministerium der Justiz verlinkt unter der Überschrift „Wesentliche Rechte für Verbraucherinnen und Verbraucher“ (Bundesministerium der Justiz 2023) anschauliche Erklärungen, Anleitungen und Musterschreiben und erläutert darüber hinaus die wesentlichen Rechte auf der Website (Recht auf Information und Auskunft; Recht auf Benachrichtigung und Löschung; Einwilligung; Recht auf Widerruf; Recht auf Widerspruch; Rechte bei automatisierter Entscheidung im Einzelfall; Recht auf Datenübertragbarkeit; Datenschutzbehörden und Verbraucherzentralen).

2.2.2. KI-Verordnung

Während die DSGVO bereits den Status einer EU-Verordnung hat, die seit dem 25. Mai 2018 in allen EU-Mitgliedsstaaten gilt, hat die KI-Verordnung (informelle Bezeichnung; engl. AI Act) bislang den Status eines Vorschlags für eine Verordnung. Im Folgenden wird daher der Begriff KI-Verordnungsentwurf verwendet. Der Entwurf ist öffentlich einsehbar (EU-Kommission 2021b; EU-Kommission 2021a).

Der KI-Verordnungsentwurf will einen Rechtsrahmen für eine sichere, vertrauenswürdige und ethisch vertretbare KI schaffen. Er soll einerseits die Nutzung von KI fördern und Unternehmen zur Entwicklung von KI-Systemen anregen. Andererseits sollen „die mit bestimmten Anwendungen dieser Technologie verbundenen Risiken eingedämmt werden können“ ((EU-Kommission 2021b), S.1). Auf diese Weise soll sichergestellt werden, dass KI-Systeme im Einklang mit den Werten, Grundrechten und Prinzipien der Europäischen Union stehen.[4]

[4] Der KI-Verordnungsentwurf berücksichtigt den Schutz von Ethikgrundsätzen gemäß der Entschließung 2020/2012(INL) (Europäisches Parlament 2020). Er adressiert zudem die Vereinbarkeit von KI-Systemen mit den Grundrechten der EU-Grundrechtecharta (Rat der Europäischen Union 2020): Grundrechte wie die Würde des Menschen (Artikel 1), die Achtung des Privatlebens und der Schutz personenbezogener Daten (Artikel 7 und 8), die Nichtdiskriminierung (Artikel 21) und die Gleichheit von Frauen und Männern (Artikel 23) sollen durch die Anforderungen an vertrauenswürdige KI noch stärker geschützt werden. Genannt werden als typische, zu lösende Probleme die „Undurchsichtigkeit, Komplexität, der sogenannte „Bias“, ein gewisses Maß an Unberechenbarkeit und teilweise autonomes Verhalten einiger KI-Systeme“ ((EU-Kommission 2021b), vgl. S.2; s.a. (EU-Kommission 2019) für Ethikleitlinien bzw. (EU-Kommission 2020) für eine Bewertungsliste für vertrauenswürdige KI). Weitere Entschließungen zur KI des Europäischen Parlaments betreffen die zivilrechtliche Haftung, das Urheberrecht, das Strafrecht sowie Bildung, Kultur und den audiovisuellen Bereich ((EU-Kommission 2021b), vgl. S.2-3).

Als beste Option, diese Ziele zu erreichen, wird ein detaillierter Rechtsrahmen ausschließlich für Hochrisiko-KI-Systeme angesehen. Hochrisikosysteme sind solche, die „erhebliche Risiken für die Gesundheit und Sicherheit oder die Grundrechte von Personen bergen“ ((EU-Kommission 2021b), S.4). Ergänzt werden soll dieser Rechtsrahmen durch einen Verhaltenskodex für KI-Systeme, die kein hohes Risiko darstellen ((EU-Kommission 2021b), S.11).

Für Hochrisiko-KI-Systeme werden Anforderungen gestellt „an hohe Datenqualität, Dokumentation und Rückverfolgbarkeit, Transparenz, menschliche Aufsicht, Präzision und Robustheit“. KI-Systeme ohne hohes Risiko müssen hingegen nur weniger Transparenzpflichten erfüllen, bspw. derart, dass „bei der Interaktion mit Menschen der Einsatz von KI-Systemen angezeigt werden muss“ ((EU-Kommission 2021b), S.8). Darüber hinaus identifiziert der KI-Verordnungsentwurf auch KI-Systeme, die aufgrund der Gefahren für betroffene Personen schlicht verboten sind. Diese Unterscheidung von Anwendungen von KI, die unannehmbares Risiko, hohes Risiko oder geringes/minimales Risiko[5] darstellen, wird auch in den Festlegungen zum Gegenstand des KI-Verordnungsentwurfs berücksichtigt (Art. 1, vgl. (EU-Kommission 2021b) S.44f):

(a) Vorschriften für das Inverkehrbringen, die Inbetriebnahme und die Verwendung von KI-Systemen in der EU;
(b) Verbote bestimmter Praktiken im Bereich der KI (Art. 5, vgl. (EU-Kommission 2021b), S.50-52). Darunter fallen bspw. KI-Systeme, die Personen unterschwellig beeinflussen wollen, um deren Verhalten in einer Weise zu beeinflussen, dass sie sich selbst oder anderen Personen physischen oder psychischen Schaden zufügt, oder auch KI-Systeme, die für Behörden die Vertrauenswürdigkeit von Personen auf Grundlage ihres sozialen Verhaltens klassifizieren, mit der Konsequenz, dass die Person bspw. ungerechtfertigt benachteiligt wird („Social Scoring“).
(c) besondere Anforderungen an Hochrisiko-KI-Systeme und Verpflichtungen für deren Betreiber (s.u.)
(d) Transparenzvorschriften für KI-Systeme, die mit natürlichen Personen interagieren sollen, KI-Systeme zur Emotionserkennung und biometrischen Kategorisierung und KI-Systeme, die zum Erzeugen oder Manipulieren von Bild-, Ton- oder Videoinhalten verwendet werden (s.u.)
(e) Vorschriften für die Marktbeobachtung und -überwachung.

Die Einstufung eines KI-Systems als Hochrisiko-KI-System hängt von seiner Funktion ab, von seinem Zweck und von den Anwendungsmodalitäten. Berücksichtigt werden dabei auch bestehende EU-Produktsicherheitsvorschriften ((EU-Kommission 2021b), S.15).

In die Kategorie Hochrisiko-KI-Systeme fallen alle KI-Systeme, die ein Produkt oder eine Sicherheitskomponente eines Produkts aus folgenden Bereichen sind (Art. 6, vgl. (EU-Kommission 2021b) S.52): Maschinen, Spielzeuge, Sportboote und Wassermotorräder, Aufzüge, Geräte in

[5] Erläuterung der Risikostufen s.a. https://digital-strategy.ec.europa.eu/de/policies/regulatory-framework-ai

explosionsgefährdeten Bereichen, Funkanalgen, Druckgeräten, Seilbahnen, persönliche Schutzausrüstung, Geräte zur Verbrennung gasförmiger Brennstoffe, Medizinprodukte und In-vitro-Diagnostika, des Weiteren Systeme für die Sicherheit der Zivilluftfahrt, zwei-, drei- und vierrädrige Fahrzeuge, land- und forstwirtschaftliche Fahrzeuge, Schiffsausrüstung und das Eisenbahnsystem ((EU-Kommission 2021a), Anhang II).

Zusätzlich gelten als Hochrisiko-KI-Systeme ((EU-Kommission 2021a), Anhang III) solche

1. zur Biometrischen Identifizierung und Kategorisierung natürlicher Personen (Nutzung in Echtzeit oder nachträglich);
2. zur Verwaltung und Betrieb kritischer Infrastrukturen (Straßenverkehr, Wasser-, Gas-, Wärme-, Stromversorgung);
3. zur allgemeinen und beruflichen Bildung (bspw. Entscheidungen bzgl. Zugang zu Einrichtungen);
4. zur Beschäftigung, Personalmanagement und Zugang zur Selbstständigkeit (bspw. Einstellung, Auswahl, Beförderung, Kündigung, Leistungsbewertung);
5. zur Zugänglichkeit und Inanspruchnahme grundlegender privater und öffentlicher Dienste und Leistungen (bspw. um festzustellen, ob einer Person öffentliche Unterstützungsleistungen zustehen, Kreditwürdigkeitsprüfung, Priorisierung von Not- und Rettungsdiensten);
6. zur Strafverfolgung (bspw. Risikoabschätzung, ob eine Person eine Straftat begehen wird, Lügendetektor);
7. im Zusammenhang mit Migration, Asyl, Grenzkontrolle (bspw. Bewertung, ob eine Person ein Gesundheitsrisiko ist, Echtheit von Nachweisunterlagen);
8. zur Rechtspflege und für demokratische Prozesse;

Hochrisiko-KI-Systeme müssen verschiedene Anforderungen erfüllen. Im Einzelnen sind dies:

- ein Risikomanagementsystem, das als kontinuierlicher iterativer Prozess den gesamten Lebenszyklus eines KI-Systems begleitet und sicherstellen soll, dass es stets bestimmungsgemäß funktioniert (Art. 9, vgl. (EU-Kommission 2021b) S.54f);
- Anforderungen bzgl. Daten und Daten-Governance: Wenn Techniken eingesetzt werden, bei denen Modelle mit Daten trainiert werden, müssen Trainings-, Validierungs- und Testdatensätze entwickelt werden, die bestimmten (hohen) Qualitätskriterien entsprechen, bspw. es müssen die Datensätze relevant, repräsentativ, fehlerfrei und vollständig sein (Art. 10, vgl. (EU-Kommission 2021b) S.55f)
- Technische Dokumentation (Art. 11, vgl. (EU-Kommission 2021b) S.56f);
- Aufzeichnungspflichten, bspw. die Protokollierung des Betriebs des KI-Systems (Art. 12, vgl. (EU-Kommission 2021b) S.57);
- Transparenz und Bereitstellung von Informationen für die Nutzenden (Art. 13, vgl. (EU-Kommission 2021b) S.57f):
    - Der Betrieb des KI-Systems muss hinreichend transparent sein, „damit die Nutzer die Ergebnisse des Systems angemessen interpretieren und verwenden können“

    - Gebrauchsanweisungen müssen bereitgestellt werden und „präzise, vollständige, korrekte und eindeutige Informationen in einer für die Nutzer relevanten, barrierefrei zugänglichen und verständlichen Form enthalten“, u.a. bzgl. Merkmalen, Fähigkeiten, Leistungsgrenzen und Zweckbestimmung des KI-Systems, Spezifikationen für die Eingabedaten, Wartungs- und Pflegemaßnahmen oder auch bzgl. technischer Maßnahmen, „die getroffen wurden, um den Nutzern die Interpretation der Ergebnisse von KI-Systemen zu erleichtern“;
- Menschliche Aufsicht (Art. 14, vgl. (EU-Kommission 2021b) S.58f): „Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie während der Dauer der Verwendung des KI-Systems – auch mit geeigneten Werkzeugen einer Mensch-Maschine-Schnittstelle – von natürlichen Personen wirksam beaufsichtigt werden können“. Bspw. müssen Anzeichen von Anomalien oder Fehlfunktionen so bald wie möglich erkannt und behoben werden können und die menschliche Aufsicht muss in den Systembetrieb eingreifen oder ihn mit einer „Stopptaste“ oder einem ähnlichen Verfahren unterbrechen können;
- Genauigkeit, Robustheit, Cybersicherheit (Art. 15, vgl. (EU-Kommission 2021b) S.59f).

Der KI-Verordnungsentwurf adressiert in Kapitel 3 (Art. 16 bis 29, vgl. (EU-Kommission 2021b) S.60-67) ausführlich die Pflichten der Anbieter:innen und Nutzenden von Hochrisiko-KI-Systemen. Für Anbieter:innen gehört dazu bspw., die zuletzt genannt Anforderungen zu erfüllen sowie verschiedene Prüf- und Kontrollmechanismen durchzuführen (Art. 16 bis 25). Pflichten für Einführer:innen, Händler:innen und Nutzende des KI-Systems sind gesondert beschrieben (Art. 26 bis 29).

In einer Beispielrechnung werden die zu erwartenden Rechtsbefolgungskosten für ein Hochrisiko-KI-System folgendermaßen beziffert ((EU-Kommission 2021b), S.11f). Bei einem System im Wert von etwa 170 000 EUR fallen geschätzt Kosten von 6000 EUR bis 7000 EUR dafür an, die Anforderungen und Verpflichtungen einzuhalten. Hinzu kommen jährliche Kosten von 5000 EUR bis 8000 EUR für die menschliche Aufsicht an sowie Überprüfungskosten zwischen 3000 EUR bis 7500 EUR. Als wichtig wird erachtet, dass diese sogenannten Rechtsbefolgungskosten nicht zu einer Verhinderung oder Verzögerung der Nutzbarkeit von KI-Systemen führen. Denn es ist erklärtes Ziel, die europäische Wettbewerbsfähigkeit und Industriebasis im Bereich der KI zu stärken, um die digitale Autonomie zu vergrößern ((EU-Kommission 2021b), S.11).

Um die Einhaltung der Vorgaben bewerten und überprüfen zu können, soll eine öffentliche, unionsweite Datenbank für Hochrisiko-KI-Systeme realisiert werden. KI-Anbieter:innen sollen zudem verpflichtet werden, schwerwiegende Vorfälle oder Fehlfunktionen an ihre nationalen Behörden zu melden, die die Vorfälle untersuchen und an die Europäische Kommission weiterleiten ((EU-Kommission 2021b), S.14).

Neben den Anforderungen und Pflichten für Hochrisiko-KI-Systeme formuliert der KI-Verordnungsentwurf zusätzlich Transparenzpflichten für bestimmte KI-Systeme (Art. 52, vgl. (EU-Kommission 2021b) S.78f):

- KI-Systeme für die Interaktion mit natürlichen Personen: Die Anbieter:innen müssen hier sicherstellen, dass die KI-Systeme „so konzipiert und entwickelt werden, dass natürlichen Personen mitgeteilt wird, dass sie es mit einem KI-System zu tun haben, es sei denn, dies ist aufgrund der Umstände und des Kontexts der Nutzung offensichtlich“ ((EU-Kommission 2021b), S.78);
- KI-Systeme zur Emotionserkennung oder zur biometrischen Kategorisierung: Verwender:innen solcher Systeme müssen die davon betroffenen natürlichen Personen über den Betrieb des Systems informieren;
- KI-Systeme, die Bild-, Ton- oder Videoinhalte erzeugen oder manipulieren: Nutzende solcher KI-Systeme müssen offenlegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden, wenn die erzeugten/manipulierten Inhalte „[…] wirklichen Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähneln und einer Person fälschlicherweise als echt oder wahrhaftig erscheinen würden („Deepfake“)“ ((EU-Kommission 2021b), S.78).

Diese Vorgaben stellen sicher, dass die betroffenen Personen bewusste Entscheidungen bzgl. der Nutzung solcher KI-System treffen bzw. bestimmte Situationen vermeiden können.

KI-Systeme, die nicht in die Kategorie der Hochrisiko-KI-Systeme fallen, sondern für die EU-Bürger ein eher geringes oder minimales Risiko darstellen, werden in Artikel 69 adressiert ((EU-Kommission 2021b), S.91). Zum einen soll gefördert und erleichtert werden, Verhaltenskodizes aufzustellen in der Form, dass mittels technischer Spezifikationen und Lösungen die Anforderungen, die für die Hochrisiko-KI-Systeme formuliert wurden, ebenfalls gewährleistet sind. Zum anderen wird gefördert und erleichtert, Verhaltenskodizes aufzustellen, die freiwillig darüber hinaus gehende Anforderungen erfüllen wie ökologische Nachhaltigkeit, barrierefreie Zugänglichkeit, Beteiligung von Interessenträgern bei der Systemkonzeption und -entwicklung sowie Vielfalt der Entwicklungsteams.

3. Erklärbare Künstliche Intelligenz

Der erfolgreiche Einsatz von erklärbarer KI hängt nicht nur davon ab, was man erklärt, sondern auch davon, wie man die Informationen einem Menschen präsentiert. Für KI-Entwickler:innen sind die meist recht technischen Ausgaben gängiger XAI-Methoden auch ohne weitere Aufbereitung nutzbar und verständlich. Für Endnutzende kann es hingegen notwendig sein, diese Ausgaben zunächst in eine einfacher verständliche Form zu überführen. Eine Erklärung muss möglicherweise auch korrekt interpretiert werden können, um die richtigen Schlussfolgerungen daraus ziehen zu, was entsprechendes Expertenwissen voraussetzt. Das Bereitstellen der Erklärungen und Vermitteln der notwendigen Informationen ist daher keine triviale Aufgabe. Insbesondere ist die Gewährleistung der Verständlichkeit einer Erklärung vom konkreten Anwendungskontext wie auch von den Fähigkeiten der Nutzenden abhängig.

3.1. Menschliche Erklärungen

Wie genau Menschen ihre Entscheidungen erklären und was eine verständliche Erklärung ausmacht, beschäftigt die Wissenschaft schon seit Jahrzehnten. Auch wenn man den menschlichen Erklärungsprozess noch nicht zur Gänze versteht, konnten Eigenschaften identifiziert werden, wie Menschen Erklärungen erfahren (Gerlings et al. 2021; Miller 2017).

Erklärungen sind:

1. vergleichend und gegenüberstellend,
2. voreingenommen und unvollständig,
3. kausal und nicht wahrscheinlichkeitsbasiert,
4. sozial und interaktiv.

Aus diesen Eigenschaften lassen sich Anforderungen an die Gestaltung einer verständlichen Benutzeroberfläche für Erklärungen (explanation user interface, XUI) ableiten: Der Umfang und die Komplexität einer Erklärung sollen so gering wie möglich gehalten werden. Den Benutzenden soll es jedoch möglich sein, bei Bedarf zusätzliche Informationen zu erhalten. Dadurch kann Interaktivität mit dem System geschaffen werden. Erklärungen sollen durch Vergleiche und Gegenüberstellungen kausale Zusammenhänge aufzeigen und somit Schlussfolgerungen über das KI-Verhalten ermöglichen.

Diese Anforderungen können eine Hilfestellung bei der Entwicklung der initialen XUI darstellen. Einen etablierten Gestaltungsleitfaden für eine verständliche XUI gibt es jedoch noch nicht. Dies liegt auch daran, dass sich die technischen Wissenschaftsdisziplinen auf die Entwicklung neuer XAI-Methoden konzentrieren, dabei jedoch deren Evaluation am Menschen vernachlässigen (Chromik und Butz 2021). Dadurch gibt es noch zu wenig Evidenz, welche Me-

thoden und welche Darstellungsformen in einem konkreten Anwendungsszenario auch zu einer für den Menschen verständlichen Erklärung des jeweiligen KI-Verhaltens führen (Nauta et al. 2023).

3.2. XAI-Zielgruppen

Die Intention und Darstellung einer Erklärung über das Verhalten eines KI-Modells hängt stark von der genauen Zielgruppe und deren Bedarfen ab. Wieso benötigt eine Person eine Erklärung und welche Fragen sollen dadurch beantwortet werden können? Welches Vorwissen besitzt eine Person, welche Informationen werden benötigt und wie müssen sie dargestellt werden, damit die Person die Erklärung verstehen kann? Bei der Entwicklung und Gestaltung eines sich selbst erklärenden Systems ist daher die Betrachtung der XAI-Zielgruppe notwendig und kritisch für eine erfolgreiche Umsetzung. Allerdings gibt es bisher noch keine einheitliche Definition, welche XAI-Zielgruppen es gibt und durch welche Eigenschaften sich diese unterscheiden.

Davis et al. stellt eine Kategorisierung anhand der Beziehung zum KI-System vor (Davis et al. 2020). Entwickler:innen (developer) trainieren das KI-Modell, Endnutzende (end user) nutzen das KI-System und interagieren damit. Die dritte Kategorie stellen Personen dar, die von der KI-Entscheidung betroffen sind (imposed user).

Die Einteilung der XAI-Zielgruppen nach Mohseni et al. folgt einem auf dem Vorwissen einer Person basierenden Ansatz (Mohseni et al. 2021): KI-Expert:in (AI expert), Datenexpert:in (data expert) und KI-Neuling (AI novice). KI-Expert:innen sind hier die Entwickler:innen des KI-Modells, Datenexpert:innen beinhalten Datenanalyse-Expert:innen sowie Domänenexpert:innen, während KI-Neuling allgemein jede Person einbezieht, die als Endnutzende das KI-System verwendet.

Gunning et al. fassen alle Zielgruppen als Endnutzende zusammen und stellen bei ihrer Unterscheidung vor allem den Zeitpunkt der KI-Entwicklung heraus, zu dem eine Erklärung benötigt wird (Gunning et al. 2019). Beispiele dieser Unterteilung sind: Entwickler:innen, Systemtester:innen, Anwendende, Richter:innen oder politische Entscheidungsträger:innen (law makers), die gegebenenfalls die Fairness eines KI-Systems beurteilen können müssen.

Die hier vorgestellte Kategorisierung der XAI-Zielgruppen vereint die drei bereits existierenden Einteilungen. Die Zielgruppen und Beispiele für die typischen Domänen Versicherung, Finanzen und Gesundheit sind in Tabelle 1: Beispiele der drei XAI-Zielgruppen Lai:in, Domänenexpert:in, KI-Expert:in für die typischen Anwendungsdomänen Versicherung, Finanzen und Gesundheit sowie typische Bedarfe der XAI-Zielgruppen. dargestellt.

Tabelle 1: Beispiele der drei XAI-Zielgruppen Lai:in, Domänenexpert:in, KI-Expert:in für die typischen Anwendungsdomänen Versicherung, Finanzen und Gesundheit sowie typische Bedarfe der XAI-Zielgruppen.

	Lai:in	Domänenexpert:in	KI-Expert:in
Versicherung	Versicherte Person	Versicherungsvertreter:in	Modell-/ Softwareentwickler:in, Data Scientist
Finanzen	Kreditnehmer:in	Kreditsachbearbeiter:in	
Gesundheit	Patient:in	Ärzt:in	
Bedarfe	Entscheidungen über sich selbst verstehen	Sinnhaftigkeit und Konsistenz eines Modells prüfen	Modell explorieren, evaluieren, anpassen

Bei Lai:innen werden weder Vorkenntnisse im Bereich KI noch Vorwissen in der betrachteten Domäne vorausgesetzt. Die Person kann dabei selbst Endnutzende der KI-Anwendung sein oder nur von der KI-Entscheidung betroffen sein. Domänenexpert:innen besitzen spezielles Wissen in einem bestimmten Bereich. Dies kann beispielsweise ein:e Kreditsachbearbeiter:in, aber auch ein:e Richter:in oder politische:r Entscheidungsträger:in sein. In diese Gruppe gehören daher nicht nur Expert:innen aus den KI-Anwendungsdomänen, sondern auch aus übergreifenden Gebieten wie zum Beispiel Ethik und Recht. KI-Entwickler:innen sind Software-Entwickler:innen und Data Scientists, die die KI-Anwendung entwickeln und dazu entsprechende Kenntnisse im Bereich KI und insbesondere ML benötigen. Domänenwissen wird hier hingegen nicht angenommen.

Neben dem vorhandenen Vorwissen unterscheiden sich die drei XAI-Zielgruppen auch in ihren Zielen, die durch Erklärungen erfüllt werden sollen. KI-Entwickler:innen müssen möglichst vollständig das Verhalten von KI-Modellen verstehen können, um es zu evaluieren und in nachfolgenden Iterationen anpassen und verbessern zu können. Domänenexpert:innen beurteilen das KI-Verhalten hinsichtlich bestimmter domänenspezifischer Aspekte zu denen nur sie entsprechendes Vorwissen besitzen. Dadurch soll vor der Inbetriebnahme sichergestellt werden, dass die KI-Anwendung den erforderlichen Qualitätsansprüchen genügt und mit bestehendem Recht konform ist. Aber auch während des Betriebs sollen Domänenexpert:innen als Endnutzende in Unternehmen die Sinnhaftigkeit und Konsistenz des KI-Verhaltens prüfen, um rechtswidrige oder schädliche KI-Entscheidungen zu verhindern. Die Zielgruppe der Lai:innen wird insbesondere dann relevant, wenn eine Person eine Erklärung zu einer KI-Entscheidung fordert, die sie selbst betrifft. Das kann beispielsweise der Fall sein, wenn sie ihr Recht auf Erklä-

rung in Anspruch nimmt, welches ihr gemäß der DSGVO zusteht. Für Lai:innen als Privatpersonen können Erklärungen jedoch auch nützlich sein, um die allgemeine Benutzung einer KI-Anwendung und die menschliche Entscheidungsfindung zu unterstützen.

3.3. Korrektheit vs. Verständlichkeit

Zwei grundlegende Eigenschaften einer Erklärung sind ihre Korrektheit und Verständlichkeit. Auch wenn man beide im ersten Moment als selbstverständlich für jede Erklärung ansehen könnte, ist dies in der Realität nicht automatisch der Fall. Der zentrale Unterschied ist, dass eine korrekte Erklärung die Nutzenden außer Acht lässt (Kim 2018). Das oberste Ziel ist hier, eine korrekte und präzise Erklärung zu liefern, ungeachtet dessen, ob sie auch von jedem nachvollzogen werden kann. Dies ist jedoch das Hauptkriterium einer verständlichen Erklärung. Einem Kindergartenkind würde man beispielsweise erklären, dass der Mensch vom Affen abstammt, auch wenn dies nicht korrekt ist, da beide lediglich gemeinsame Vorfahren besitzen. Einem Kindergartenkind, für das das Konzept der Evolution überfordernd sein könnte, hilft diese Vereinfachung jedoch, um die Kernaussage zu verstehen: Menschen und Affen sind sich sehr ähnlich.

Die beiden Eigenschaften lassen sich auch auf das technische Äquivalent einer durch XAI-Methoden erzeugten Erklärung abbilden. Korrektheit bedeutet hier Modelltreue (fidelity): Wie treu bildet die erzeugte Erklärung die tatsächliche Logik des ursprünglichen KI-Modells ab? Modelltreue ist nicht bei jeder XAI-Methode trivial nachzuweisen und steht gegebenenfalls im Widerspruch zu weiteren Eigenschaften wie beispielsweise dem Rechenaufwand, aber auch der Verständlichkeit (Gunning et al. 2019). Die Verständlichkeit einer erzeugten Erklärung hängt auch hier wieder vom Nutzenden ab. Die XUIs müssen daher an der Zielgruppe hinsichtlich ihrer Verständlichkeit evaluiert werden.

3.4. XAI-Beispielmethoden

Verschiedene XAI-Methoden haben unterschiedliche Vor- und Nachteile und sind somit für unterschiedliche Anwendungskontexte geeignet. In den letzten Jahren wurde eine Vielzahl von neuen XAI-Methoden vorgestellt und das dazugehörige Forschungsfeld ist weiterhin sehr aktiv. Hier werden daher nur Beispiele gängiger und typischer Methoden angeführt.

3.4.1. SHAP

Shapley Additive Explanations (SHAP) ist ein Verfahren für erklärbare KI, das auf den Shapley-Werten aus der Spieltheorie basiert. Das SHAP-Framework zielt darauf ab, den Beitrag der unterschiedlichen Merkmale zur Vorhersage eines Modells zu bestimmen (Lundberg und Lee 2017). Die Shapley-Werte, auf denen das Verfahren basiert, können den fairen Beitrag von

Spielerinnen und Spielern in einem kooperativen Spiel messen. In Bezug auf XAI kann ein KI-Modell als "Spiel" betrachtet werden, bei dem jedes Merkmal als "Spieler:in" fungiert (Molnar 2020).

SHAP liefert sehr komplexe Visualisierungen als Ergebnis. Die Methode ist daher nicht gut für Endnutzende geeignet liefert aber für Modellentwickler:innen und Domänenexpert:innen sehr viele Informationen. Es können verschiedene Visualisierungen generiert werden, die unterschiedliche Informationen liefern, z.B. die Wichtigkeit einzelner Merkmale, die Wichtigkeit in Abhängigkeit von Merkmalswerten oder dem Zusammenspiel unterschiedlicher Merkmale. Die Berechnung von SHAP ist recht aufwändig und der Rechenaufwand steigt exponentiell mit der Anzahl der Merkmale. Es gibt jedoch auch optimierte Verfahren wie z.B. TreeSHAP, die die Berechnungskomplexität für bestimmte Modelle deutlich reduzieren (Yang 2021).

SHAP bietet eine mathematisch fundierte Methode zum Erklären von KI-Modellen. Durch die Verwendung von Shapley-Werten kann SHAP den Beitrag einzelner Merkmale zur Modellvorhersage ermitteln und transparente, konsistente und lokal interpretierbare Erklärungen liefern.

3.4.2. Kontrafaktische Erklärungen

Kontrafaktische Erklärungen basieren auf kontrafaktischer Logik. Diese befasst sich mit der Untersuchung von Aussagen, die sich auf Ereignisse beziehen, die nicht tatsächlich eingetreten sind. Es werden hypothetische Szenarien und Bedingungen betrachtet, unter denen diese Szenarien wahr oder falsch wären. Kontrafaktische Erklärungen sind eine Anwendung der kontrafaktischen Logik in der erklärbaren KI, die darauf abzielt, die Ursache-Wirkungs-Beziehungen von Entscheidungen von KI-Modellen zu verstehen. Kontrafaktische Erklärungen geben Aufschluss darüber, wie sich eine Entscheidung ändert, wenn sich bestimmte Merkmale im Eingangsszenario ändern. Sie dienen dazu, das Verständnis von Entscheidungen oder Vorhersagen von KI-Modellen zu verbessern, indem sie alternative Szenarien betrachten und erklären, warum bestimmte Ergebnisse aufgetreten sind oder nicht. (Burkart und Huber 2021)

Als Erklärungen für KI-Modelle können kontrafaktische Erklärungen helfen, die Bedeutung und den Einfluss von Merkmalen auf die Modellvorhersagen zu verstehen. Sie ermöglichen es, hypothetische Szenarien zu erstellen, in denen bestimmte Merkmale geändert werden, und die Auswirkungen dieser Veränderungen auf die Entscheidung des Modells zu analysieren. Dies bietet Einblicke in die Sensitivität des Modells gegenüber unterschiedlichen Merkmalswerten und kann helfen, mögliche Vorurteile, Diskriminierungen oder unerwünschte Abhängigkeiten zu identifizieren. Kontrafaktische Erklärungen können dabei helfen, die Entscheidungslogik von komplexen Modellen besser zu verstehen und Vertrauen in die Entscheidungsfindung zu schaffen (Molnar 2020).

Durch die Analyse von kontrafaktischen Erklärungen können potenzielle Vorurteile oder Muster aufgedeckt werden, die zu unerwünschten oder unfairen Entscheidungen führen könnten. Dies ermöglicht es den Entwicklerinnen und Entwicklern, die Modelle zu verbessern. Es ermöglicht Betroffenen, die Entscheidungen zu verstehen, und gibt Hinweise auf Änderungsmöglichkeiten, die die Vorhersage verbessern können. Kontrafaktische Erklärungen können in tabellarischer Form oder in natürlicher Sprache dargestellt werden. Die Darstellung in natürlicher Sprache ist besonders für Endnutzerinnen und Endnutzer geeignet, die kein technisches Vorwissen haben (Molnar 2020).

Kontrafaktische Erklärungen bieten eine Möglichkeit, die Funktionsweise von KI-Modellen besser zu verstehen und Transparenz in deren Entscheidungsfindung zu schaffen. Sie helfen, das Vertrauen in KI-Systeme zu stärken und potenzielle Probleme oder Vorurteile aufzudecken, die durch die Entscheidungsprozesse entstehen könnten. Sie können dazu beitragen, die Vertrauensbildung in und Verantwortlichkeit von KI-Systemen zu verbessern, potenzielle Risiken und Einschränkungen in den Modellen zu erkennen und diese zu adressieren (Wachter et al. 2017).

4. Literaturverzeichnis

Adadi, Amina/Berrada, Mohammed (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Bao, Manuela/Vugrincic, Aline/Dreher, Ann-Katrin/Vonderau, Daniel/Tran, Hoa/Rill, Maria/Balaban, Silvia (2023). Datenschutz bei Künstlicher Intelligenz. Kompetenzzentrum KARL. https://doi.org/10.5445/IR/1000161675.

Bathaee, Yavar (2017). The Artificial Intelligence Black Box and the Failure of Intent and Causation. Harvard Journal of Law & Technology (Harvard JOLT) 31, 889.

Bundesministerium der Justiz (2023). Datenschutz-Grundverordnung. Online verfügbar unter https://www.bmj.de/DE/themen/digitales/DSGVO/DSGVO_artikel.html (abgerufen am 15.09.2023).

Burkart, Nadia/Huber, Marco F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317.

Card, Stuart K. (2018). The psychology of human-computer interaction. Crc Press.

Chromik, Michael/Butz, Andreas (2021). Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In: Carmelo Ardito/Rosa Lanzilotti/Alessio Malizia et al. (Hg.). Human-Computer Interaction – INTERACT 2021, Cham, 2021. Cham, Springer International Publishing, 619–640.

Davis, Brittany/Glenski, Maria/Sealy, William/Arendt, Dustin (2020). Measure Utility, Gain Trust: Practical Advice for XAI Researchers. In: Lisa O'Conner (Hg.). 2020 IEEE Workshop on TRust and EXpertise in Visual Analytics. TREX 2020 : virtual event, 25 October 2020 : proceedings, 2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX), Salt Lake City, UT, USA, 10/25/2020 - 10/30/2020. Piscataway, NJ, IEEE, 1–8.

Dwivedi, Rudresh/Dave, Devam/Naik, Het/Singhal, Smiti/Omer, Rana/Patel, Pankesh/Qian, Bin/Wen, Zhenyu/Shah, Tejal/Morgan, Graham/Ranjan, Rajiv (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Computing Surveys 55 (9), 1–33. https://doi.org/10.1145/3561048.

EU-Kommission (2019). Ethics guidelines for trustworthy AI vom 8.4.2019. Online verfügbar unter https://data.consilium.europa.eu/doc/document/ST-11481-2020-INIT/de/pdf (abgerufen am 15.09.2023).

EU-Kommission (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment vom 17.7.2020. Online verfügbar unter https://data.consilium.europa.eu/doc/document/ST-11481-2020-INIT/de/pdf (abgerufen am 15.09.2023).

EU-Kommission (2021a). Anhänge des Vorschlags für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung Harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, vom 21.4.2021 – COM(2021) 206 final. (abgerufen am 15.09.2021).

EU-Kommission (2021b). Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung Harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, vom 21.4.2021 – COM(2021) 206 final. (abgerufen am 15.09.2023).

Europäisches Parlament (2020). Framework of ethical aspects of artificial intelligence, robotics and related technologies 2020/2012(INL). Online verfügbar unter https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/2012(INL) (abgerufen am 15.09.2023).

Europäisches Parlament und Rat der Europäischen Union (2016). 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 (Datenschutz-Grundverordnung). (abgerufen am 15.09.2023).

Gerlings, Julie/Shollo, Arisa/Constantiou, Ioanna (2021). Reviewing the Need for Explainable Artificial Intelligence (xAI). Proceedings of the 54th Hawaii International Conference on System Sciences, 1284–1293. https://doi.org/10.24251/HICSS.2021.156.

Gunning, David/Stefik, Mark/Choi, Jaesik/Miller, Timothy/Stumpf, Simone/Yang, Guang-Zhong (2019). XAI-Explainable artificial intelligence. Science robotics 4 (37). https://doi.org/10.1126/scirobotics.aay7120.

Kim, Tae Wan (2018). Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review 39 (4), 261–283. https://doi.org/10.1007/s10462-011-9272-4.

Lackes, Richard/Markus, Siepermann (2018). Was ist "Künstliche Intelligenz (KI)"? Technische Universität Dortmund. Online verfügbar unter https://wirtschaftslexikon.gabler.de/definition/kuenstliche-intelligenz-ki-40285/version-263673 (abgerufen am Juni 2023).

London, Alex John (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report 49 (1), 15–21. https://doi.org/10.1002/hast.973.

Lundberg, Scott M./Lee, Su-In (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Matzka, Stephan (2021). Künstliche Intelligenz in den Ingenieurswissenschaften. Springer.

Miller, Tim (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences.

Mohseni, Sina/Zarei, Niloofar/Ragan, Eric D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems 11 (3-4), 1–45. https://doi.org/10.1145/3387166.

Molnar, Christoph (2020). Interpretable machine learning. Lulu.com.

Nauta, Meike/Trienes, Jan/Pathak, Shreyasi/Nguyen, Elisa/Peters, Michelle/Schmitt, Yasmin/Schlötterer, Jörg/van Keulen, Maurice/Seifert, Christin (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys 55 (13s), 1–42. https://doi.org/10.1145/3583558.

Rat der Europäischen Union (2020). Schlussfolgerungen des Vorsitzes – Die Charta der Grundrechte im Zusammenhang mit künstlicher Intelligenz und dem digitalen Wandel vom 21.10.2020 11481/20. Online verfügbar unter https://data.consilium.europa.eu/doc/document/ST-11481-2020-INIT/de/pdf (abgerufen am 15.09.2023).

Turing, Alan M. (1950). Computing machinery and intelligence. Mind 59 (236), 433–460.

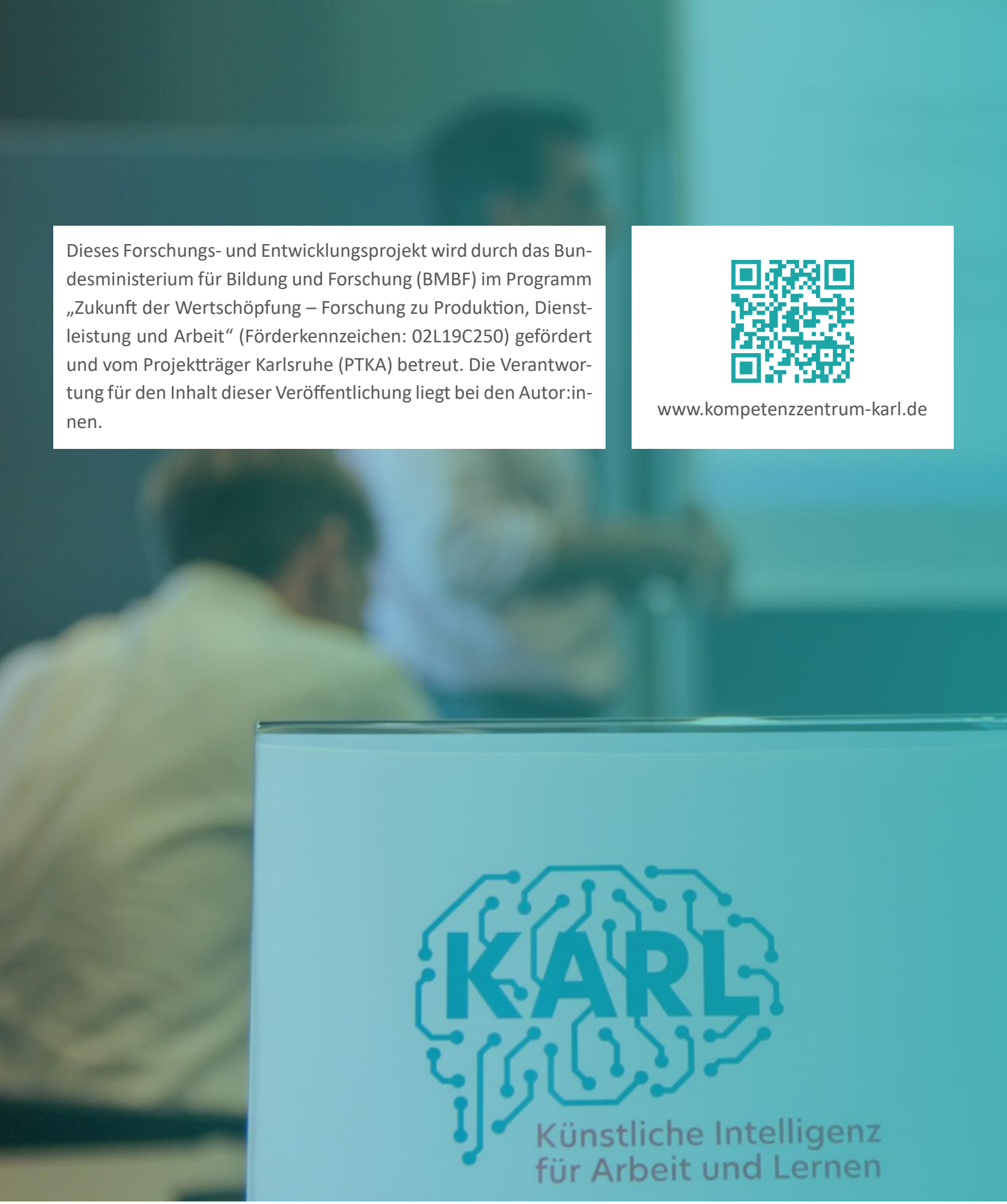
Wachter, Sandra/Mittelstadt, Brent/Russell, Chris (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech. 31, 841.

Wahlster, Wolfgang (2017). Die Speerspitze der Digitalisierung-Künstliche Intelligenz und ihre Entwicklung 2017.

Weber, Mathias/Buschbacher, Florian (2017). Künstliche Intelligenz-Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung. Bitkom e. V., DFKI, Berlin, Kaiserslautern.

Wilmott, Paul (2020). Grundkurs Machine Learning. Rheinwerk Verlag.

Yang, Jilei (2021). Fast treeshap: Accelerating shap value computation for trees.



Dieses Forschungs- und Entwicklungsprojekt wird durch das Bundesministerium für Bildung und Forschung (BMBF) im Programm „Zukunft der Wertschöpfung – Forschung zu Produktion, Dienstleistung und Arbeit“ (Förderkennzeichen: 02L19C250) gefördert und vom Projektträger Karlsruhe (PTKA) betreut. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen.

www.kompetenzzentrum-karl.de

KARL
Künstliche Intelligenz
für Arbeit und Lernen

Kompetenzzentren
Arbeitsforschung

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung