HOCHSCHULE RUHR WEST
UNIVERSITY OF APPLIED SCIENCES

Erklärbare KI im Bildungswesen: Konzeption eines Prüfrahmenwerks für die Erklärbarkeit von KI-Systemen

Bachelorarbeit

Studiengang Wirtschaftsinformatik
der Hochschule Ruhr West

Emre Selim Kosar
10015087

Erstprüfer:	Michael Schellenbach
Zweitprüfer:	Nils Fischer

Düsseldorf, Dezember 2024

Kurzfassung

Die rapide Entwicklung von Künstlicher Intelligenz (KI) hat eine transformative Wirkung auf zahlreiche Lebensaspekte, einschließlich des Bildungswesens. Die Integration von KI-Technologien im Bildungskontext bietet vielfältige Möglichkeiten, Lehr- und Lernmethoden effizienter zu gestalten. Um diese Potenziale jedoch voll auszuschöpfen und die Technologie vertrauenswürdig einsetzen zu können, ist eine erklärbare Gestaltung der KI-generierten Entscheidungsprozesse unerlässlich. Dennoch hat sich bisher kein Prüfrahmenwerk etabliert, anhand dessen Auditoren mangelhafte Erklärbarkeit und hierbei im Zusammenhang stehende potenzielle Risiken erkennen und an relevante Stakeholder kommunizieren können.

Das Ziel in der vorliegenden Arbeit ist es, ein Rahmenwerk zur Prüfung der Erklärbarkeit von KI-Systemen zu konzipieren und mit einem konkreten Anwendungsbeispiel aus dem Bildungswesen zu veranschaulichen.

Hierfür bietet Kapitel drei dieser Arbeit zunächst ein einheitliches Verständnis der Erklärbarkeit von KI-Systemen im Bildungswesen, indem es die Anwendungsgebiete, die Notwendigkeit, die Zielgruppen und den Informationsgehalt sowie potenzielle Risiken Erklärbarer KI im Bildungskontext analysiert.

In Kapitel 4 wird schließlich das Prüfrahmenwerk mit der direkten Demonstration der Kriterien am Anwendungsbeispiel vorgestellt. Um ein umfassendes Prüfrahmenwerk zu konzipieren, wurde die Literatur rund um Erklärbare KI untersucht und die Kriterien extrahiert, die andere Autoren für die Bewertung von Erklärbarkeit vorgeschlagen oder in ihrer Forschung verwendet haben. Das Prüfrahmenwerk wird hierfür in drei Dimensionen unterteilt, um die Vorgehensweise zu strukturieren: Die soziale, sozio-technische und technische Dimension. Das Werk kann als Leitfaden oder Orientierungshilfe zur Prüfung der Erklärbarkeit von KI-Systemen dienen und leistet einen Beitrag zur Etablierung verantwortungsvoller KI-Nutzung, die den Bedürfnissen und Erwartungen der Gesellschaft gerecht wird.

Abstract

The rapid development of artificial intelligence (AI) is having a transformative effect on many aspects of life, including education. The integration of AI technologies in the educational context offers diverse opportunities to make teaching and learning methods more efficient. To fully realize these potentials and use the technology in a trustworthy manner, it is essential to design AI-generated decision-making processes in an explainable manner. However, no widely accepted framework has yet been established to assist auditors in identifying insufficient explainability, evaluating associated potential risks, and communicating these to relevant stakeholders.

The aim of this thesis is to develop a framework for auditing the explainability of AI systems and to illustrate it through a specific use case in the educational context.

For this purpose, chapter three of this thesis provides a unified understanding of the explainability of AI systems in education by analyzing the application areas, the necessity, the target audiences and the informational content as well as potential risks of explainable AI in the educational context.

Chapter four introduces the audit framework and demonstrates its application through the use case. To create the comprehensive audit framework, existing literature on explainable AI was reviewed, and criteria proposed or used by other authors for evaluating explainability were extracted. The framework is divided into three dimensions in order to structure the approach: The social, socio-technical and technical dimension. It can serve as a guide or reference for auditing the explainability of AI systems and contributes to the establishment of responsible AI usage that meets societal needs and expectations.

Inhaltsverzeichnis

Kurzfassung ..... **II**

Abstract ..... **III**

Inhaltsverzeichnis ..... **IV**

Abbildungsverzeichnis ..... **VI**

Tabellenverzeichnis ..... **VI**

Abkürzungsverzeichnis ..... **VII**

Gender-Hinweis ..... **VIII**

1 Einleitung **1**

1.1 Relevanz und Motivation 1

1.2 Zielsetzung und Aufbau 1

1.3 Methodik und Forschungsdesign 2

1.4 Verwendete Literatur 3

2 Theoretische Grundlagen **4**

2.1 Künstliche Intelligenz 4

2.1.1 Formen Künstlicher Intelligenz 4

2.1.2 Funktionsweise von KI-Anwendungen 5

2.2 Erklärbare Künstliche Intelligenz 6

2.2.1 Transparenz vs. Erklärbarkeit vs. Interpretierbarkeit 7

2.2.2 Kategorien und Methoden zur Erklärbarkeit von KI-Modellen 8

3 (Erklärbare) Künstliche Intelligenz im Bildungswesen **10**

3.1 Einsatzgebiete von KI im Bildungswesen 11

3.1.1 KI bei Administrativen Aufgaben 11

3.1.2 KI in der Lehre 14

3.1.3 KI beim Lernen 16

3.2 Notwendigkeit der Erklärbarkeit von KI im Bildungswesen 18

3.2.1 Erklärbarkeit zur Entlarvung von Verzerrungen (Bias) 18

3.2.2 Erklärbarkeit zur kontinuierlichen Verbesserung 19

3.2.3 Erklärbarkeit zur Stärkung von Vertrauen und Akzeptanz der Ergebnisse 19

3.2.4 Erklärbarkeit als Grundlage für die Erarbeitung neuer Erkenntnisse 20

3.2.5 Erklärbarkeit als Teil der Gesetzgebung 20

3.2.6 Erklärbarkeit zur Klärung der Verantwortlichen 23

3.3 Empfänger und Informationsgehalt der Erklärungen im Bildungswesen 24

3.3.1 Bediener, Entscheider, Entscheidungs- und Datenbetroffene 24

3.3.2 Schöpfer 25
3.3.3 Prüfer 25
3.4 Potenzielle Risiken im Zusammenhang mit Erklärbarer KI im Bildungswesen 26
3.5 Konkretes Anwendungsbeispiel im Bildungswesen 28
3.5.1 GradeAI Solutions: Ein KI-System zur automatisierten Bewertung und Benotung von Prüfungsleistungen 28
3.5.1.1 Übersicht 28
3.5.1.2 Stakeholder/Akteure 29
3.5.1.3 Hochrisiko-KI-System gemäß der KI-VO 30
3.5.1.4 Wer ist für den Audit der Erklärbarkeit zuständig? 30

4 Prüfrahmenwerk für Erklärbare KI mit einem Anwendungsbeispiel aus dem Bildungswesen 31

4.1 Soziale Dimension: Fokus auf die menschenzentrierte Perspektive der Erklärbarkeit 32
4.2 Sozio-technische Dimension: Fokus auf die Mensch-Technik-Interaktion der Erklärbarkeit 37
4.3 Technische Dimension: Fokus auf die technologischen Grundlagen der Erklärbarkeit 42

5 Schlussbetrachtung 47

Anhang A 49

Literaturverzeichnis 50

Erklärung 63

Abbildungsverzeichnis

Abbildung 1: Vorlage eines möglichen Chatbots für Bildungseinrichtungen..... 14
Abbildung 2: Bündelung verschiedener KI-Funktionen in eine Anwendung 15
Abbildung 3: Lehrer Chatbot Alena..... 18

Tabellenverzeichnis

Tabelle 1: Übersicht über White-Box-/Black-Box-Charakter von Modellen, die für das maschinelle Lernen eingesetzt werden..... 9

Abkürzungsverzeichnis

Abs.	Absatz
Anh.	Anhang
Art.	Artikel
Engl.	Englisch
KI	Künstliche Intelligenz
KI-VO	KI-Verordnung
orig.	original

Gender-Hinweis

In dieser Arbeit wird ein generisches Maskulinum verwendet. Sofern die Aussagen es erfordern, sind immer alle Geschlechtsidentitäten mitgemeint. Die Verwendung der maskulinen Form dient nur der besseren Lesbarkeit und beinhaltet keine Wertung.

1 Einleitung

1.1 Relevanz und Motivation

In jüngster Vergangenheit hat das Thema Künstliche Intelligenz eine hohe Aufmerksamkeit in der Öffentlichkeit erfahren. Diese Entwicklung ist darauf zurückzuführen, dass KI-Systeme in einer Vielzahl von Branchen zunehmend zum Einsatz kommen, die für das tägliche Leben vieler Menschen relevant sind, wie etwa in der Medizin oder gar bei der Wettervorhersage. Auch im Bildungswesen hat KI erheblich an Bedeutung gewonnen und zeigt das Potenzial, das Bildungswesen nachhaltig zu revolutionieren. Von intelligenten Tutorensystemen bis hin zu automatisierten Lernumgebungen bietet KI zahlreiche Vorteile für Schüler, Lehrkräfte und die administrativen Abteilungen schulischer Einrichtungen.

Gleichzeitig birgt der revolutionäre Charakter dieser Technologie Herausforderungen und Notwendigkeiten, insbesondere im Hinblick auf die Erklärbarkeit der Entscheidungsprozesse einer KI. Im Vergleich zu weniger sensiblen Bereichen, in denen die Erklärbarkeit von KI-Systemen weniger relevant ist, wie etwa bei den Empfehlungsalgorithmen von Netflix oder Instagram, ist im Bildungswesen eine nachvollziehbare Entscheidungsfindung unerlässlich. Das ist insbesondere in Situationen der Fall, in denen Schüler die Auswirkungen der Entscheidungen der KI wahrnehmen können, jedoch nicht die Beweggründe der dahinterstehenden Entscheidungsfindung. Erklärungen für KI-Systeme tragen unter anderem dazu bei, Vertrauen in KI-gestützte Bildungstechnologien zu stärken, rechtlichen Vorgaben zu erfüllen und falsche Klassifikatoren aufzudecken. Diese Aspekte werden unter dem Begriff der Erklärbaren KI erforscht.

Für risikobehaftete KI-Systeme im Bildungswesen, die direkte Auswirkungen auf Lern- und Bildungswege haben, ist somit das Verständnis und die Validierung ihrer internen Entscheidungsprozesse von zentraler Bedeutung. Trotz der schnell wachsenden Relevanz von KI, dem zunehmenden Einfluss auf das Bildungswesen und der allgemeinen Notwendigkeit von Erklärungen hat sich bisher kein umfassendes Prüfrahmenwerk mit Kriterien zur Auditierung der Erklärbarkeit von KI-Systemen etabliert, ganz zu schweigen im Kontext eines Anwendungsfalls im Bildungswesen.

1.2 Zielsetzung und Aufbau

Ziel und Nutzen dieser Arbeit besteht darin, ein Prüfrahmenwerk zu konstruieren, anhand dessen Kriterien Auditoren die Erklärbarkeit von jeglichen KI-Systemen überprüfen können. Zu den Auditoren gehören neben speziell ausgebildeten internen und externen Prüfern auch weitere Stakeholder, die in das KI-Projekt eingebunden sind wie bspw. Anwender. Die einzelnen Kriterien des Prüfrahmenwerks werden im Kontext eines Anwendungsfalls im Bildungswesen veranschaulicht, um dessen praktische An-

wendbarkeit zu demonstrieren, einen möglichen Ausgangspunkt für die Prüfung der Erklärbarkeit von KI-Systemen in realen Bildungsszenarien zu bieten und als Grundlage oder Inspiration für zukünftige Ansätze zu dienen. Dabei wird die Erklärbarkeit aus einer anwendungsbezogenen Perspektive untersucht, indem soziale, sozio-technische und technische Aspekte einbezogen werden, die für die Erklärbarkeit relevant sind. Es erfolgt jedoch keine detaillierte Betrachtung spezifischer Erklärbarkeitsmodelle wie LIME oder SHAP.

Konkret lässt sich daraus folgende Forschungsfrage ableiten:

Anhand welcher Kriterien können Auditoren die Erklärbarkeit von KI-Systemen überprüfen, und wie lässt sich das entwickelte Prüfrahmenwerk im Kontext eines Anwendungsfalls im Bildungswesen veranschaulichen.

Der Aufbau der vorliegenden Arbeit gliedert sich in fünf Kapitel. Kapitel zwei führt in das Themenkomplex der KI mit den dazugehörigen zentralen Begriffen wie dem Maschinellen Lernen ein und vermittelt die Grundlagen des Forschungsgebiets der Erklärbaren KI. In Kapitel drei werden die Einsatzgebiete von KI im Bildungswesen vorgestellt und die Notwendigkeit der Erklärbarkeit von KI im Bildungskontext herausgearbeitet. Anschließend werden die Empfänger und der Informationsgehalt der Erklärungen im Bildungswesen sowie potenzielle Risiken im Zusammenhang mit Erklärbarer KI im Bildungsbereich untersucht. Zum Abschluss von Kapitel drei wird ein konkretes Anwendungsbeispiel vorgestellt, anhand dessen das Prüfrahmenwerk veranschaulicht werden soll. Das entwickelte Prüfrahmenwerk sowie die direkte Demonstration anhand des Anwendungsbeispiels wird anschließend in Kapitel vier vorgestellt. Kapitel fünf schließt die vorliegende Arbeit mit einer Schlussbetrachtung ab.

In dieser Arbeit wird der Begriff „erklärbare Künstliche Intelligenz“ mit „Erklärbare KI“ abgekürzt. Um die Lesbarkeit zu fördern, wird das Adjektiv „Erklärbare“ großgeschrieben, da es sich um einen festen Begriff handelt.

Zudem werden zur Vereinfachung die Begriffe *Lernende* und *Lehrende* verwendet, um sowohl Schüler als auch Studierende bzw. Lehrkräfte unabhängig von der Art der Bildungseinrichtung anzusprechen. Diese neutralen Begriffe wurden bewusst gewählt, um die Vielseitigkeit der Zielgruppen zu verdeutlichen und Wiederholungen zu vermeiden.

1.3 Methodik und Forschungsdesign

Die Forschungsmethode dieser Arbeit basiert auf einer umfassenden Literaturanalyse, die verschiedene Perspektiven und Ansätze zusammenführt, um ein umfassendes Verständnis für die Thematik zu entwickeln.

Es werden breit gefächerte Aspekte von KI und Erklärbarer KI im Bildungsbereich bzw. auch allgemein ohne Bildungskontext betrachtet und in Kapitel drei zusammengeführt.

Zudem werden bereits bestehende Kriterien zur Prüfung der Erklärbarkeit von KI-Systemen recherchiert und auf Grundlage dessen in das Prüfrahmenwerk in Kapitel vier integriert. Es ist nicht das Ziel, neue Kriterien bzw. Metriken zu schaffen, sondern

die bereits vorliegenden jedoch verstreuten Metriken in diesem Prüfrahmenwerk anhand einer Einteilung von drei Dimensionen zusammenzuführen. Jedes Kriterium wird auf Basis des Anwendungsbeispiels aus dem Bildungswesen veranschaulicht.

1.4 Verwendete Literatur

Um einen umfassenden Überblick über das Gebiet der KI und Erklärbaren KI im Bildungswesen zu schaffen, greift die vorliegende Arbeit auf eine vielfältige Auswahl an Literaturquellen (Akademische Veröffentlichungen, Online-Artikel und Gesetze) zurück.

Eine zentrale Grundlage für Kapitel drei bietet zum einen der Schlussbericht des mmb Instituts zu KI in der Bildung, der im Auftrag der Deutschen Telekom Stiftung (2021) erstellt wurde, als auch die Arbeit von Samek & Müller (2019). Ergänzend dazu liefern Webseiten von Anbietern der jeweiligen KI-Systeme im Bildungswesen einen Überblick über Einsatzmöglichkeiten und Vorteile der Anwendung. Zur Beschreibung der Empfänger und Informationsgehälter der Erklärungen eignet sich die Klassifikation der Akteure von Tomsett et al. (2018) hervorragend. Rechtliche Aspekte beginnend von der DSGVO bis zu der KI-Verordnung verweisen entweder direkt die Gesetzestexte oder werden auf Basis verschiedener Quellen formuliert. Besonders bedeutsam ist hierbei die Diplomarbeit von Moser (2023), die einen Überblick über die rechtlichen Anforderungen an KI-Systeme in Bezug auf Erklärbarkeit verschafft.

Das in Kapitel vier vorgestellte Prüfrahmenwerk basiert auf einer umfassenden Analyse zu Kriterien bzw. Metriken im Kontext der Erklärbaren KI aus unterschiedlichen Quellen, darunter Arbeiten von Chatti et al. (2022), Hoffman et al. (2018), Rosenfeld (2021), Krause et al. (2016), Coroama & Groza (2022), Langer et al. (2021), Kulesza et al. (2015), Arrieta et al. (2019) und Zhou et al. (2024). Jedoch liegen die Metriken sehr verstreut und weniger organisiert vor, was eine praxisnahe Nutzung als Richtlinie oder Prüfrahmenwerk erschwert. Sokol & Flach (2019) und Langer, Baum et al. (2021) hingegen bieten jeweils ein Prüfrahmenwerk mit kategorisierter Herangehensweise zur Prüfung der Erklärbarkeit von KI-Systemen an und wurden in der vorliegenden Arbeit als Inspiration herangezogen. Jedoch fehlt dem einem Rahmenwerk die Tiefe, dem anderen die Berücksichtigung rechtlicher Aspekte, und beiden mangelt es an Praxisnähe.

2 Theoretische Grundlagen

Zum Verständnis von Kapitel drei und vier werden grundlegende zentrale Begriffe vorausgesetzt, die im folgenden Kapitel erläutert werden.

2.1 Künstliche Intelligenz

KI ist ein vielseitiger Begriff, der in verschiedenen Kontexten unterschiedlich interpretiert und definiert wird, weshalb „es keine kohärente Definition der KI gibt“ (Richter et al., 2019, S. 36).

In Wörterbüchern gibt es teilweise mehrere Begriffsdefinitionen, wobei z.B. das Merriam-Webster Dictionary KI als „die Fähigkeit von Computersystemen oder Algorithmen, intelligentes menschliches Verhalten nachzuahmen“ definiert (Merriam-Webster, Inc., 2024, Übersetzung durch den Autor). Aufgaben wie das Erkennen von Mustern oder das Lernen aus Erfahrungen, die normalerweise menschliche Intelligenz erfordern, können durch KI ausgeführt werden.

Der Begriff „künstlich“ in Künstlicher Intelligenz bezieht sich auf ihren Ursprung als menschengemachtes Produkt und unterscheidet sich von natürlicher Intelligenz, die biologisch oder evolutionär entstanden ist. KI ist somit ein Produkt menschlichen Einfallsreichtums und Genialität (Fetzer, 1990).

In der Geschichte der KI gab es Phasen des Rückgangs, die als „KI-Winter“ bezeichnet werden. Diese Perioden sind gekennzeichnet durch sinkendes Interesse und Investitionen in KI-Technologien, oft bedingt durch enttäuschte Erwartungen und stagnierende Fortschritte (Richter et al., 2019, S. 35). Das die derzeitige KI-Forschung mitten im KI-Frühling ist, lässt sich durch das hohe Interesse und die Finanzierungsrunden nicht von der Hand weisen. Der Umsatz des weltweiten KI-Marktes wird für das Jahr 2024 auf etwa 279 Milliarden US-Dollar geschätzt, was eine Zunahme von etwa 80 Milliarden US-Dollar im Vergleich zum Vorjahr und eine Steigerung von rund 220 Milliarden US-Dollar im Vergleich zu 2020 darstellt (Grand View Research, 2024).

2.1.1 Formen Künstlicher Intelligenz

Zur besseren Verständlichkeit der „Intelligenz“ einer KI hat sich die Einteilung in starke und schwache KI als hilfreich erwiesen. Die **schwache KI** ist darauf beschränkt, bestimmte, vordefinierte Anwendungsfälle zu lösen, wie die Spracherkennung durch Systeme wie Alexa oder die Bereitstellung von Korrekturvorschlägen bei Suchanfragen. Ein weiteres Anwendungsbeispiel ist die Generierung textbasierter Antworten, wie es bei ChatGPT der Fall ist. Es beantwortet lediglich die gestellten Fragen aber zeigt keinen eigenen Willen oder ein Bewusstsein. Die meisten der heutzutage in verschiedenen gesellschaftlichen Bereichen genutzten KI-Anwendungen entsprechen der Definiti-

on einer schwachen KI. Im Gegensatz dazu setzt die **starke KI** intelligentes Handeln voraus, die mindestens genauso gut, wenn nicht sogar besser ist als die menschliche Intelligenz (Richter et al., 2019, S. 37). „Dazu gehören bspw. das Entscheiden unter Unsicherheit, Kreativität, Empathie und ein eigenes Bewusstsein“ (Harwardt & Köhler, 2023, S. 23).

2.1.2 Funktionsweise von KI-Anwendungen

Die Nutzung technischer Hilfsmittel zur Unterstützung menschlicher Entscheidungsprozesse ist kein neues Phänomen. Neben Instrumenten und Systemen wie dem Abakus, den Logarithmentafeln und dem Taschenrechner hat die KI als Teilgebiet der Informatik in diesem Kontext bereits in der Vergangenheit eine zentrale Rolle eingenommen.

Regelbasierte Systeme, einschließlich vieler Decision Support Systems (DSS) und deren intelligenter Varianten mit KI-Funktionen (Intelligent Decision Support Systems, IDSS) wurden entwickelt, um Experten bei der Entscheidungsfindung zu unterstützen, indem sie auf Grundlage umfangreicher Datenanalysen Empfehlungen gaben (Xaqt, Inc., 2023). Diese Systeme beruhen häufig auf fest programmierten Regeln, die durch menschliche Programmierer vorgegeben wurden (Gillis & Moore, 2023).

Allerdings stießen diese Systeme an ihre Grenzen, vor allem in Bezug auf die Verarbeitung und das Lernen aus den immer größer werdenden Datenmengen. Ein zentrales Problem dieser Systeme war ihre Unfähigkeit, eigenständig zu lernen oder sich flexibel an neue Bedingungen und Herausforderungen anzupassen. Traditionelle regelbasierte Systeme basieren auf starren Algorithmen, die nur die vordefinierten Regeln befolgen und keine eigenständige Anpassung an neuartige Situationen ermöglichen (Johnston, 2024). Mit dem rasanten Anstieg von Datenmengen (Engl. Big Data) wurde es notwendig, einen neuen Ansatz für die informationstechnische Verarbeitung von Daten zu entwickeln, um somit den Herausforderungen gerecht zu werden.

Hier setzt das maschinelle Lernen (Engl. **Machine Learning**) an. Machine Learning ermöglicht es Systemen selbstständig aus Daten zu lernen, anstatt strikt vorgegebene Regeln zu befolgen. Statt jede Entscheidung explizit durch Softwareentwickler kodieren und explizieren zu lassen, lernen die Systeme durch die Analyse großer Datensätze, Muster und Gesetzmäßigkeiten zu erkennen und Vorhersagen oder Entscheidungen zu treffen. Im Gegensatz zu den traditionellen regelbasierten Systemen, die nur für spezifische und vordefinierte Probleme geeignet sind, zeichnet sich Machine Learning durch seine Flexibilität und Anpassungsfähigkeit aus. Es lernt nicht nur aus den vorhandenen Daten, sondern kann auch neue Problemstellungen und unbekannte Situationen bewältigen, indem es kontinuierlich seine Algorithmen auf Basis der Trainingsdaten optimiert (Harwardt & Köhler, 2023).

Machine Learning Algorithmen können in verschiedenen Formen auftreten, wie etwa überwachtes, unüberwachtes oder durch bestärkendes Lernen. Das wesentliche Prinzip bleibt jedoch bei allen Formen gleich: „Die richtigen Trainingsdaten zu verwenden,

um die richtigen Algorithmen zu erstellen, um die richtigen Aufgaben zu erfüllen“ (Flach, 2012, S. 20, Übersetzung durch den Autor).

Während bei den Formen des maschinellen Lernens menschliche Eingriffe erforderlich sind, insbesondere in den Phasen der Daten- oder Merkmalsauswahl, um sicherzustellen, dass die Algorithmen effektiv lernen können, verfolgt das Verfahren des „Tiefen Lernens“ (Engl. **Deep Learning**) die Strategie einer möglichst geringen manuellen Eingriffsnotwendigkeit. Deep Learning ist eine moderne Ausprägung des maschinellen Lernens und reduziert den Bedarf an manuellen Eingriffen in gewissem Maße, da Verfahren des Deep Learnings in der Lage sind, komplexe Muster in großen Datenmengen selbstständig zu lernen und automatisch Merkmale aus den Daten zu extrahieren, was die manuelle Merkmalsauswahl in der Regel obsolet macht (Alpaydin, 2022, S. 329).

Künstlich neuronale Netze (Engl. **Artificial Neural Networks**) ahmen die Funktionsweise des menschlichen Gehirns und dessen neuronale Verbindungen nach. Sie ermöglichen im Vergleich zu den bisherigen Formen des maschinellen Lernens die Erkennung von Zusammenhängen, die diesen zuvor verborgen geblieben waren. Darüber hinaus profitieren sie erheblich von einer größeren Menge an Trainingsdaten (Harwardt & Köhler, 2023, S. 26). „Diese Datensätze können sowohl strukturierte als auch unstrukturierte Daten enthalten” (Richter et al., 2019, S. 40). Diese neuronalen Netze sind die Grundlage für die Verfahren des Deep Learning, welches diese Fähigkeit zur Verarbeitung großer Datenmengen weiter nutzt um komplexe Muster zu lernen. Ein Beispiel für die Anwendung von Deep Learning ist der Online-Händler Zalando, der dieses Verfahren in seinen Logistikzentren einsetzt, um die Effizienz zu steigern. Diese Methode hat dazu beigetragen, die Laufwege der Mitarbeiter zu verkürzen und die “Pickrouten“ zu optimieren, was zu optimalen Wegen durch die verschiedenen Regale führt (Harwardt & Köhler, 2023; Seward, 2015).

2.2 Erklärbare Künstliche Intelligenz

Im Bereich des maschinellen Lernens gelten viele komplexe KI-Modelle, wie tiefe neuronale Netze, als **Black-Box-Modelle**, da sie aufgrund ihrer komplizierten internen Strukturen und Berechnungen Außenstehenden nicht offenlegen, wie die Vorhersagen zustande gekommen sind (Zednik, 2021).

Diese Undurchsichtigkeit stellt ein zentrales Problem mit unter anderem praktischen, rechtlichen und ethischen Folgen dar, besonders bei KI-Anwendungen, die „soziale Konsequenzen für die Teilhabe Einzelner (z.B. Ablehnung von Bewerbern)“, als auch Folgen für die gesamte Gesellschaft haben können (Zweig, 2018, S. 33). Burrell (2016) und Ribeiro et al. (2016) argumentieren, dass Endbenutzer Maschinen, deren Funktionsweise sie nicht verstehen, wahrscheinlich weder vertrauen noch ihnen die Kontrolle überlassen werden. Die Undurchsichtigkeit von KI-Systemen kann es Regulierungsbehörden außerdem erschweren, festzustellen, ob ein KI-System Daten fair und sicher verarbeitet (Rieder & Simon, 2017, nach Zednik, 2021, S. 266). Ethisch gesehen ist

Erklärbare KI zudem daher von Bedeutung, da sie sicherstellt, dass die Entscheidungen von KI-Systemen transparent und gerecht nachvollziehbar sind, wodurch mögliche Diskriminierungen oder Ungleichbehandlungen vermieden werden.

Infolgedessen wurde die Entwicklung des Forschungsfeldes namens **Erklärbare Künstliche Intelligenz** (Engl. Explainable Artificial Intelligence, abgekürzt **XAI**) vorangetrieben. Die DARPA definiert das Forschungsfeld der Erklärbaren KI als „KI-Systeme, die einem menschlichen Benutzer ihre Beweggründe erklären, ihre Stärken und Schwächen beschreiben und ein Verständnis dafür vermitteln können, wie sie sich in Zukunft verhalten werden“ (Gunning & Aha, 2021, Übersetzung durch den Autor). Hierbei sollen die Erklärbaren KI-Systeme den Black-Box-Modellen entgegenwirken, indem sie die Nachvollziehbarkeit und Benutzerfreundlichkeit erhöhen, ohne dabei die Leistungsfähigkeit der KI-Modelle zu beeinträchtigen (Escalante et al., 2018; Minh et al., 2022, S. 3504).

Solche Informationen bzw. Erklärungen sind in einigen Fällen bereits vor der Nutzung oder bevor das KI-Modell eine Entscheidung generiert hat notwendig (ex-ante-Information). In weiteren Fällen kann es zudem erforderlich sein, eine Erklärung zu generieren, nachdem die Entscheidung bereits getroffen wurde (ex-post-Information) (Martini, 2019).

2.2.1 Transparenz vs. Erklärbarkeit vs. Interpretierbarkeit

Die Informationen der erklärbaren KI-Modelle müssen nicht nur zugänglich, sondern auch verständlich sein (HEG-KI, 2019). In Bezug auf KI unterscheidet man hierbei konkret zwischen den Begriffen **Transparenz**, **Erklärbarkeit** und **Interpretierbarkeit**.

Transparenz im Allgemeinen bedeutet Sichtbar-Sein oder Sichtbar-Machen und zielt darauf ab, dass das Verhalten der KI-Anwendung vollständig nachvollziehbar ist (Döbel et al., 2018, S. 30). Transparenz ist das Gegenteil der Undurchsichtigkeit bzw. der „Black-Box-Natur“ von KI-Modellen (Lipton, 2018, S. 12). Das Ziel, die Transparenz zu erhöhen, wird bspw. durch das schlichte Offenlegen von Daten, wie etwa der Enthüllung des Quellcodes erfüllt (Authority of the House of Lords, 2018, S. 38). Ein weiteres Beispiel ist das Informieren des Nutzers, wenn sie mit einem KI-System interagieren, es sei denn, es ist offensichtlich (Future of Life Institute, 2024). In der Praxis ist die Forderung nach vollständiger Transparenz jedoch oft ungewünscht. Es reicht oft nicht aus, lediglich den Quellcode oder andere technische Details offenzulegen, da es dem Anwender nicht automatisch erklärt, warum eine spezifische Entscheidung getroffen wurde (Authority of the House of Lords, 2018, S. 38; Döbel et al., 2018, S. 30). Das lässt darauf schließen, dass die Forderung nach „mehr Daten“ nicht immer vorteilhaft ist, sondern letztendlich sogar dem Ziel einer höheren Transparenz entgegenwirken kann. Somit ist Transparenz zwar ein wichtiges Konzept, reicht aber nicht immer aus, um alle Fragen zu beantworten, die sich aus der Nutzung von KI-Anwendungen ergeben (Kroll et al., 2017, S. 639). Bis zu einem gewissen Grad besteht ein Zusammenhang zwischen Transparenz und Erklärbarkeit, da eine Erklärung teilweise durch Transparenz bedingt ist (Transparenz ist Voraussetzung für Erklärbarkeit). Beide

Grundsätze sollten jedoch unabhängig voneinander betrachtet werden, da die Anforderungen der Erklärbarkeit tiefgründiger sind und über das bloße Offenlegen von Informationen hinausgehen (Moser, 2023, S. 34).

Erklärbarkeit geht einen Schritt weiter. Der Duden definiert das Wort „erklären“ als „deutlich machen“ oder „erläutern, dass der bzw. die andere die Zusammenhänge versteht“ (*Duden, „erklären“*, o. J.). Im Kontext der KI wird Erklärbarkeit so beschrieben, dass ein KI-Modell seine Entscheidungen auf eine Art darlegt, die für Menschen nachvollziehbar sind. Dabei steht im Vordergrund die Frage, *warum* das KI-Modell diese spezifische Entscheidung getroffen hat. Der Nutzer soll auf eine klare und intuitive Weise die Begründung für die bestimmte Entscheidung des Algorithmus verstehen können, wobei sowohl die Wahrnehmung und der Wissensstand der Zielperson als auch die Ausrichtung der Fragestellung beim Erstellen von Erklärungen notwendigerweise zu berücksichtigen sind (Kraus et al., 2021, S. 18). Im Gegensatz zur Transparenz, die sich auf die Offenlegung des gesamten Modells und dessen Funktionsweise bezieht, ist die Erklärbarkeit oft auf die wesentlichen Einflussfaktoren einer spezifischen Entscheidung fokussiert, was technisch einfacher zu erreichen ist (Döbel et al., 2018, S. 30).

Interpretierbarkeit hingegen bezieht sich auf die Funktionsweise der inneren Logik und der Mechaniken des KI-Modells. Es geht darum, das „Wie“ zu verstehen, also wie die Komponenten des KI-Modells eine Zuordnung zwischen den Eingaben und Ausgaben herstellen.

Beim Vergleich der beiden gängigen Konzepte, Erklärbarkeit und Interpretierbarkeit, wird deutlich, dass sich die Erklärbarkeit auf die Erläuterung der getroffenen Entscheidungen konzentriert, während die Interpretierbarkeit sich auf das Verständnis der inneren Funktionsweise der Modelle konzentriert. Somit erfordert die Interpretierbarkeit ein höheres Maß an Detailgenauigkeit als die Erklärbarkeit und richtet sich somit eher an KI-Experten oder -Forscher, die das Modell verbessern oder debuggen müssen, wohingegen die Erklärbarkeit besonders für Endnutzer wichtig ist, die kein Domänenwissen haben (Giosmin, 2023; Raza, 2024).

In dieser Bachelorarbeit wird die Bezeichnung „erklärbare KI“ (anstelle von „interpretierbarer“ oder „transparenter“ KI) genutzt, um widerzuspiegeln, das hierbei der Fokus auf den Einsatz wirksamer KI-Erklärungen liegt, die für den Menschen ohne technisches Fachwissen verständlich sind.

2.2.2 Kategorien und Methoden zur Erklärbarkeit von KI-Modellen

KI-Modelle werden grundsätzlich in zwei Kategorien unterteilt, die **White-Box-** und **Black-Box-Modelle**. Um zu verstehen, wie gut ein Modell nachvollziehbar ist und welche Erklärungsmethoden notwendig sind, um die Ergebnisse eines KI-Modells erklärbar zu machen, ist diese Klassifizierung notwendig.

Die sogenannten **White-Box-Modelle** zeichnen sich dadurch aus, dass ihre interne Struktur und Funktionsweise „von Haus aus“ offen, leicht verständlich und interpretier-

bar ist. Sie erlauben einen klaren Einblick in den Entscheidungsprozess, d.h. wie Eingabe auf Ausgabe abgebildet wird und welche Variablen den größten Einfluss haben (Žlahtič et al., 2024, S. 2). Gängige Beispiele für White-Box-Modelle sind Entscheidungsbäume oder lineare/logistische Regressionsmodelle (siehe Tabelle 1).

Im Gegensatz dazu stehen die **Black-Box-Modelle**, welche hochkomplex sind und ihre internen Strukturen, Verhaltensweisen und Grundlagen für die Vorhersagen schwer zu durchschauen sind. Lediglich die Beziehung zwischen Ein- und Ausgabedaten ist bekannt. Somit ist es von Hause aus nicht möglich, einen Einblick in die Algorithmen zu bekommen (Žlahtič et al., 2024, S. 2). Ein Beispiel hierfür sind neuronale Netze, „da die von ihnen gelernten Zusammenhänge und Datenrepräsentationen so komplex und abstrakt sind, dass Menschen – und selbst Experten – sie nicht mehr nachvollziehen können“ (Schaaf, 2020). Tabelle 1 bietet eine Übersicht über die White-Box- und Black-Box-Charakteristiken von Modellen, die für das maschinelle Lernen eingesetzt werden (Arrieta et al., 2019, S. 14, nach Kraus et al., 2021, S. 20).

Zur Bewältigung des sogenannten „Black-Box-Problems“ wurden verschiedene Ansätze wie **LIME** (Local Interpretable Model-Agnostic Explanations) oder **SHAP** (Shapley Additive exPlanations) entwickelt. Diese Methoden schaffen Transparenz, ohne die gesamte innere Logik offenzulegen. Sie fügen eine zusätzliche Erklärungsebene hinzu, indem sie die Merkmale des Modells analysieren und diese Analysen mit den Entscheidungen des Black-Box-Modells verknüpfen. Mithilfe dieser können komplexe Black-Box-Modelle verständlich gemacht werden, indem sie Nutzern zeigen, welche Faktoren die Entscheidungen des Modells beeinflusst haben (Žlahtič et al., 2024, S. 2).

Es ist zu erwähnen, dass Black-Box-Modelle zwar schwerer zu erklären sind als White-Box-Modelle, jedoch performanter sind. Das macht es umso wichtiger, undurchsichtige KI-Modelle durch geeignete Ansätze verständlicher zu gestalten.

Tabelle 1: Übersicht über White-Box-/Black-Box-Charakter von Modellen, die für das maschinelle Lernen eingesetzt werden

KI-Modell	Transparenz			White-Box/ Black-Box
	Simulierbarkeit	Unterteilbarkeit	Algorith. Transparenz	
Neuronale Netze	X	X	X	Black-Box
Ensemble-Modelle (z. B. Tree Ensembles)	X	X	X	Black-Box
Support Vector Machines	X	X	X	Black-Box
Bayes-Netze	(✓)	(✓)	✓	White-Box*
Lineare/logistische Regressionsmodelle	(✓)	(✓)	✓	White-Box*
Entscheidungsbäume (Decision Trees)	(✓)	(✓)	✓	White-Box*

Tabelle von Arrieta et al. (2019), nach Kraus et al. (2021).
* Gilt im Falle nachvollziehbarer Eingangsparameter und generell bei Unterteilbarkeit. ✓= Erfüllt, X=Nicht Erfüllt

3 (Erklärbare) Künstliche Intelligenz im Bildungswesen

Auch das Bildungswesen blieb von der rasanten Entwicklung der KI nicht unberührt. KI-basierte Technologien verändern grundlegend, wie Wissen sowohl vermittelt als auch angeeignet wird. Bereits heute werden digitale Nachhilfe-Programme mit KI-Funktionen angeboten, um Lernenden überall und zu jeder Zeit Unterstützung zu bieten, dabei deutlich günstiger als herkömmliche Nachhilfelehrer (Wenig, 2024).

Jedoch hat die rasante Entwicklung der KI und ihre Anwendung in Branchen wie dem Bildungswesen nicht nur technologische Fortschritte, sondern auch rechtliche, ethische und gesellschaftliche Fragestellungen aufgeworfen. Lehrkräfte, Schüler und Eltern müssen die Entscheidungswege der eingesetzten KI-Modelle verstehen, um Vertrauen in die noch relativ junge Technologie zu entwickeln. Auditoren müssen sicherstellen können, dass die Lernprozesse fair, nachvollziehbar und individuell angepasst sind. Hierbei machen Regulierungen wie die DSGVO oder die KI-VO besonders Druck, um möglichen ethischen Problemen entgegenzuwirken. Besonders, da Entscheidungen der KI-Systeme einen direkten Einfluss auf den Bildungserfolg und die Zukunftschancen von Lernenden haben, ist die Erklärbarkeit der KI im Bildungswesen unentbehrlich und hat ihre eigenen unverwechselbaren Bedingungen, die sich von anderen Bereichen unterscheidet.

Dieses Kapitel gibt als erstes einen Überblick über die verschiedenen Einsatzgebiete von KI im Bildungswesen, um zu beleuchten, wie KI in der Lehre, beim Lernen und bei Administrativen Aufgaben Anwendung findet oder finden kann. Als nächstes geht es um die konkrete Notwendigkeit der Erklärbarkeit von KI im Bildungskontext, und folglich die verschiedenen Empfänger und der Informationsgehalt der Erklärungen von KI-Systemen im Bildungswesen. Anschließend werden die potenziellen Risiken im Zusammenhang mit Erklärbarer KI-Systeme im Bildungsbereich untersucht. Als letztes wird ein konkreter Anwendungsfall vorgestellt, anhand dessen das Prüfrahmenwerk veranschaulicht werden soll.

3.1 Einsatzgebiete von KI im Bildungswesen

Durch eine Analyse der aktuellen Forschungsliteratur sowie der neusten Entwicklungen im Bildungswesen wird deutlich, dass KI-Anwendungen bereits in vielfältigen Bereichen des Bildungswesens eingesetzt werden und sich kontinuierlich weiterentwickeln. Diese Anwendungsmöglichkeiten ermöglichen es Lehrenden, ihre Effizienz und Effektivität zu steigern, während den Lernenden personalisierte und optimierte Lernerfahrungen geboten werden.

Die Einsatzgebiete von KI im Bildungswesen lassen sich in drei zentrale Kategorien unterteilen**: KI bei Administrativen Aufgaben, in der Lehre und beim Lernen**. Hierbei ist wichtig zu betonen, dass im Folgenden speziell im schulichen Bildungskontext KI-Technologien beschrieben werden, welche teilweise bereits heute in verschiedenen Ländern Anwendung finden, während manche noch weit von einer Implementierung im Alltag entfernt sind.

3.1.1 KI bei Administrativen Aufgaben

Um die Schulorganisation, sprich das Management und die Planung von Schulen und Hochschulen, voranzutreiben, kann sich das Erfassen und Analysieren von schulischen Datenbeständen als wichtig erweisen (**Educational Data Mining**). Diese schulichen Datenbestände basieren auf Verwaltungsdaten wie Ausfall- und Vertretungszeiten, Lernzeiten und Lerndauer, Leistungsniveaus und Anmeldungen für Prüfungen und Wiederholungsprüfungen. Zusätzlich werden MOOC-Daten, die aus massiven Online-Kursen (Massive Open Online Courses) wie Coursera for Campus stammen, analysiert, um das Lernverhalten der Lernenden in digitalen Kursen, ihre Fortschritte und Teilnahmen zu verstehen. Ebenso liefern LMS-Daten (Learning Management System), die aus Plattformen wie Moodle gewonnen werden, Informationen zur Nutzung von Lernmaterialien, zu Aktivitäten in Diskussionsforen und zur Zusammenarbeit der Lernenden in digitalen Lernumgebungen (Kramer, 2019, S. 12). Educational Data Mining, unterstützt durch KI-Funktionen, ermöglicht eine effizientere Analyse, Diagnostik und Prognose der schulischen Datenbestände. Es erkennt schneller und automatisiert Muster, deckt wertvolle Erkenntnisse auf und passt sich kontinuierlich an, indem es aus den Daten lernt und sich weiterentwickelt (Barilla, 2024). Die gewonnen Daten können die Schulverwaltungen bei administrativ-planerischen Aufgaben datenbasiert unterstützen, etwa bei Wirtschaftlichkeitsanalysen, Schulinspektionen, Zielvereinbarungsgesprächen, Schulentwicklungsberatungen oder Empfehlungen für Schulleitungen, Schüler und Lehrkräfte (Schmid et al., 2021, S. 16). Zudem ermöglichen intelligente Analysen und Prognosen die Vorhersage der Lernendenanzahl und eine fundierte Planung des entsprechenden Lehrkräftebedarfs, um eine optimale Ressourcenzuweisung sicherzustellen. Ein weiteres Beispiel ist die “Unterstützung von Lehrkräften bei der Verwaltung und Vorstrukturierung von Elternabenden und -gesprächen, indem z.B Einladungen und Elternbriefe vorformuliert

werden (z.B. basierend auf aktuellen Leistungsdaten der Schülerinnen und Schüler)" (Schmid et al., 2023, S. 50).

Ein weiteres Feld, das sich mit der Analyse von Bildungsdaten beschäftigt, ist das **Learning Analytics**. Hierunter versteht man die "kontinuierliche Messung und Sammlung, Analyse und Berichterstattung von Daten über Lernende und ihre Aktivitäten zum besseren Verständnis und zur Optimierung des Lernens in den jeweiligen digitalen Lernumgebungen" (Schmid et al., 2021, S. 12). Auf Grundlage der schulischen Datenbestände, wie sie auch im Educational Data Mining genutzt werden, können "einerseits didaktische Interventionsmaßnahmen und Anreize, andererseits aber auch maßgeschneiderte, personalisierte Lernwege, Hilfestellungen und Lernziele generiert werden" (Schmid et al., 2021, S. 12). So ist die Anwendung von Learning Analytics in Bildungseinrichtungen in der Lage, die Qualität der Dienstleistungen zu erhöhen und gleichzeitig die Noten und Leistungen der Schüler zu verbessern (Hwang et al., 2017). Dies wird durch den Einsatz von KI unterstützt, die die Effizienz und Personalisierung der Lernprozesse steigern kann (Theodotou, 2024). Im Gegensatz zum Educational Data Mining, das stärker forschungsorientiert ist, liegt der Fokus des damit eng verwandten Feldes der Learning Analytics stärker auf dem individuellen Lernenden (Schmid et al., 2021, S. 13).

Eine relevante Ebene der Learning Analytics sind die **Predictive Learning Analytics**, bei denen laut Ruhland et al. (2023, S. 51) „Voraussagen über denkbare zukünftige Lernzustände und Lernverläufe" getroffen werden können. Diese Voraussagen ermöglichen es Lehrenden, erwartete Schulerfolge zu prognostizieren, „ihren Unterricht frühzeitig anzupassen und so negative Entwicklungen vorzubeugen" (Ruhland et al., 2023, S. 51). Ein Anwendungsbeispiel, an dem momentan umfangreich geforscht wird, ist das KI-Tool zum Vorhersagen, welche Schüler am meisten gefährdet sind, die Schule früh zu verlassen. So hat die Regierung in Argentinien ein solches System eingeführt, das Daten wie Anwesenheit, Noten und sozioökonomische Faktoren analysiert, um Schüler zu identifizieren, die ein erhöhtes Risiko für einen Schulabbruch aufweisen, sodass gezielte Unterstützungsmaßnahmen ergriffen werden können (Páez, 2023).

Eine weitere wesentliche administrative Aufgabe im Bildungswesen ist die Bewertung von Prüfungen und Hausaufgaben. Obwohl diese Aufgabe traditionell von Lehrenden durchgeführt wird, handelt es sich dabei um eine Routinearbeit, die nicht direkt zur Wissensvermittlung beiträgt. **KI-gestützte Systeme zur automatisierten Korrektur und Bewertung** (Engl. Automated Assessment/Grading) können diese Aufgabe effizienter gestalten, indem sie vor allem in Zeiten des Lehrermangels Entlastung bieten und den Lehrenden mehr Zeit für die direkte Interaktion mit den Lernenden ermöglichen. Dabei basiert der Ansatz darauf, dass "ein Lernalgorithmus die richtige Lösung einer Aufgabe auf Grund markierter Trainingsdaten verlässlich identifizieren kann, und dem Lernenden eine entsprechende Rückmeldung bzw. Bewertung geben kann" (Schmid et al., 2021, S. 12). Es wird bereits in vielen Forschungsprojekten versucht, den "Einsatz der KI für die automatische Testgenerierung und -bewertung zu untersuchen" (Schmid et al., 2021, S. 30). Eine Studie des mmb Instituts hat ergeben,

dass 78% der Befragten die Entlastung der Lehrkräfte durch automatisierte Bewertungen gut finden und 78% dies für technisch realisierbar halten. Allerdings halten nur 57% der Befragten diese automatisierten Systeme für wünschenswert (Schmid et al., 2021, S. 30). Einer der Gründe hierfür ist die unbeabsichtigte Folge, dass "Tests mit hohen Anforderungen zunehmend durch häufige formative Beurteilungen mit niedrigen Anforderungen ersetzt werden, da der Aufwand und die Kosten für die Beurteilung abnehmen" (Schmid et al., 2021, S. 30). Ein weiterer Grund ist möglicherweise die Auffasung, dass solch eine Arbeit bzw. Aufgabe menschliches Urteilsvermögen und Fachwissen erfordert, was die Überwachung und das Eingreifen des Menschen in KI-gestützte Bildungssysteme erforderlich macht (Chaushi et al., 2024, S. 52).

In der heutigen Online-Lernumgebung stehen Lehrenden Programme mit KI-Funktionen wie Turnitin zur Verfügung, die sie bei der Benotung von studentischen Arbeiten unterstützen können. Hier wird bspw. eine Funktion angeboten, bei dem automatisch ähnliche Antworten zusammengefasst werden, "wodurch unnötiger Aufwand auf ein Minimum reduziert wird und ein qualitativ hochwertigeres, konsistenteres Feedback ermöglicht wird" (Turnitin, LLC., 2024).

Programme wie Turnitin, PaperRater und Grammarly bieten zudem **Plagiatsprüfungen** an, indem sie studentische Arbeiten auf Ähnlichkeiten mit bestehenden Quellen überprüfen, bieten aber auch eine **Grammatik- und Rechtschreibunterstützung**. Mithilfe dieser KI-Funktionen können Lehrende effizienter arbeiten, indem sie Fehler in Texten schneller identifizieren und beheben sowie potenzielle Plagiate zuverlässiger erkennen (Grammarly Inc., 2024).

Auch die Verwendung von **Chatbots** können zur Beantwortung von Wissensfragen aller Art vielfältig genutzt werden. Der Chatbot fungiert als textbasierter, dialogfähiger KI-Agent, der mittels Verarbeitung natürlicher Sprache (Engl. Natural Language Processing), Machine Learning und manchmal auch generativer KI kontextbezogene und flexible Antworten erstellt und aus Interaktionen lernt. Studierende haben durch den Einsatz von Chatbots rund um die Uhr einfachen Zugang zu Kursinformationen, was bedeutet, dass sie auch außerhalb der üblichen Arbeitszeiten Unterstützung bei administrativen Anfragen erhalten können. Die Chatbots dienen als erste Anlaufstelle für Fragen zu Abgabeterminen, Aufgaben und anderen Einzelheiten des Kurses. Dies führt nicht nur dazu, dass die Studierenden sofort Antworten auf ihre Fragen erhalten, sondern entlastet auch das Lehrpersonal, da wiederholte Anfragen vermieden werden (Stanford Education, 2021). Dass Chatbots ein großes Potenzial bieten, hat die scieneers GmbH in Zusammenarbeit mit der RWTH Aachen bewiesen, indem sie einen Chatbot entwickelt haben, der an den Lehrstühlen für Mathematik, Bauingenieurwesen und Informatik mithilfe eines speziell entwickelten Plugins in Moodle integriert wurde (Krieg, 2023). Dieser bietet "einen nahtlosen Zugang zu Vorlesungsinhalten für Studierende und Mitarbeitende direkt in der Lernplattform Moodle. Somit können themenspezifische Fragen mit Inhalten aus einer Vorlesung beantwortet werden. Eine

zusätzliche Quellenangabe erlaubt außerdem das schnelle Finden der relevanten Vorlesungsinhalte in den Unterlagen" (Höllmer, 2023).

Abbildung 1: Vorlage eines möglichen Chatbots für Bildungseinrichtungen

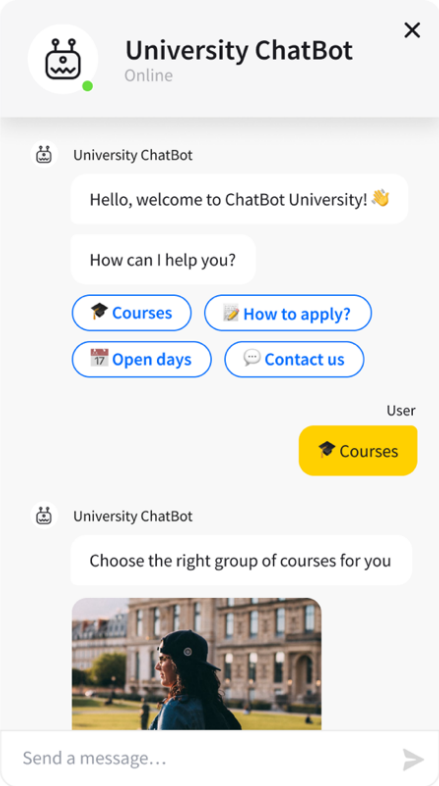


Abbildung von Text, Inc. (o. J.)

3.1.2 KI in der Lehre

Adaptive Lernplattformen, auch als Adaptive Lernsysteme bekannt, sind eine Schlüsseltechnologie im Bereich der KI-unterstützten Lehre. Diese Plattformen passen Lernaktivitäten und -inhalte in Echtzeit an die individuellen Bedürfnisse, Präferenzen und das Lernverhalten der Lernenden an. Ermöglicht wird das durch Verfahren des Machine Learnings, Learning Analytics und des Educational Data Minings. Dadurch können Lehrende die Lernfortschritte präzise überwachen, während das System Lernenden bspw. Empfehlungen gibt, "ein bestimmtes Kapitel zu wiederholen oder auch zu überspringen" (Fraunhofer IOSB, 2023). Besonders in den MINT-Fächern und beim Spracherwerb können adaptive Selbstlernangebote unterrichtsbegleitend zur individuellen Förderung eingesetzt werden (Schmid et al., 2021, S. 17). Ein Beispiel für eine solche Plattform ist Knewton, die den Ansatz des adaptiven Lernens folgendermaßen beschreibt: "Knewton adaptive learning combines a data-driven approach to course materials with a 360° view of an individual student to create optimized learning paths that help all learners reach their highest potential" (John Wiley & Sons, Inc., 2024).

Während Adaptive Lernplattformen Inhalte und Aktivitäten an das individuelle Lernverhalten anpassen, bieten **intelligente Tutorensysteme (ITS)** eine persönlichere, tiefere Unterstützung, indem sie den Lernenden gezielt durch den Lernprozess führen, ähnlich wie ein menschlicher Tutor. Ein etabliertes Beispiel hierfür ist Khanmigo, das als „Tutor für Lernende“ und „Assistent für Lehrkräfte“ dient (Avid Center, 2023). Lernende profitieren dadurch, dass ein KI-Tool wie Khanmigo sie dazu leitet, die Antwort selbst zu finden, statt dass sie durch ein KI-Tool wie ChatGPT direkt eine finale Antwort bekommen. Ermöglicht wird das durch Funktionen wie Schreiben und Debattieren unter Anleitung mit sofortigem Feedback, aber auch durch Echtzeit-Feedback beim programmieren (Khan Academy, 2024a). Für Lehrende beschleunigt Khanmigo zahlreiche Aufgaben. Wie in Abbildung 2 zu sehen ist umfasst dies die Erstellung von Unterrichtsplänen, Quizfragen, schnellen Entwürfen für E-Mails an Familien, die Gruppierung von Schülern sowie die Ausarbeitung von Einstiegsfragen, Abschlussfragen und Bewertungsrichtlinien (Khan Academy, 2024b). Des Weiteren kann das KI-Tool den Lehrenden dabei helfen, mit den neuesten Fortschritten in ihrem Lehrbereich Schritt zu halten. Ein Physiklehrer könnte bspw. mit Hilfe des Tools über die neusten Durchbrüche in der Astrologie informiert werden, welcher dieser an seine Schüler weitergibt. Eine Alternative zu Khanmigo ist der KI-Assistenz fobizz.

Abbildung 2: Bündelung verschiedener KI-Funktionen in eine Anwendung

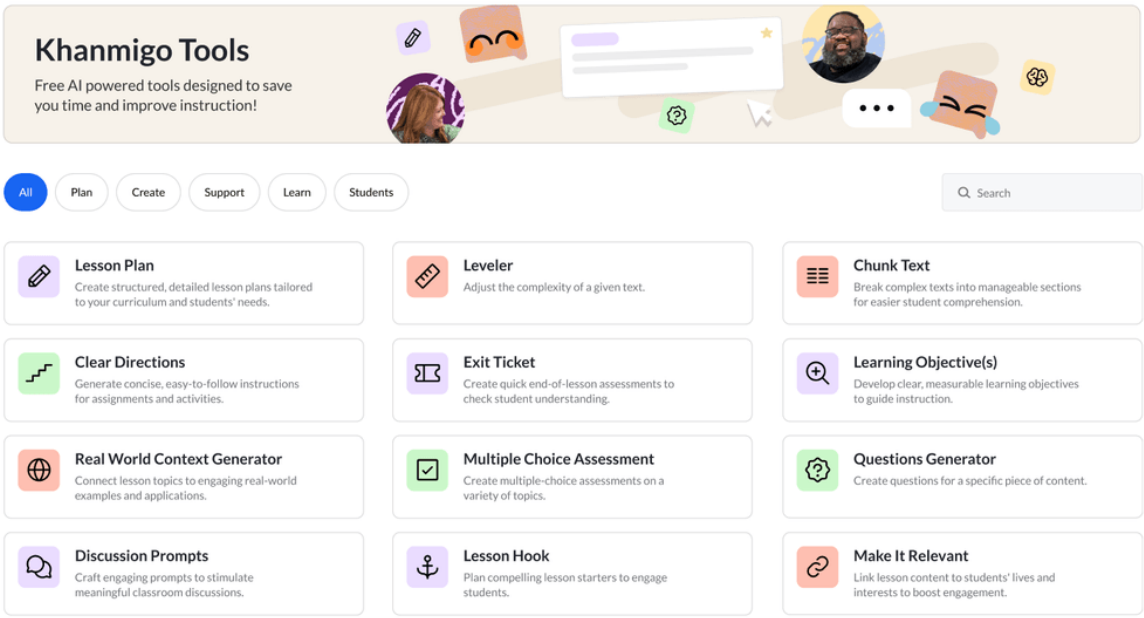


Abbildung von Khan Academy, 2024b

Intelligente Tutorensysteme wie Khanmigo zeigen in **Flipped Classroom Settings**, wozu sie noch in der Lage sind. In diesem Lernsetting werden traditionelle Unterrichtsmethoden umgekehrt: Schüler studieren Lernmaterialien wie Videos oder Texte zu Hause, während die Unterrichtszeit für kollaboratives und interaktives Lernen genutzt wird. KI spielt hierbei eine Schlüssrolle, indem Anwendungen wie Khanmigo die notwendigen für die Schüler personalisierten Inhalte bereitstellen und ihnen von zu Hause aus gezieltes Feedback sowie Unterstützung bieten können. Da die notwendige

KI-gestützte Vorarbeit bereits zu Hause geleistet wurde, sind Lehrkräfte im Klassenzimmer entlastet, da sie weniger Zeit mit der Vermittlung von Lehrinhalten verbringen müssen. Sie können nun die Lehre effizienter und flexibler gestalten, indem sie sich mehr auf die persönliche Betreuung und die Vertiefung des Verständnisses konzentrieren (Mollick, 2024).

Lernspiele sind in der Bildung längst anerkannt, da sie das Lernen auf spielerische Weise fördern und die Motivation der Lernenden steigern (Staube et al., 2012). Mithilfe von KI werden diese Lernspiele noch effektiver, wie es z.B. die Plattform Kahoot! zeigt. Bei Kahoot! handelt es sich um eine „spielbasierte Lernplattform, die das Erstellen, Teilen und Spielen von Quizfragen in wenigen Minuten ermöglicht“ (Kahoot!, o. J.). Statt wie früher jede Frage manuell zu erstellen oder ein fertiges Quiz zu wählen, bietet Kahoot! nun die Möglichkeit, mithilfe des integrierten OpenAI-Modells GPT-4 automatisch passende Fragen und Quizze zu generieren. So lassen sich vollständige Quizze durch das Eintippen der gewünschten Thematik, das Hochladen einer PDF-Datei oder dem Einfügen einer URL erstellen (Kahoot!, 2024). Somit sparen Lehrende viel Zeit und werden zudem durch die eigenständige Erstellung von Lernmaterialien entlastet.

3.1.3 KI beim Lernen

Spätestens seit OpenAI im November 2022 ChatGPT vorgestellt hat, hat KI den Alltag insbesondere von Schülern und Studenten maßgeblich verändert. **Generative KI-Sprachmodelle** wie etwa ChatGPT, Gemini oder Claude werden vielfältig genutzt, etwa zur Unterstützung bei Hausaufgaben, beim Verfassen von Essays, dem Beantworten von Multiple-Choice-Fragen oder sogar beim Schummeln in Klausuren. So standen bspw. einige Schüler in Hamburg unter Verdacht, in den Abiturprüfungen 2023 mit Hilfe von KI, wie etwa ChatGPT, geschummelt zu haben (Norddeutscher Rundfunk, 2023).

Somit hat die Nutzung von Programmen wie ChatGPT im Bildungswesen, die zwar viele Chancen bietet, Unsicherheit und Skepsis geweckt. Eine von der Civey GmbH im Auftrag der Vodafone Stiftung durchgeführte Studie, bei der 5000 Bundesbürger ab 18 Jahren und rund 500 Eltern von Kindern bis 18 Jahren befragt wurden, zeigt zwar, dass 54% der Deutschen glauben, KI werde den Schulunterricht maßgeblich verändern. Gleichzeitig zeigen sich 57% der Befragten Programmen wie ChatGPT gegenüber skeptisch und sehen KI in Schulen eher mit einer Gefahr als mit einer Chance (Vodafone Stiftung Deutschland gGmbH, 2023, S. 5).

Trotz der damit verbunden Bedenken und Gefahr setzt das nordrhein-westfälische Schulministerium auf einen Leitfaden für den Umgang mit KI anstatt eines Verbots von ChatGPT an Schulen in NRW. Aus diesem Leitfaden geht hervor, dass Schüler, die zur Erledigung ihrer Aufgaben KI nutzen, dies angeben müssen, um eine Wertung als Täuschversuch und der Verwendung unzulässiger Hilfsmittel zu vermeiden (Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen, 2023, S. 8).

Im Bereich der „Generativen KI“, zu der eben auch textgenerierende Systeme wie ChatGPT gehören, haben sich ebenfalls **material- und bildgenerierende Systeme** etabliert. Durch kurze Prompts lassen sich innerhalb weniger Sekunden Bild-, Video- oder Präsentationsmaterialien erstellen. KI-Präsentationsersteller wie Slidesgo, Canva oder Simplified ermöglichen es, Präsentationen in erstaunlicher Qualität und Geschwindigkeit zu generieren, die gut als Ausgangspunkt und Inspiration für weitere Ergänzungen und Anpassungen dienen können. Mithilfe von KI-Bildgeneratoren wie DALL-E, Stable Diffusion oder starryai können Lernende ihre Wörter und Sätze in eindrucksvolle Fotos und Kunstwerke verwandeln, um bspw. ihre Präsentationen mit individuell erzeugten Bildern zu erweitern. Im Kunstunterricht können KI-Bildgeneratoren ebenfalls Anwendung finden, indem berühmte Künstler nachgeahmt oder Vorlagen für eigene Werke geschaffen werden. Die wachsende Popularität der KI-Generatoren ist unter anderem darauf zurückzuführen, dass viele von ihnen als App oder Browseranwendung schnell und einfach nach der Einrichtung eines kostenlosen Accounts nutzbar sind (Schmid et al., 2023, S. 28). Die Humanwissenschaftliche Fakultät der Universität Potsdam hat hierfür eine Übersicht mit verschiedensten KI-Anwendungen bereitgestellt, die den Unterricht bereichern könnten (Humanwissenschaftilche Fakultät Universität Potsdam, 2023). Hierbei ist zu erwähnen, dass material- und bildgenerierende Systeme natürlich auch in der Lehre genutzt werden können.

Neben den bereits beschriebenen generativen KI-Modellen haben sich mittlerweile auch spezialisierte **KI-Tools für wissenschaftliche Arbeiten** etabliert. Adressiert werden hierbei insbesondere die Schwächen von allgemeinen Sprachmodellen im akademischen Kontext. Während bspw. ChatGPT bei Zitaten und Quellenangaben häufig unzuverlässig ist und sogar teilweise fiktive Quellen generiert, bieten spezialisierte Werkzeuge eine deutlich präzisere Unterstützung für wissenschaftliche Arbeiten. Der entscheidende Unterschied liegt zum einen darin, dass die spezialisierten Werkzeuge generative Sprachmodelle mit direkter Anbindung an wissenschaftliche Datenbanken kombinieren. Dadurch entsteht eine konsolidierte und vertrauenswürdigere Datenbasis im Vergleich zu generellen Sprachmodellen wie ChatGPT. Zum anderen nutzen die spezialisierten Systeme integrierte Prompt-Engineering-Methoden, die ihre Ausgaben automatisch für den akademischen Kontext optimieren, statt das der Nutzer das manuell machen muss. Ein Beispiel ist Perplexity.ai, welches Internetsuche und Chatbot-Funktionalität mit einer kuratierten Datengrundlage kombiniert. Alle Aussagen in den Antworten sind mit einem Quellenhinweis versehen. Weitere Beispiele sind Semantic Scholar, Elicit und Research Rabbit (Sperl, 2024).

KI-Funktionen in **adaptiven Lernplattformen und intelligenten Tutorensystemen** haben nicht nur neue Wege der Lehre ermöglicht, sondern auch den Lernprozess der Lernenden verbessert. Angewandte Technologien wie Learning Analytics und Educational Data Mining ermöglichen eine personalisierte Förderung, die auf die individuellen Stärken und Schwächen der Lernenden zugeschnitten ist. Durch das kontinuierliche Erheben von Leistungsdaten wird das Schwierigkeitsniveau von Übungsaufgaben individuell angepasst und es wird gezielt an Kompetenzrückständen gearbeitet. Adaptive

Lernplattformen wie Knewton, bettermarks und Century Tech haben diese Technologien integriert, um automatisch individualisierte Lernangebote und Empfehlungen bereitzustellen (Daniela Hau, o. J.; Schmid et al., 2023, S. 44). Zu den intelligenten Tutorensystemen zählt unter anderem das KI-Nachhilfe-Programm Alena, das jederzeit und ortsunabhängig nicht nur Antworten bietet, sondern auch die Lösungswege erklärt. Lernende können Arbeitsblätter oder Fotos von Rechenwegen hochladen, um mithilfe von Alena Fehler zu erkennen und Erklärungen zu einzelnen Rechenschritten zu erhalten. Alena verspricht zudem, eine unbegrenzte Anzahl an zielgerichteten und prüfungsrelevanten Übungsaufgaben zu generieren, die an den bundeslandspezifischen Lernplan angepasst sind (TutorSpace GmbH, 2024). Technisch basiert es auf datenschutzkonformen Zugängen zu ChatGPT 4.0 in Kombination mit dem Mathematik-Plugin Wolfram Alpha (Bronner, 2024).

Abbildung 3: Lehrer Chatbot Alena

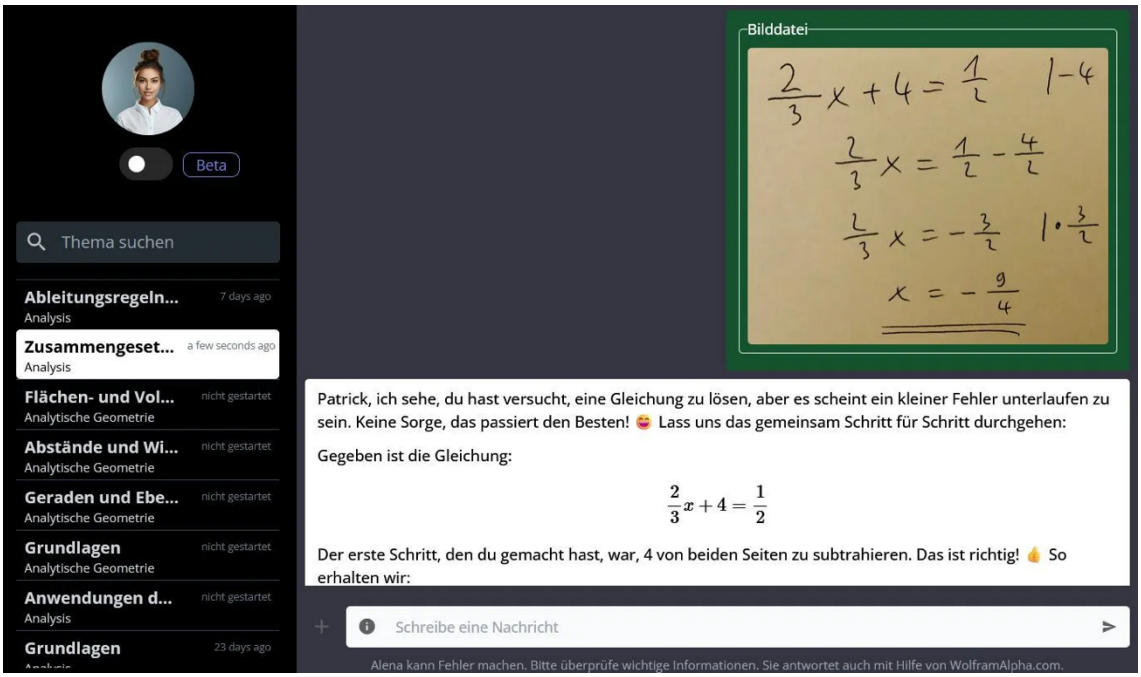


Abbildung von Bronner (2024)

3.2 Notwendigkeit der Erklärbarkeit von KI im Bildungswesen

Seit dem KI-gestützte Systeme für eine Reihe von Funktionen wie personalisiertes Lernen, die Bewertung sowie die Unterstützung von Lernenden eingesetzt werden, gewinnt das Forschungsgebiet der Erklärbaren KI im Bildungskontext zunehmend an Bedeutung. Verschiedene Faktoren tragen zur Bedeutung und Notwendigkeit von Erklärbarer KI im Bildungsbereich bei, die im Folgenden dargelegt werden.

3.2.1 Erklärbarkeit zur Entlarvung von Verzerrungen (Bias)

„Der kluge Hans“, ein Pferd, das scheinbar mathematische Fragen beantworten konnte, galt in den 1900er Jahren als wissenschaftliche Sensation. Es wurde jedoch später festgestellt, dass Hans nicht wirklich die Mathematik beherrschte, sondern in 90% der

Fälle die korrekte Antwort aus der unbewussten Reaktion des Fragestellers interpretierte (Baranzke, 2021). Sogenannte Verzerrungen (Engl. Bias) sind auch in modernen KI-Systemen zu beobachten. Ähnlich wie beim Phänomen des „klugen Hans“ treten Verzerrungen in modernen KI-Systemen auf, wenn Algorithmen auf Basis irreführender und oberflächlicher Merkmale lernen, ohne sinnvolle Kausalzusammenhänge zu erkennen (Samek & Müller, 2019, S. 3).

In KI-gestützten Bildungssystemen kann das zu ungleichen und unfairen Ergebnissen für Lernende führen, wie etwa ungerechte Bewertungen von Klausuren. Gründe für Verzerrungen in solchen KI-Systemen sind unter anderem unzureichend diverse Trainingsdaten oder vorhandene Vorurteile in den Daten (Chaushi et al., 2024, S. 54).

Ein Beispiel für die Auswirkungen von Verzerrungen in einem Algorithmus ist der A-Level-Bewertungsskandal, bei dem im Jahr 2020 in Großbritannien die A-Level-Prüfungen (vergleichbar mit dem Abitur) aufgrund der COVID-19-Pandemie abgesagt wurden. Die Benotung erfolgte stattdessen durch ein KI-gestütztes Bewertungssystem, welches Schüler basierend auf früheren Leistungen bewerte. Wer auf einer Schule war, deren Abschlussjahrgänge in der Vergangenheit schwächer waren, dessen Prüfungsergebnisse wurden herabgestuft, was somit zu einer Benachteiligung von Schülern aus sozial benachteiligten Gebieten führte. Diese Verzerrungen waren nicht offensichtlich, bis die Ergebnisse veröffentlicht wurden. Folglich kam es zu Protesten und die Benotungen durch den Algorithmus wurden zurückgenommen (Hau, 2023; Jochum, 2020; Kolkman, 2020).

Erklärbarkeit kann somit dabei helfen, Verzerrungen in KI-Modellen bzw. in den Daten frühzeitig zu erkennen und zu beheben, und dadurch ethische Implikationen zu vermeiden (Samek & Müller, 2019, S. 8).

3.2.2 Erklärbarkeit zur kontinuierlichen Verbesserung

Potenzielle Verzerrungen (Bias) sind eine von vielen möglichen Schwächen eines KI-Systems, die mithilfe von Erklärbarkeit identifiziert und verstanden werden können. Modelle, die ihre Schwächen umfassend offenlegen und ihre Ergebnisse verständlich erklären, lassen sich in der Regel leichter verbessern. Durch diese Transparenz werden häufig potenzielle, noch unbekannte Schwachstellen und Fehler schneller erkannt und behoben (Andrae, 2021, S. 18).

3.2.3 Erklärbarkeit zur Stärkung von Vertrauen und Akzeptanz der Ergebnisse

Schlagzeilen wie das A-Level-Bewertungsskandal und rassistische Verzerrungen in KI-generierten Bildern tragen nicht dazu bei, dass ohnehin bereits bestehende Misstrauen der Bevölkerung gegenüber den getroffenen Entscheidungen von KI-Systemen zu verringern (Nicoletti & Bass, 2023). So hat eine Studie der Firma KPMG im Jahr 2023 ergeben, dass 61% der Befragten Bedenken haben, KI-Systemen zu vertrauen und 67% berichten von einer geringen bis mittleren Akzeptanz von KI (KPMG, 2023).

Um das Vertrauen in KI-Systeme zu stärken und deren Akzeptanz zu fördern, ist es unerlässlich, algorithmische Entscheidungen den Empfängern transparent bzw. erklärend offenzulegen und zu überprüfen. Besonders im Bereich des Bildungswesens, wo der Einsatz von KI noch relativ jung ist und viele Beteiligte Unsicherheiten und Bedenken haben, ist Vertrauen und somit eine Erklärbarkeit unverzichtbar. Lehrkräfte, Schüler und Eltern müssen die Funktionsweise der eingesetzten KI-Modelle verstehen können, um Vertrauen zu entwickeln und sicherzustellen, dass die Lernprozesse fair, nachvollziehbar und individuell angepasst sind. Erst durch eine klare und verständliche Erklärung der KI-Funktionsweise kann das notwendige Vertrauen geschaffen werden, um die Akzeptanz dieser Technologien im Bildungswesen zu fördern (Chaushi et al., 2024, S. 52; Ribeiro et al., 2016; Samek & Müller, 2019, S. 4)

3.2.4 Erklärbarkeit als Grundlage für die Erarbeitung neuer Erkenntnisse

KI-Systeme können Muster und Zusammenhänge in Daten identifizieren, die für Menschen oft schwer zugänglich sind. Beispiele sind etwa Spielstrategien oder unbekannte Assoziationen zwischen Variablen wie Genen und Krankheiten. So betonen Samek und Müller (2019, S. 5), dass in manchen Szenarien das Identifizieren von Mustern bedeutender ist als die reine Vorhersage selbst, da sie zu neuen Erkenntnissen führt. Für diese Identifizierung neuer Muster ist es jedoch essenziell, erklären und interpretieren zu können, welche Merkmale vom KI-System zur Vorhersage genutzt werden (Samek & Müller, 2019, S. 5).

Besonders im Bildungswesen könnte eine Erklärbare KI bisher verborgene Erkenntnisse über Lernprozesse und -hindernisse aufdecken. Dadurch könnte man besser nachvollziehen, wie bestimmte Lernstrategien wirken oder welche Faktoren am meisten den Lernerfolg beeinflussen.

3.2.5 Erklärbarkeit als Teil der Gesetzgebung

KI-Systeme unterliegen rechtlichen Bestimmungen und Regulierungen und agieren somit keineswegs in einem rechtsfreien Raum. Diese Rechtsvorschriften gelten bereits auf nationaler, internationaler und insbesondere auf europäischer Ebene und beeinflussen die Entwicklung, Einführung und Nutzung von KI-Systemen (Moser, 2023, S. 65).

EU-weit greifen bereits Rechtsvorschriften zur Förderung einer vertrauenswürdigeren KI ein, darunter Grundrechte (wie Nichtdiskriminierung, Datenschutz und Schutz der Privatsphäre), Verbraucherschutz und Produktsicherheit und -haftung (HEG-KI, 2019, S. 8).

Des Weiteren wurde 2016 die DSGVO erlassen, um personenbezogene Daten und die Privätsphäre der Anwender zu schützen. Die Artikel 12, 13, 14 und 22 der DSGVO sind für diese Arbeit von Bedeutung, da sie Informationspflichten und somit konkrete Bestimmungen zur Transparenz von KI-Systemen beim Bestehen von automatisierten Entscheidungsfindungen und Profiling vorschreiben. Gemäß Art. 12 Abs. 1 DSGVO

müssen Unternehmen, die personenbezogene Daten verarbeiten, Betroffenen Auskunft darüber in “präziser, transparenter, verständlicher und leicht zugänglicher Form“ zur Verfügung stellen (DSGVO, 2016, Art. 12 Abs. 1; Weik, 2021). Art. 13 Abs. 2 DSGVO verpflichtet Unternehmen, Betroffene nicht nur über die „Existenz des Verarbeitungsvorgangs und seine Zwecke“ zu unterrichten, sondern sie darüber hinaus „über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung“ aufzuklären (DSGVO, 2016, Art. 13 Abs. 2 lit f; Moser, 2023, S. 50; Weik, 2021). Art. 14 DSGVO fungiert als Auffangtatbestand für Art. 13 DSGVO. Es legt fest, in welchem Umfang der Verantwortliche Informationen bereitstellen muss, wenn die personenbezogenen Daten (im Gegensatz zu Art. 13 DSGVO) nicht direkt bei der betroffenen Person erhoben werden (Moser, 2023, S. 65). Der Art. 22 DSGVO befasst sich speziell mit Entscheidungen, die „ausschließlich auf automatisierten Verarbeitungen beruhen, einschließlich Profiling“, und die rechtliche Wirkungen für die betroffenen Personen haben oder sie in ähnlicher Weise erheblich beeinträchtigen (DSGVO, 2016, Art. 22 Abs. 1; Moser, 2023, S. 60). Der Ausdruck „ausschließlich“ impliziert hierbei, dass grundsätzlich alle jene Fälle zu verstehen sind, bei denen KI-Systeme Entscheidungen ohne menschliches Eingreifen treffen (Engl. „human in the loop“) (Moser, 2023, S. 61). Auch in Fällen, wo ein Mensch eine automatisierte Vorentscheidung formal bestätigt, gilt dies als rein automatisierte Entscheidung, da keine inhaltliche Einflussnahme stattfindet (Moser, 2023, S. 62). So hat bspw. eine Person das Recht, bei einer Kreditablehnung durch ein KI-System einer Bank die Gründe für diese Entscheidung zu erfahren (Samek & Müller, 2019, S. 5).

Somit regeln die genannten Artikel der DSGVO „zahlreiche Mitteilungs- und Offenlegungspflichten, um den Grundsatz einer fairen und transparenten Verarbeitung personenbezogener Daten zu fördern“, umfassen aber auch „eine Kennzeichnungspflicht für das Bestehen einer vollständigen automatisierten Entscheidungsfindung“ (Moser, 2023, S. 78).

Jedoch gilt die DSGVO nur für die Verarbeitung personenbezogener Daten und findet daher keine Anwendung, wenn Menschen von durch KI-Systemen generierten Entscheidungen betroffen sind, die ohne solche Daten auskommen. Des Weiteren sind die Regelungen zur Erklärbarkeit begrenzt. Sie bieten lediglich Ansätze zur allgemeinen Erklärung der Entscheidungslogik vollautomatisierter KI-Systeme, jedoch keine Möglichkeit für Betroffene, eine spezifische Erklärung für individuelle Entscheidungen zu erhalten (Moser, 2023, S. 73).

Die DSGVO allein genügt somit nicht, um den Herausforderungen gerecht zu werden und der Menge an Risiken, die KI-Systeme mit sich bringen, entgegenzuwirken. Deshalb sind zusätzliche, gezielte regulatorische Maßnahmen erforderlich, um den Schutz des Einzelnen und der Gesellschaft angemessen zu gewährleisten.

So hat die Europäische Union das Gesetzespaket der KI-Verordnung (kurz. KI-VO, engl. EU AI Act, kurz AIA) in die Wege geleitet. Dieser zielt darauf ab, einen rechtlichen Rahmen in Zusammenhang mit KI zu schaffen, den Schutz der Grundrechte zu gewährleisten und mögliche Verletzungen dieser durch KI hinzuhalten und die

Entwicklung einer vertrauenswürdigen KI zu fördern (Moser, 2023, S. 79). Die KI-VO trat am 01.08.2024 in Kraft, die Gesetztesvorgaben selbst werden jedoch schrittweise eingeführt (Generaldirektion Kommunikation, 2024). Die Verordnung gilt für Anbieter, die KI-Systeme in der EU in Verkehr bringen oder in Betrieb nehmen, unabhängig davon, ob diese Anbieter in der EU niedergelassen sind oder nicht (Langer, Baum, et al., 2021, S. 166).

Die KI-VO konzentriert sich nicht auf spezifische KI-Technologien oder deren Funktionsweisen, sondern verfolgt einen risikobasierten Ansatz, indem Anwendungsfälle von KI-Systemen reguliert werden, die Gefahren für die Grundrechte und die Sicherheit von Personen darstellen (Verch, 2024, S. 6). Der risikobasierte Ansatz unterteilt KI-Systeme in vier Risikokategorien – unannehmbar/inakzeptabel, hoch, begrenzt und minimal. Jede Kategorie hat ihre eigenen Vorschriften, wobei das "Inverkehrbringen, die Inbetriebnahme oder die Verwendung von KI-Systemen, die ein unannehmbares Risiko darstellen", verboten ist (Heynike, 2024). Es gilt, je höher das ausgehende Risiko eines KI-Systems für einen Betroffenen ist, desto höher ist die Risikoeinstufung und die Strenge der Vorschriften (Europäische Kommission, 2024; Moser, 2023, S. 80).

Welche KI-Systeme in die Kategorie der Hochrisiko-KI-Systeme fallen, wird zum einen in Art. 6 KI-VO beschrieben und zum anderen durch den Anh. III zu Art. 6 Abs. 2 KI-VO, in welcher konkrete Anwendungsfälle aufgeführt sind, die ein KI-System als "risikoreich" qualifizieren würden. So werden in Anh. III zu Art. 6 Abs. 2 KI-VO insgesamt acht allgemeine Anwendungsbereiche aufgeführt, in denen KI-Systeme als risikoreich eingestuft werden können.

Ein Anwendungsbereich davon ist die "allgemeine und berufliche Bildung" (KI-VO, 2024, Anh. III Abs. 3). Allerdings gilt nicht jede in diesen Bereich fallende KI als Hochrisiko-KI-System (Moser, 2023, S. 87). Anh. III Abs. 3 zu Art. 6 Abs. 2 KI-VO führt vier verschiedene Anwendungsfälle aus, die konkret im Bildungswesen eine KI risikoreich machen. Es werden KI-Systeme in der Bildung als Hochrisiko-KI-Systeme eingestuft, wenn sie "für die Auswahl und Zulassung von Bewerber*innen sowie für die Bewertung von Lernergebnissen und des Bildungsniveaus verwendet werden". Ebenfalls risikoreich sind KI-Systeme in der Bildung, "die bei Prüfungen zur Aufdeckung von Täuschungsversuchen verwendet werden" (Verch, 2024, S. 9).

Für alle als hochriskant eingestuften KI-Systeme, einschließlich die genannten Anwendungsfälle im Bereich der Bildung, gelten strenge organisatorische und technische Vorgaben, darunter hohe Ansprüche an die Qualität der Datensätze, mit denen die KI trainiert wurde (Art. 10 KI-VO), eine verständliche Dokumentation (Art. 11 KI-VO) sowie technische und organisatorische Maßnahmen zur Förderung der Cybersicherheit und Robustheit von KI-Systemen (Art. 15 KI-VO) (Verch, 2024, S. 10).

Des Weiteren verpflichtet die KI-VO zur Transparenz und Bereitstellung von Informationen für Nutzer, um so das Vertrauen der Menschen KI-Systeme zu stärken.

Im Speziellen schreibt Art. 13 KI-VO einen Grundsatz zur Transparenz vor, insbesondere für Hochrisiko-KI-Systeme.

So sind gemäß Art. 13 Abs. 1 KI-VO KI-Systeme so zu konzipieren und zu entwickeln, “dass ihr Betrieb hinreichend transparent” gestaltet ist, sodass die “Anwender die Ergebnisse des Systems interpretieren und angemessen nutzen können” (KI-VO, 2024, Art. 13 Abs. 1).

Allerdings bleibt unklar, wie die "Interpretierbarkeit" in der KI-VO konkret definiert ist. Während frühere EU-Kommunikationen von "Erklärbarkeit" sprachen, bleibt offen, ob die europäische Regierungsbehörde die Begriffe nun gleichsetzt oder einen Unterschied dazwischen sieht (Ebers et al., 2021, nach Moser, 2023, S. 89). Obwohl meines Erachtens “Interpretierbarkeit“ mit “Erklärbarkeit” gleichzusetzten ist, fordert die KI-VO rechtlich gesehen lediglich eine Transparenz. So lässt sich festhalten, dass Art. 13 KI-VO keine Verpflichtung beinhaltet, die “allgemeine (innere) Entscheidungslogik des Systems zu erklären oder dem Nutzer Informationen darüber mitzuteilen” (Moser, 2023, S. 93). Die Transparenzvorgaben der KI-VO ändern wenig an den bereits in der DSGVO bestehenden, jedoch unzureichenden Informationspflichten, um eine umfassende Erklärbarkeit von KI-Systemen sicherzustellen. Sie bieten vielmehr eine erweiterte, aber für eine wirklich vertrauenswürdige KI noch immer ungenügende Form von Transparenz (Moser, 2023, S. 98).

3.2.6 Erklärbarkeit zur Klärung der Verantwortlichen

Neben den datenschutzrechtlichen Aspekten und der KI-Verordnung müssen auch Fragen zur Verantwortlichkeit und zivilrechtlichen Haftung für Schäden, die durch KI verursacht werden, berücksichtigt werden.

Um dies rechtlich zu verankern, hat die Europäische Kommission im September 2022 einen Vorschlag zur Gestaltung der KI-Haftungsrichtlinie veröffentlicht, mit welcher „europaweit einheitliche Regeln zur außervertraglichen Haftung für Schäden durch den Einsatz von künstlicher Intelligenz“ eingeführt werden sollen (Noerr PartGmbB, 2024). Die KI-Haftungsrichtlinie ist ergänzend zu der Produkthaftungsrichtlinie notwendig und hat einen eigenen Anwendungsbereich, „insbesondere für Schäden durch Diskriminierung, bei Verletzung von Persönlichkeits- oder IP-Rechten und bei rein wirtschaftlichen Schäden“ (Noerr PartGmbB, 2024).

So soll eine Offenlegungspflicht der Undurchsichtigkeit von Black-Box-Modellen entgegenwirken. Geschädigte hätten somit das Recht, Informationen zu KI-Systemen anzufordern, um potenzielle Ansprüche und Anspruchsgegner zu identifizieren sowie das KI-System auf Fehler oder Sicherheitslücken zu untersuchen (Noerr PartGmbB, 2024).

Ein Beispiel hierfür ist die Nutzung von KI zur Bewertung von Klausuren, bei der möglicherweise Bias behaftete Daten in großem Umfang eingeschleust wurden, um die Benotung zu beeinflussen. In solchen Fällen ist Erklärbare KI von großer Bedeutung, da sie es ermöglicht festzustellen, welche Bias wo genau zum Einsatz gekommen sind und möglicherweise auch ob diese absichtlich eingeführt wurden, um das System zu

beeinflussen. Auf Grundlage der Offenlegungspflicht und den gesammelten Informationen können Geschädigte rechtlich vorgehen.

3.3 Empfänger und Informationsgehalt der Erklärungen im Bildungswesen

Der Detailierungsgrad und Informationsgehalt einer Erklärung sollte je nach Empfänger variieren und auf die Sachkenntnisse des jeweiligen Adressanten angepasst sein. Daher ist zu berücksichtigen, ob es sich beim Adressanten bspw. um einen KI-Entwickler, einen Forscher, einer nicht fachkundigen Person oder einer Regulierungsbehörde handelt. Ein Bediener benötigt und versteht vielleicht nur Erklärungen, die die Eingabe- und Ausgabezustände des Systems auf einfache Weise verdeutlichen, während KI-Entwickler möglicherweise Erklärungen benötigen, die beschreiben, wie die Komponenten des KI-Modells eine Zuordnung zwischen den Eingaben und Ausgaben hergestellt haben (Samek & Müller, 2019, S. 6; Zednik, 2021, S. 266).

Tomsett et al. (2018) bieten eine nützliche Klassifikation der Stakeholder/Akteure im Machine-Learning-Ökosystem, also jener Personen und Institutionen, die von einem durch maschinelles Lernen gesteuerten Computersystem abhängig sind oder mit ihm interagieren. Hierbei identifizieren sie sechs verschiedene Stakeholder-Typen, welche sich hervorragend mit dem Bildungswesen verbinden lassen und im Folgenden dargestellt werden (Tomsett et al., 2018).

3.3.1 Bediener, Entscheider, Entscheidungs- und Datenbetroffene

Schüler, Lehrkräfte und administrative Bereiche einer Schuleinrichtung sind **Bediener** (orig. „Operators“) und/oder **Entscheider** (orig. „Executors“). Bediener sind Akteure, die direkt mit dem ML-System interagieren und das System mit Eingaben versorgen, wobei sie direkt die Ausgaben bekommen. Beispiele hierfür sind Schüler, die einen Chatbot benutzen oder Lehrkräfte, die anhand eines Tools die Lernstandsanalyse ihrer Schüler beobachten. Entscheider hingegen sind Akteure, die finale Entscheidungen treffen basierend auf den Entscheidungen des ML-Systems und den Informationen der Bediener. Speziell im Bildungswesen verschmelzen beide Stakeholder-Rollen, da es bislang keine KI-basierten Ergebnisse oder Entscheidungen im Bildungskontext gibt, die eine Bestätigung einer höheren Instanz als die der Lernenden, Lehrenden oder Eltern benötigen.

Ein weiterer wichtiger Akteur ist der **Entscheidungsbetroffene** (orig. „Decision-subjects“), der aktiv von den Entscheidungen des Entscheiders betroffen ist. Eng dazu verwandt ist der **Datenbetroffene** (orig. „Data-subjects“). Hierbei handelt es sich um Akteure, deren persönliche Daten zum Trainieren der ML-Systeme verwendet wurden. Diese Akteure sind nur dann betroffen, wenn das ML-System mit personenbezogenen Daten trainiert wurde. So sind z.B. Schüler und Lehrer Datenbetroffene, wenn das Sekretariat mittels KI-gestütztem Educational Data Mining und Learning Analytics schuliche Datenbestände erfasst und analysiert.

Für die vier genannten Akteure sind Erklärungen im Bildungswesen von besonderer Bedeutung, da sie aktiv die Folgen der Entscheidungen zu spüren bekommen, bspw. wenn ein Schüler auf Grund einer KI-generierten Note nicht versetzt wird oder wenn eine KI rassistische Verzerrungen aufweist. Da die Akteure hierbei hauptsächlich Lehrer und Schüler sind, sollten die Erklärungen leicht interpretierbar und kompakt sein, da umfangreiche KI-Kenntnisse keine Voraussetzung zur Nutzung von KI-Systemen im Bildungswesen sind. So kann es KI-Experten zufolge in vielen Fällen nicht hilfreich sein, „jedem alles zu sagen“ (Authority of the House of Lords, 2018, S. 38). Im Einzelfall müssen somit häufig „Kompromisse zwischen einer verbesserten Erklärbarkeit eines Systems (was die Präzision beeinträchtigen kann) und mehr Präzision (auf Kosten der Erklärbarkeit) eingegangen werden“ (HEG-KI, 2019, S. 22).

3.3.2 Schöpfer

Abseits aktiven Bedienern und Entscheidern eines KI-Systems wie etwa Schüler und Lehrkräfte sind die **Schöpfer** (orig. „Creators“) wichtige Stakeholder. Gemeint sind hierbei Akteure, die das ML-System erschaffen haben, sein es die Eigentümer (orig. „Owners“), die Eigentumsrechte am System besitzen, oder die Implementierenden (orig. „Implementers“), die das System technisch realisieren, meist als Angestellte oder beauftragte Dienstleister der Eigentümer. Dazu gehören die Softwareentwickler, die die KI-Implementierung sowie Erklärbarkeitsmethoden für das System entwerfen und entwickeln, aber auch Systemadministratoren, die für die Wartung und möglicherweise für die Feinabstimmungen des Systems verantwortlich sind.

Auf Grund der Tatsache, dass die Algorithmen eines Systems als Geschäftsgeheimnis oder geistiges Eigentum geschützt sind, sind die Entwickler bzw. Implementierenden die einzigen Akteure, die das jeweilige KI-System und die Erklärbarkeit entwickeln und integrieren können. Um zu gewährleisten, dass die Erklärungen des KI-Systems optimal gestaltet sind und kontinuierlich verbessert werden können, benötigen die Entwickler Zugang zu allen verfügbaren Informationen, einschließlich Beweise über die Entscheidungen des KI-Systems in der höchsten Auflösung (Samek & Müller, 2019, S. 6). Nur durch umfassende Informationen über das KI-Modell und ihren Bewegründen bzw. Parametern können sie das Verhalten des Systems verstehen und erklären.

3.3.3 Prüfer

Abschließend klassifizieren Tomsett et al. (2018) **Prüfer** (orig. „Examinors“) als Akteure, die mit der „Konformitäts-/Sicherheitsprüfung, der Auditierung oder der forensischen Untersuchung“ eines Systems beauftragt sind (Tomsett et al., 2018).

Regulatorische Institutionen wie Datenschützer, externe Wirtschaftsprüfungsgesellschaften wie Deloitte, interne Revisionsteams und Forschungseinrichtungen sind Akteure, die das System überprüfen, um sicherzustellen, dass es den gesetzlichen, ethischen und betrieblichen Standards entspricht. Anbetracht der rechtlichen Lage in der EU spielen die DSGVO, KI-VO und die KI-Haftungsrichtlinie im Kontext der Erklärbar-

keit eine bedeutende Rolle. Vor allem im Bildungswesen, bei welchem hochriskante KI-Systeme Anwendung finden, ist die Erklärbarkeit zunehmend von Bedeutung und Relevanz und es ist anzunehmen, dass Prüfer wahrscheinlich eine immer wichtigere Rolle in künftigen KI-Systemen spielen werden.

Während Sicherheitsprüfungen idealerweise vor der Inbetriebnahme eines Systems erfolgen, untersuchen Auditoren und forensische Ermittler hingegen vergangene Systemausgaben, unter anderem Erklärungen, um festzustellen, ob diese den Standards und Richtlinien entsprechen und nachvollziehbar sind. Prüfer könnten außerdem das System mit neuen Daten testen, um die Sensitivität der Entscheidungen durch hypothetische „Was-wäre-wenn"-Szenarien zu untersuchen. Anhand der untersuchten Daten können Prüfer darüber hinaus wertvolles Feedback an die Entwickler geben, um das System zu verbessern (Tomsett et al., 2018).

3.4 Potenzielle Risiken im Zusammenhang mit Erklärbarer KI im Bildungswesen

Automatisierte administrative Aufgaben, personalisierte Lernerfahrungen und neue Tutorensystemen bieten das Potenzial, das Bildungswesen mithilfe von Erklärbaren KI-Systemen grundlegend zu verändern. Dennoch birgt die Implementierung von Erklärbarer KI im Bildungswesen potenzielle Risiken, die bedacht und vermieden werden sollten und im Folgenden aufgeführt sind.

Herausforderung / Risiko / Begrenzung	Strategien zur Bewältigung
Abschreckung potenzieller Entwickler in der EU durch zu strenge KI-Vorschriften wie die KI-VO, besonders wenn sie eine umfassende Erklärbarkeit und Transparenz von KI-Systemen verlangen (Potenzielles Hindernis für Innovation in der EU)	• Gezielte Förderprogramme und Anreize für Unternehmen anbieten, insbesondere für die die umfassende Erklärbare KI integrieren • KI-VO derart ausgestalten, dass sie einen echten Standortvorteil für Länder wie Deutschland bedeutet • Unterstützung für Unternehmen durch bspw. KI-Reallabore (D16-Digitalminister, 2022, S. 4) • Investitionen in KI-Bildungsprogramme, um das Verständnis für KI-Technologien und deren ethische Implikationen in der Bevölkerung zu erhöhen
Notwendigkeit, im Einzelfall einen Kompromiss entweder zugunsten hö-	• Gezielte Kommunikation mit den Stakeholdern eines KI-Systems,

herer Präzision der Ergebnisse des KI-Modells (zum Nachteil der Erklärbarkeit), **oder zugunsten besserer Erklärbarkeit** (auf Kosten der Präzision) einzugehen	um die genauen Anforderungen nachvollziehen zu können • Forschung im Bereich der Erklärbaren KI vorantreiben, um sowohl die Präzision als auch die Interpretierbarkeit zu verbessern • Techniken wie LIME einsetzten, um die Erklärbarkeit komplexer Modelle zu gewährleisten, ohne deren Leistung zu beeinträchtigen
Unnötiger Einsatz komplexer Erklärbarkeitsmethoden, die möglicherweise nicht erforderlich sind oder sogar weniger verständliche Erklärungen liefern als einfachere Modelle (Khosravi et al., 2019, nach Rachha & Seyam, 2023, S. 236)	• Entwicklung mit einfachen, interpretierbaren Modellen beginnen und die Komplexität nur bei Bedarf erhöhen ⇒ Leistung in Bezug auf Genauigkeit als auch Interpretierbarkeit messen und vergleichen, um den besten Kompromiss zwischen Modellkomplexität und Erklärbarkeit zu finden (Conati et al., 2020)
Erklärungen, die Schülern **zu viele Details über die Funktionsweise des KI-Modells** bieten und so das **Risiko erhöhen**, dass sie versuchen, das **System zu manipulieren oder zu betrügen** (Rachha & Seyam, 2023, S. 236)	• Verstärkter Fokus auf aggregierte Erklärungen statt detaillierter Einblicke • Verwendung adaptiver Erklärungsmethoden, um Manipulationen zu erkennen
Einsatz ungeeigneter Modelle, die aufgrund ihrer Beschaffenheit nur unzureichende Erklärungen liefern oder zu plausiblen, aber falschen Erklärungen verleiten (Rachha & Seyam, 2023, S. 236)	• Sorgfältige Entwicklung der Modelle (gründliches Training, Validierung, Testen) • Erklärungen müssen auf mathematischen Grundlagen basieren, die das System wirklich akkurat beschreiben (Rachha & Seyam, 2023, S. 236) • Nicht der Versuchung nachgeben, "schöne" Erklärungen zu generieren, die nicht korrekt/akkurat beschreiben, wie das Modell tatsächlich zu der Entscheidung gekom-

	men ist
Unzureichende Berücksichtigung der Bedürfnisse und Fachkenntnisse der Stakeholder bezüglich Erklärungen; Risiko von Missinterpretationen bei zu komplexen oder zu stark vereinfachten Erklärungen	• Durchführung umfassender Stakeholder-Analysen vor KI-Implementierung • Regelmäßige Konsultationen mit Bildungsexperten, Lehrenden und Lernenden • Entwicklung von KI-Systemen und der Gestaltung der Erklärungen unter Berücksichtigung verschiedener Nutzerperspektiven • Implementierung von Feedback-Mechanismen zur kontinuierlichen Verbesserung

3.5 Konkretes Anwendungsbeispiel im Bildungswesen

In diesem Abschnitt wird ein Anwendungsbeispiel vorgestellt, das den Einsatz eines möglichen KI-Systems im Bildungswesen darstellt und in Kapitel vier zur Veranschaulichung der Kriterien zur Messung der Erklärbarkeit von KI-Systemen genutzt wird.

3.5.1 GradeAI Solutions: Ein KI-System zur automatisierten Bewertung und Benotung von Prüfungsleistungen

3.5.1.1 Übersicht

GradeAI Solutions ist eine innovative Lösung zur automatisierten Bewertung und Benotung von Prüfungsleistungen für Bildungseinrichtungen aller Art, von allgemeinbildenden Schulen bis hin zu Universitäten. Mithilfe des Systems, das durch den Einsatz moderner Technologien der KI ermöglicht wird, können Tests, Klausuren und andere schriftliche Aufgaben bewertet und benotet werden, wodurch Lehrkräfte entlastet und Ressourcen für andere Aufgaben frei werden.

GradeAI Solutions bietet nicht nur schnelle und vertrauenswürdige Benotungen, sondern liefert auch umfassende Rückmeldungen, die Lernenden dabei helfen, ihre Leistungen besser zu verstehen und sich gezielt zu verbessern. Die Flexibilität von GradeAI Solutions erlaubt es Bildungseinrichtungen, das KI-System an spezifische Anforderungen wie unterschiedliche Lehrpläne, Kurse oder Sprachen anzupassen, wodurch eine nahtlose Integration in bereits bestehende Lehr- und Lernprozesse gewährleistet wird.

3.5.1.2 Stakeholder/Akteure

Das Ökosystem von GradeAI Solutions umfasst verschiedene Stakeholder/Akteure, die auf unterschiedliche Weise in die Entwicklung, den Einsatz und die Überprüfung des KI-Systems eingebunden sind. Die verschiedenen Rollen lassen sich treffend anhand der Klassifikation von Tomsett et al. (2018) beschreiben (siehe Kapitel 3.3).

Zu den Nutzern des Systems gehören die **Bediener** (orig. “Operators”), also Lehrkräfte, Administratoren und Schüler/Studenten, die direkt mit GradeAI Solutions interagieren und das System mit Eingaben versorgen, wobei sie direkt die Ausgaben bekommen. Lehrkräfte laden somit noch nicht bewertete Klausuren hoch, können optional Bewertungskriterien definieren und analysieren die Ergebnisse. Lernende hingegen können mit dem System interagieren, indem sie bspw. nach individuellen Erklärungen fragen.

Lehrende sind in diesem Anwendungsbeispiel sowohl Bediener als auch **Entscheider** (orig. „Executors“), da sie die vom KI-System bereitgestellten Informationen nutzen und die finale Entscheidung treffen, ob die Benotung und Bewertung gerechtfertigt ist und so in den Lehrplan eingetragen werden kann. Sieht der Entscheider die Benotung nicht ein, kann er die Bewertung von GradeAI Solutions verwerfen und eine manuelle Bewertung vornehmen.

Die **Entscheidungsbetroffenen** (orig. „Decision-subjects“), in diesem Fall die Lernenden, sind unmittelbar von den Bewertungen des Systems betroffen. Die Benotungen haben direkten Einfluss auf Noten, Lernfortschritte, Versetzungen und möglicherweise die Motivation der Lernenden.

Da personenbezogene Daten wie etwa frühere Prüfungsleistungen oder Fehlermuster für das Training der KI-Modelle verwendet werden, spielen auch die **Datenbetroffenen** (orig. „Data-subjects“) eine wichtige Rolle. Der Schutz ihrer Daten ist von Bedeutung, um ethische und regulatorische Anforderungen zu erfüllen.

Eine zentrale Gruppe stellen außerdem die **Schöpfer** (orig. “Creators”) dar, die für die Erschaffung, Entwicklung und Wartung des KI-Systems verantwortlich sind. Innerhalb dieser Gruppe sind die **Eigentümer** (orig. ”Owners”) von GradeAI Solutions, die die Eigentumsrechte am System besitzen und für die strategische Ausrichtung zuständig sind, während die **Implementierenden** (orig. “Implementers”) als Entwickler und Techniker die technische Umsetzung, das Training und die Wartung der KI-Modelle sicherstellen. Ebenfalls liegt in ihrer Verantwortung die Integration der Erklärbarkeitsmodelle in das KI-System.

Abschließend gibt es die **Prüfer** (orig. “Examinors”), wie Auditoren oder externe Prüfer, und sind mit der „Konformitäts-/Sicherheitsprüfung, der Auditierung oder der forensischen Untersuchung“ eines Systems beschäftigt (Tomsett et al., 2018). Diese Stakeholder-Rolle sorgt dafür, dass GradeAI Solutions den hohen Ansprüchen gerecht wird, darunter rechtliche Ansprüche wie die der DSGVO und der KI-VO, ethische Ansprüche wie etwa Nichtdiskriminierung oder auch technische Ansprüche. Wer konkret für die Prüfung von GradeAI Solutions verantwortlich ist, wird in Kapitel 3.5.1.4 erläutert.

3.5.1.3 Hochrisiko-KI-System gemäß der KI-VO

Gemäß den Einstufungsvorschriften in Art. 6 KI-VO ist das KI-System von GradeAI Solutions als Hochrisiko-KI-System einzustufen. Diese Einstufung ergibt sich unter anderem aus Anh. III Abs. 3b, dem zufolge „KI-Systeme, die zur Bewertung von Lernergebnissen eingesetzt werden“, als hochriskant einzustufen sind. Zudem bestätigt Anh. III Abs. 3c diese Einstufung, wonach „KI-Systeme, die zur Bewertung des angemessenen Bildungsniveaus, das eine Person erhalten wird oder zu dem sie Zugang haben wird“, ebenfalls als Hochrisiko-KI-System einzustufen sind (KI-VO, 2024, Anh. III Abs. 3 KI-VO).

Für KI-Systeme mit der Risikoklasse „hoch“ gelten besondere rechtliche Anforderungen, darunter die Pflicht zur Transparenz, welche im Prüfrahmenwerk in Kapitel vier aufgegriffen wird.

3.5.1.4 Wer ist für den Audit der Erklärbarkeit zuständig?

Das Audit für GradeAI Solutions wird von einer Vielzahl von Beteiligten durchgeführt, die aus verschiedenen Perspektiven beurteilen, ob eine zufriedenstellende Erklärbarkeit gegeben ist.

Zu diesen Beteiligten gehören Auditoren wie KI-Experten von unabhängigen spezialisierten Beratungsfirmen oder Forschungseinrichtungen, aber auch Entwickler und Ingenieure, die direkt bei GradeAI Solutions arbeiten. Diese Personen sind mit KI-Systemen bestens vertraut und können insbesondere die technischen Kriterien der Erklärbarkeit nachvollziehen und kontrollieren.

Zusätzlich überprüfen regulatorische und rechtliche Prüfer, ob das KI-System die gesetzlichen Anforderungen erfüllt, wie etwa den datenschutzkonformen Umgang mit personenbezogenen Daten sowie die Einhaltung der Transparenzpflicht gemäß der KI-VO. Speziell für das Überwachen der Einhaltung der Vorgaben der KI-VO, darunter auch das Durchführen von Audits und Inspektionen im Zusammenhang mit der Transparenz von KI-Systemen, hat die Bundesregierung Deutschland die zentrale nationale Rolle für die KI-Aufsicht der Bundesnetzagentur übergeben. Parallel dazu sind auch die Datenschutzbehörden in diese Prozesse eingebunden (Laudien, 2024).

Endbenutzer wie Lehrende und Lernende spielen auch eine wichtige Rolle im Auditierungsprozess, da sie eine aktive Rolle bei der Nutzung und Interpretation von GradeAI Solutions spielen. Sie können konkret prüfen, ob die Erklärungen ihren Bedürfnissen entsprechen und ob sie verständlich und praktikabel gestaltet sind. Somit würden sie sich eher auf die menschenzentrierte Perspektive also auch auf die Mensch-Technik-Interaktion fokussieren.

4 Prüfrahmenwerk für Erklärbare KI mit einem Anwendungsbeispiel aus dem Bildungswesen

Das in diesem Kapitel dargestellte Prüfrahmenwerk adressiert einer der größten Herausforderungen für heutige KI-Anwendungen: Eine Erklärbarkeit zu gewährleisten, die technologische Komplexität in Einklang mit menschlichen Bedürfnissen und gesellschaftlichen Anforderungen gleichermaßen berücksichtigt. Mit dem breiten Einsatz von KI-Systemen in zunehmend kritischeren Anwendungsbereichen wie dem Bildungswesen fehlen ausgereifte und standardisierte Rahmenwerke, die detaillierte und klare Erklärbarkeitskriterien definieren. Somit wird eine systematische Prüfung von KI-Systemen durch Auditoren verzögert, um deren Transparenz, Fairness und Nachvollziehbarkeit sicherzustellen.

Das vorliegende Prüfrahmenwerk setzt genau hier an, indem es Erklärbarkeitskriterien in drei Dimensionen strukturiert: die Soziale, Sozio-technische und Technische Dimension.

Das Rahmenwerk dient einerseits als mögliche Grundlage für Auditoren, mangelhafte Erklärbarkeit und hierbei im Zusammenhang stehende potenzielle Risiken zu bewerten, zu verstehen und an relevante Stakeholder zu kommunizieren. Anderseits kann das Prüfrahmenwerk von Entwicklern, Nutzern, Fachleuten und Regulierungsbehörden als Orientierungshilfe genutzt werden. Es bietet die Möglichkeit, das Prüfrahmenwerk weiterzuentwickeln und an spezifische Anforderungen und Anwendungsfälle anzupassen. Zudem ist hervorzuheben, dass rechtliche Regulation wie die KI-Verordnung klar verdeutlichen, dass die Relevanz und die Anforderungen an die Erklärbarkeit von KI-Systemen in Zukunft weiter zunehmen wird.

Eine weitere Besonderheit des hier vorgestellten Prüfrahmenwerks ist die Integration eines konkreten Anwendungsbeispiels: GradeAI Solutions ist ein KI-System zur automatisierten Bewertung und Benotung von Prüfungsleistungen (siehe Kapitel 3.5). Dieses Beispiel dient dazu, die Anwendung der abstrakten Erklärungskriterien zu veranschaulichen und dient zudem als Orientierung für Auditoren in der Praxis, die vor der Aufgabe stehen, KI-Systeme, darunter auch wie hier, Hochrisiko-KI-Systeme auf ihre Erklärbarkeit zu prüfen.

Die in diesem Prüfrahmenwerk aufgegriffenen Kriterien basieren auf einer umfassenden Analyse der bestehenden Literatur zur Messung der Erklärbarkeit von KI-Systemen (siehe Kapitel 1.4). Diese Kriterien wurden sorgfältig ausgewählt, teilweise zusammengeführt und in drei Dimensionen strukturiert, um eine systematische und umfassende Bewertung der Erklärbarkeit von KI-Systemen zu ermöglichen.

Das Prüfrahmenwerk ist grundsätzlich branchenübergreifend anwendbar wie z.B. im Gesundheitswesen, Finanzwesen oder in der Logistik. Durch das konkrete Anwen-

dungsbeispiel bekommt es jedoch einen spezifischen Bezug zum Bildungskontext und kann so besser als Orientierung für Auditoren in der Praxis genutzt werden.

Eine vollständige Übersicht aller Kriterien sowie der dazugehörigen Audit-Fragen zur Messung der Erklärbarkeit von KI-Systemen, gegliedert nach den drei Dimensionen (Soziale, Sozio-technische und Technische Dimension), findet sich in Anhang A. In den folgenden Unterabschnitten werden diese Metriken im Detail erläutert und anhand des Anwendungsbeispiels veranschaulicht.

4.1 Soziale Dimension: Fokus auf die menschenzentrierte Perspektive der Erklärbarkeit

In dieser Dimension steht die Bedeutung der Erklärbarkeit für individuelle Nutzer im Vordergrund. Es geht hierbei um die rechtlichen, ethischen und sozialen Anforderungen der Nutzer an die Erklärbarkeit, die sicherstellen sollen, dass KI-Systeme nachvollziehbar, gerecht und nicht diskriminierend agieren. Es liegt der Fokus darauf, wie Erklärungen dazu beitragen, Vertrauen zu stärken und die Rechte der Betroffenen zu wahren.

S1 ***Funktionale Erklärbarkeit (Engl. Functionality):*** Ist das KI-System so konzipiert, dass es von den Nutzern verstanden werden kann, oder bietet es Mechanismen wie Erklärbare KI-Modelle, um seine Entscheidungsprozesse nachträglich verständlich zu machen? Ein zentraler Aspekt der funktionalen Erklärbarkeit besteht darin, sicherzustellen, dass ein KI-System entweder direkt verständlich ist oder durch zusätzliche Mechanismen wie Erklärbare Modelle erklärt werden kann (Langer et al., 2021, S. 165).

Anwendungsbeispiel: GradeAI Solutions vermittelt seine Entscheidungsprozesse auf eine für die Nutzer nachvollziehbare und verständliche Weise. Beispielsweise sollte es transparent darlegen können, welche Kriterien zur Bewertung beigetragen haben. Dadurch wird die Nachvollziehbarkeit der Ergebnisse für Lehrende und Lernende gewährleistet.

S2 ***Vertrauensvolle/Treue Erklärbarkeit (Engl. Trust/Fidelity):*** Werden Informationen bereitgestellt, die die Entscheidungsprozesse des KI-Systems genau und wahrheitsgemäß beschreiben? Es ist essenziell, dass ein KI-System präzise und wahrheitsgemäße Einblicke in seine Entscheidungsprozesse bietet, um sowohl die Vertrauenswürdigkeit des Systems als auch die Zuverlässigkeit der gelieferten Informationen sicherzustellen (Hoffman et al., 2018, S. 19; Langer, Baum, et al., 2021, S. 165). Beispiele für solche Informationen sind das Offenlegen kausaler Zusammenhänge, bei denen das System erklärt, welche spezifischen Eingabedaten direkt zu einer bestimmten Entscheidung geführt haben. Ein weiteres Beispiel ist die

Transparenz in der Gewichtung von Entscheidungsparametern, bei der das System angibt, wie verschiedene Faktoren in die Entscheidung einfließen.

Anwendungsbeispiel: Gerade im Bildungswesen, wo noch Misstrauen in der noch relativ jungen Technologie steckt, ist die vertrauensvolle/treue Erklärbarkeit von großer Bedeutung, um die Vertrauenswürdigkeit des Systems und die Zuverlässigkeit der gelieferten Informationen sicherzustellen. So könnte GradeAI Solutions aufzeigen, dass eine bestimmte Note hauptsächlich aufgrund der inhaltlichen Korrektheit der Antwort vergeben wurde, während stilistische Schwächen weniger Einfluss hatten, z.B. "50% der Punktzahl basieren auf inhaltlicher Richtigkeit, 30% auf Stil und 20% auf Rechtschreibung".

S3 Zufriedenheit mit Erklärungen (Engl. Satisfaction): Sind die Erkärungen für den Nutzer subjektiv zufriedenstellend? Die Zufriedenheit mit einer Erklärung hängt davon ab, wie gut die Nutzer das Gefühl haben, das System zu verstehen, und entsteht oft durch Klarheit, Kontextualität und die Berücksichtigung individueller Befürnisse (Hoffman et al., 2018, S. 5).

Anwendungsbeispiel: Auditoren können die Zufriedenheit der Lernenden und Lehrenden evaluieren, indem sie sie darüber befragen, ob die Erklärungen die Bewertungslogik von GradeAI Solutions verständlich und nachvollziehbar darstellen.

S4 *Verständlichkeit der Erklärungen (Engl. Comprehensibility):* Helfen die Erklärungen den Nutzern, die Entscheidungsprozesse des Systems besser nachzuvollziehen? Wie viel Aufwand ist nötig, damit Menschen die Erklärungen verstehen? Dieses grundlegende Kriterium ist besonders entscheidend, um sicherzustellen, dass dessen Nutzung transparent ist und um Vetrauen in das System aufzubauen (Coroama & Groza, 2022, S. 405; Langer, Oster, et al., 2021, S. 14).

Anwendungsbeispiel: GradeAI Solutions sollte nach der Bewertung einer schriftichen Prüfung nicht nur eine Punktzahl ausgeben. Das KI-System sollte eine Übersicht mit Erklärungen bieten, welche Bewertungskriterien es angewendet hat wie z.B. Vollständigkeit der Antwort oder sprachliche Qualität und wie diese zur Gesamtbewertung beigetragen haben. Eine klare und einfache Sprache, aber auch visuelle Hilfsmittel wie etwa Diagramme können dazu beitragen, dass Lehrende und Lernende die Logik des Systems nachvollziehen können.

S5 *Nichtdiskriminierung, Ethische Verantwortung und Rechenschaftspflicht (Engl. Non-Discrimination, Ethical Responsibility and Accountability):* Ermöglicht die Erklärbarkeit des KI-Systems eine verantwortungsvolle Entscheidungsfindung und eine klare Zuweisung von Verantwortung für getroffene Entscheidungen? Ein Erklärbares KI-System sollte Informationen bereitstellen, mithilfe dessen Nutzer fundierte und verantwortungsbewusste Entscheidungen treffen können, insbesondere wenn Men-

schen aktiv in den Entscheidungsprozess eingebunden sind (engl. Human-in-the-Loop). Hierbei ist es zum einen von Bedeutung, dass sowohl Entscheider als auch Entscheidungsbetroffene (siehe Kapitel 3.3.1) mithilfe der Erklärungen die zugrunde liegenden Daten und Prozesse verstehen können, um ihre Entscheidungen auf ethische Grundsätze wie Fairness und Gleichbehandlung prüfen zu können. Dadurch wird sichergestellt, dass die KI-Entscheidungen moralisch vertretbar sind. Zum anderen ist es hierbei von Bedeutung, mithilfe von Erklärungen die verantwortlichen Akteure zu identifizieren, um unter anderem bei fragwürdigen Entscheidungen oder systematischen Fehlern Rechenschaft oder gar Schadensersatz einzufordern (Noerr PartGmbB, 2024; Andrae, 2021, S. 41; Langer, Baum, et al., 2021, S. 166).

Anwendungsfall: Ein Elternteil vermutet, dass die Bewertung der Tochter durch GradeAI Solutions unfair oder diskriminierend ausgefallen ist. Somit möchten sie verstehen, wie das Ergebnis zustande gekommen ist und möchten mögliche Verantwortliche für eine fehlerhafte Bewertung identifizieren. Mithilfe der bereitgestellten Erklärung durch GradeAI Solutions können sie die angewendeten Kriterien, spezifischen Datenpunkte und Regeln einsehen. Durch diese Transparenz haben sie die Möglichkeit, Fehler oder Ungerechtigkeiten festzustellen und Verantwortliche zu identifzieren.

S6 ***Einhaltung der Anforderungen der DSGVO im Kontext der Erklärbarkeit von KI-Systemen (Engl. Compliance with the requirements of the GDPR):*** Entspricht die Datenverarbeitung den datenschutzrechtlichen Vorgaben? Sofern personenbezogene Daten verarbeitet werden, sind die Anforderungen der DSGVO zu beachten, um die Rechte betroffener Personen und den Schutz personenbezogener Daten sicherzustellen. Im Hinblick auf die Erklärbarkeit von KI-Systemen lassen sich daraus folgende Kriterien ableiten:

S6.1 ***Transparenz der Datenverarbeitung (Art. 12-14 DSGVO):*** Werden die Betroffenen umfassend über die Verarbeitung personenbezogener Daten informiert? Gemäß Art. 12 Abs. 1 DSGVO müssen Unternehmen, die personenbezogene Daten verarbeiten, Betroffenen Auskunft darüber in “präziser, transparenter, verständlicher und leicht zugänglicher Form“ zur Verfügung stellen (DSGVO, 2016, Art. 12 Abs. 1). Zudem müssen Betroffene gemäß Art. 13 DSGVO „über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung” aufgeklärt werden (DSGVO, 2016, Art. 13 Abs. 2 lit f). Außerdem muss Art. 14 DSGVO erfüllt sein, wenn die Daten nicht direkt von den Betroffenen erhoben werden.

Anwendungsbeispiel: Die Erklärungen, die GradeAI Solutions bereitstellt, enthalten personenbezogene Daten, da sie spezifische Informationen enthalten, die auf eine Person zurückzuführen sind. Darunter zählen Erklärungen, die Details zu den Antworten eines einzelnen Schülers auf spezifische Prüfungsfragen enthalten,

oder auch Erklärungen, die beschreiben, wie individuelle Leistungsdaten (z.B. Notenverläufe, Fehleranalysen) in die Bewertung eingeflossen sind. Somit muss GradeAI Solutions den Betroffenen sowohl Auskunft über die Verarbeitung personenbezogenen Daten geben, als auch über die involvierte Logik sowie die Tragweitegeben und die angestrebten Auswirkungen.

S6.2 ***Nachweis einer rechtlichen Grundlage (Art. 5-6 DSGVO):*** Basiert die Verarbeitung der personenbezogenen Daten auf einer ausreichend rechtlichen Grundlage und wurden die Grundsätze der Datenverarbeitung (z.B. Datenminimierung) beachtet? Art. 5 DSGVO schreibt Prinzipien wie etwa Datenminimierung und Speicherbegrenzung vor, die eingehalten werden müssen. Gemäß Art. 6 DSGVO ist außerdem eine rechtliche Grundlage für die Datenverarbeitung notwendig, wie z.B. Einwilligung, Vertragserfüllung oder berechtigtes Interesse.

Anwendungsbeispiel: GradeAI Solutions hat dafür zu sorgen, dass die personenbezogenen Daten rechtmäßig verarbeitet und die betroffenen Personen ordnugnsgemäß informiert wurden, wie z.B. durch Einwilligung oder Vertrag. Ebenfalls müssen die Erklärungen datenschutzkonform gestaltet sein, sodass unter anderem die Grundsätze der Datenminimierung beachtet werden.

S6.3 Anforderungen bei automatisierten Entscheidungen (Art. 22 DSGVO): Sind die betroffenen Personen vor nachteiligen Auswirkungen vollständig automatisierter Entscheidungen ohne menschliches Eingreifen geschützt? Wird eine vollautomatisierte Entscheidung durch ein KI-System nicht noch einmal von einem Menschen überprüft wird, sondern direkt in eine eigene Entscheidung umgesetzt, gilt Art. 22 DSGVO (siehe Kapitel 3.2.5). Es sind die Voraussetzungen des Art. 22 Abs. 2 DSGVO zu prüfen, unter anderem ob eine ausdrückliche Einwilligung der betroffenen Personen vorliegt, sofern keine gesetzliche Grundlage existiert. Ebenfalls sollte geprüft werden, ob eine Entscheidung plausibel durch einen Menschen geprüft wird, sodass Art. 22 DSGVO nicht greift.

Anwendungsbeispiel: Lehrende können selbst entscheiden, ob sie die Entscheidung bzw. Benotungen von GradeAI Solutions vor der Veröffentlichung überprüfen und anpassen wollen. Wenn dies nicht der Fall ist, muss geprüft werden, ob die betroffenen Schüler über ihr Recht gemäß Art. 22 DSGVO aufgeklärt wurden.

S7 ***Einhaltung der Transparenzanforderungen der KI-VO im Kontext der Erklärbarkeit von KI-Systemen (Art. 13 KI-VO) (Engl. Compliance with the requirements of the AI Act):*** Ist das Hochrisiko-KI-System hinreichend transparent gestaltet? Gemäß Art. 13 Abs. 1 KI-VO liegt eine Verpflichtungen zur transparenten Gestaltung dieser Systeme und zur Übermittlung von bestimmten Informationen an die Nutzer vor, sodass die Nutzer die Ergebnisse des KI-Systems “angemessen

interpretieren und verwenden können" (KI-VO, 2024, Art. 13 Abs. 1). Die zu gewährleistende Transparenz muss "in einem angemessenen Maß gewährleistet" werden, damit die Nutzer und Anbieter die "festgelegten einschlägigen Pflichten" nach Abschnitt 3 erfüllen können (KI-VO, 2024, Art. 13 Abs. 1).

Abschnitt 2 sieht vor, dass Hochrisiko-KI-Systeme "mit Gebrauchsanweisungen in einem geeigneten digitalen Format oder auf andere Weise beizufügen sind, die prägnante, vollständige, korrekte und klare Informationen enthalten, die für die Anwender relevant, zugänglich und verständlich sind" (KI-VO, 2024, Art. 13 Abs. 2).

Die Mindestanforderungen an die Betriebsanleitung sind in Abschnitt 3 aufgelistet. Einige davon sind relevant für die Erklärbarkeit von KI-Systemen, die meines Erachtens als ex-ante-Informationen gewertet werden können. Demnach sind nach Art. 13 Abs. 3b Nr. iv KI-VO "gegebenenfalls technische Fähigkeiten und Merkmale des Hochrisiko-KI-Systems" in der Betriebsanleitung von Bedeutung, "um Informationen bereitzustellen, die zur Erläuterung seiner Ausgaben relevant sind" (KI-VO, 2024, Art. 13 Abs. 3b Nr. iv). Ebenfalls sind gemäß Art. 13 Abs. 3b Nr. vi KI-VO "gegebenenfalls Spezifikationen für die Eingabedaten oder sonstiger relevanter Informationen über die verwendeten Trainings-, Validierungs- und Testdatensätze, unter Berücksichtigung der Zweckbestimmung des Hochrisiko-KI-Systems" von Relevanz (KI-VO, 2024, Art. 13 Abs. 3b Nr. vi).

Abschließend ist anzumerken, dass Art. 13 KI-VO den Anbieter eines KI-Systems nicht dazu verpflichtet, die von einer KI-basierten Entscheidung betroffenen Person eine Erklärung zu liefern. Es wird lediglich vorgeschrieben, dass Anbieter ihre Hochrisiko-KI-Systeme so konzipieren und entwickeln müssen, dass deren Ergebnisse von einem Nutzer interpretiert werden können. Auditoren haben somit rechtlich gesehen lediglich zu prüfen, ob die Mittel für eine Interpretation zur Verfügung gestellt werden. Ob der Nutzer die Ergebnisse letztlich richtig oder falsch interpretiert hat und somit richtig versteht, bleibt ungewiss und ist nicht die Aufgabe des Auditors (mehr unter Kapitel 3.2.5).

Anwendungsbeispiel: GradeAI Solutions fällt gemäß Anh. III Abs. 3 Nr. c und d KI-VO unter ein Hochrisiko-KI-System (siehe Kapitel 3.2.5). Auditoren haben zu prüfen, ob das KI-System hinreichend transparent konzipiert und entwickelt wurde, sodass die Nutzer die Ausgaben angemessen interpretieren können. Es wird jedoch nicht überprüft, ob die Lehrkräfte die Ergebnisse tatsächlich korrekt interpretieren, da dies außerhalb des rechtlichen Rahmens der KI-VO liegt. Ebenfalls muss GradeAI Solutions detaillierte Betriebsanleitung bereitgestellt werden, die die Anforderungen gemäß Art. 13 Abs. 3 KI-VO erfüllt.

S8 ***Klärung der Verantwortlichen durch Erklärbarkeit (KI-Haftungsrichtlinie) (Engl. Liability):*** Ist eine fundamentale Erklärbarkeit gegeben, mit welcher sich die Verantwortlichen entlang der gesamten Wertschöpfungskette des KI-Systems klären lassen? Mit Hilfe von Erklärungen lässt sich zurückführen, ob ein potenzieller Fehler

bspw. auf eine fehlerhafte Algorithmenentwicklung durch den Anbieter, eine unsachgemäße Anpassung durch den Betreiber oder eine unzureichende Datenqualität zurückzuführen ist. So wird es ermöglicht, präzise nachzuvollziehen, wer für welche Aspekte der Entwicklung, Bereitstellung und Nutzung des KI-Systems verantwortlich ist. Diese Nachvollziehbarkeit ist essenziell, um bei Streitigkeiten oder fehlerhaften Ergebnissen die jeweiligen Verantwortlichen zu identifizieren und mögliche rechtliche Konsequenzen abzuleiten. Für die KI-Haftung gibt es derzeit noch keine spezialgesetzlichen Regelungen und stellt aktuell keine direkte gesetzliche Verpflichtung im Rahmen der KI-VO dar, weshalb dieses Kriterium rechtlich gesehen noch keine Pflicht darstellt. Jedoch ist zu erwarten, dass die Bedeutung dieses Kriteriums durch die geplante Einführung der KI-Haftungsrichtlinie in Zukunft erheblich zunehmen und gesetzlich verankert wird (siehe Kapitel 3.2.6) (Noerr PartGmbB, 2024; Schreiber & Dimov, 2024).

Anwendungsbeispiel: Ein Elternteil ist besorgt und vermutet, dass der Bewertungsalgorithmus von GradeAI Solutions ihre Tochter systematisch diskriminierend benachteiligt hat. Die Eltern sind fest entschlossen, dass hierbei ein Schaden verursacht wurde, und wollen mit Hilfe der Offenlegungspflicht der neuen KI-Haftungsrichtlinie potenzielle Ansprüche und Anspruchsgegner ermitteln und das KI-System auf Fehler oder Sicherheitslücken untersuchen. Auditoren müssen sicherstellen, dass die Entscheidungen des KI-Systems verständlich dokumentiert wurden, sodass klar wird welche Schritte zu welchem Ergebnis geführt haben. So lassen sich die verantwortlichen Akteure des potenziellen Fehlers klären und Schadensersatzansprüche geltend machen.

4.2 Sozio-technische Dimension: Fokus auf die Mensch-Technik-Interaktion der Erklärbarkeit

Hier liegt der Fokus auf der Interaktion zwischen den Nutzern und des KI-Systems. Die Kriterien zeigen, wie Erklärungen gestaltet werden sollten, damit der Nutzer sie effektiv verstehen und mit der Technologie interagieren kann. Es werden sowohl die technischen Fähigkeiten der KI als auch die Bedürfnisse und Erwartungen der Nutzer berücksichtigt.

ST1 ***Erklärungsdomäne (Engl. Explanation Domain)*****:** Werden die bereitgestellten Erklärungen entweder im ursprünglichen Format der Modellparameter und Datenmerkmale (Originaldomäne) bereitgestellt, oder in einer umgewandelten, verständlicheren Form (Transformationsdomäne)? Dabei wird bewertet, ob die gewählte Darstellungsform für die Nutzer nachvollziehbar und geeignet ist, die Entscheidungen des KI-Systems verständlich zu machen (Sokol & Flach, 2019, S. 58).

Anwendungsbeispiel: GradeAI Solutions könnte die Erklärungen im zugrunde liegenden Format der Modellparameter und Merkmale in Form von numerischen Scores oder Gewichtungen darstellen (Originaldomäne). Alternativ könnte es die Ergebnisse in einer umgewandelten Form präsentieren, etwa als anschauliche

Diagramme oder als textuelle Rückmeldungen, die die Bewertungskriterien in verständlicher Sprache erläutern (Transformationsdomäne). Die Prüfung würde sicherstellen, dass die bereitgestellten Erklärungen in der jeweiligen passenden Form bzw. Domäne entsprechend der Bedürfnisse der Zielgruppen bereitgestellt werden.

ST2 ***Kausale und Handlungsorientierte Erklärungen (Engl. Causality vs. Actionability):*** Wird dem Nutzer des KI-Systems deutlich kommuniziert, ob kausale oder handlungsorientierte Erklärungen verwendet werden, um Missverständnisse und falsche Schlussfolgerungen zu vermeiden und sicherzustellen, dass keine kausale Verbindung irrtümlich angenommen wird? Kausale Erklärungen stellen eine direkte Ursache-Wirkung-Beziehung zwischen den Eingabedaten und den Entscheidungen des Systems dar. Eine Änderung der Eingabedaten führt dabei zu einer spezifischen Veränderung in der Entscheidung. Handlungsorientierte Erklärungen hingegen fokussieren sich darauf, welche Maßnahmen ergriffen werden können, um das Ergebnis zu beeinflussen, ohne eine direkte kausale Beziehung zu implizieren (Sokol & Flach, 2019, S. 59).

Anwendungsbeispiel: Eine kausale Erklärung, die die Bedeutung der Rechtschreibung für das Endergebnis verdeutlicht, könnte so lauten: „Die Punkte für Rechtschreibung und Grammatik sind entscheidend für die Note, da sie 30 % des Endergebnisses ausmachen“. Anhand dieser Erklärung wird eine direkte Ursache-Wirkung-Beziehung zwischen der Rechtschreibung und der Endbewertung hergestellt. Andererseits könnte das System auch eine handlungsorientierte Erklärung anbieten, wie zum Beispiel: „Achten Sie beim nächsten Mal verstärkt auf Rechtschreibung, um Ihre Bewertung zu verbessern“. Es wird lediglich dem Nutzer die Maßnahme empfohlen, dass eine Veränderung in der Rechtschreibung die Bewertung potenziell verbessern kann, jedoch ohne eine direkte Ursache-Wirkung-Beziehung herzustellen.

ST3 ***Anpassungsfähigkeit der Erklärungen (Engl. Adaptability):*** Sind die Erklärungen an die unterschiedlichen Bedürfnisse und das Hintergrundwissen der jeweiligen Nutzer angepasst? Das erfordert, dass Erklärungen hinsichtlich der verwendeten Sprache und Terminologie als auch in Bezug auf die Granularität der bereitgestellten Informationen zielgruppenspezifisch gestaltet werden (Walton, 2007). Fachleute benötigen oft detaillierte technische Erklärungen, während es für herkömmliche Nutzer eines KI-Systems häufig ausreicht, vereinfachte und leicht verständliche Erkärungen zu erhalten. Auditoren überprüfen, ob das System die Sprache und die Komplexität der Erklärungen an das Wissen der Zielgruppe anpasst und ob es möglich ist, zwischen unterschiedlichen Granularitätsstufen zu wechseln (Langer, Baum, et al., 2021, S. 165; Schneider & Handali, 2019).

Anwendungsbeispiel: Für Lehrende liefert GradeAI Solutions im idealen Fall detaillierte Erklärungen, die fachliche Tiefe und analytische Genauigkeit aufweisen, wie bspw.: „Die Note wurde mit einem Punktabzug von 20% aufgrund falscher

Rechenwege ermittelt. Die fehlerhafte Anwendung der Multiplikationsregel im zweiten Schritt führte zu einem falschen Ergebnis“.

Für Lernende gestaltet das System die Erklärung einfacher und motivationsfördernd: „Deine Rechnung war fast richtig, aber bei der Multiplikation ist dir ein kleiner Fehler passiert. Wenn du den zweiten Schritt noch einmal prüfst, wirst du es das nächste Mal richtig machen“.

Eltern erhalten eine Erklärung, die den Fortschritt des Kindes betont und allgemein verständlich formuliert ist, in etwa so: „Ihr Kind hat bei der Rechenaufgabe solide Ansätze gezeigt, allerdings hat ein kleiner Rechenfehler das Endergebnis beeinflusst. Dies bietet eine gute Grundlage, um weiter zu üben“.

ST4 ***Erklärungsmedium (Engl. Explanatory Medium)****:* Bietet das KI-System Erklärungen in einem Medium an, das für die Zielgruppe geeignet ist und die Art der Information verständlich vermittelt, ohne die Nuzter zu überfordern oder wichtige Details auzulassen? Die Wahl des Erklärungsmediums wie z.B. Text, Visualisierung, statistische Zusammenfassung oder eine Kombinationen davon ist entscheidend um sicherzustellen, dass die Erklärungen klar, prägnant und auf die Bedürfnisse der Nutzer abgestimmt sind (Sokol & Flach, 2019, S. 58).

Anwendungsbeispiel: Lehrende könnten eine Visualisierung nutzen, um den Bewertungsprozess nachvollziehbar zu machen. Ein Diagramm könnte bspw. zeigen, wie unterschiedliche Kriterien (z. B. inhaltliche Richtigkeit, Rechtschreibung, Struktur) zur Gesamtbewertung beitragen. Zusätzlich könnte eine kurze textuelle Erklärung bereitgestellt werden, etwa: „Die Note basiert zu 40% auf inhaltlicher Korrektheit, zu 30% auf Rechtschreibung und zu 30% auf Textstruktur. Der Fehler im dritten Rechenschritt hat die Punktzahl inhaltlich um 10% reduziert“.

ST5 ***Interaktive Kommunikation mit Erklärungen (Engl. Interactivity):*** Können Nutzer mit den bereitgestellten Erkärungen interagieren? Kann das KI-System auf spezifische Nachfragen eingehen und bei Bedarf zusätzliche Informationen oder eine tiefere Analyse bereitstellen? Es ist zu prüfen, ob das KI-System ein geeignetes Kommunikationsprotokoll bereitstellt, das den Anforderungen der Zielgruppe entspricht. Das Kommunikationsprotokoll einer Erklärbaren KI kann statisch, interaktiv oder eine Kombination aus beiden sein. Statische Erklärungen liefern grundlegende Informationen wie die wichtigsten Einflussfaktoren einer Entscheidung, folgen jedoch einem festen Schema und sind daher nicht individuell anpassbar. Interaktive Erklärungen ermöglichen es Nutzern, tiefergehende Informationen abzurufen oder spezifische Fragen zu stellen, was die Flexibilität erhöht. Ermöglicht wird das durch unter anderem interaktive Webschnittstellen und Dialogsysteme. Eine Kombination beider Ansätze bietet sowohl schnelle Einsicht als auch die Möglichkeit zur individuellen Anpassung (Sokol & Flach, 2019, S. 58).

Anwendungsbeispiel: Die statische Erklärung von GradeAI Solutions gibt eine allgemeine Bewertung, z.B.: „Die Gesamtbewertung basiert auf mehreren Kriterien, darunter Argumentationsstruktur, Kreativität und die Nutzung relevanter Quellen“. Mithilfe eines interaktiven Kommunikationsprotokolls könnte der Nutzer weitere Details anfordern, etwa: „Welche Aspekte der Argumentationsstruktur wurden als besonders überzeugend bewertet?“ oder „Wie wurden die verwendeten Quellen gewichtet?“.

ST6 Förderung von Neugier durch Erklärungen (Engl. Curiosity): Fördert das KI-System durch seine Erklärungen die Neugier der Nutzer, indem es Wissenslücken aufzeigt und zur aktiven Interaktion anregt? Neugier ist ein wichtiger Treiber für das Verlangen nach Erklärungen in KI-Systemen und nach schafft somit Vertrauen und motiviert die Nutzer zum weiterlernen, indem sie ihre Wissenslücken erkennen und diese durch verständliche und strukturierte Erklärungen schließen wollen. Erklärungen sollten zudem so gestaltet sein, dass sie nicht durch Überforderung abschreckend wirken (Hoffman et al., 2018, S. 16).

Anwendungsbeispiel: Lernende können auf dem Portal von GradeAI Solutions mit individuellen Erklärungen interagieren. Das Portal erkennt diese Interaktionen durch Benutzeranfragen und durch die Analyse von Verhaltensmustern und kann so Lernende nach Details fragen. Mithilfe dieser Rückschlüsse kann GradeAI Solutions feststellen, ob die Erklärung ausreicht oder ob sie anders gestaltet werden muss.

ST7 *Prägnanz der Erklärungen (Engl. Conciseness/Simplicity):* Sind die Erklärungen so selektiv und prägnant wie möglich gestaltet, um den Nutzer nicht mit unnötigen Informationen zu überlasten? Es ist für die Nutzer nicht nur von Bedeutung, dass die Erklärungen zugänglich und verständlich präsentiert werden, sondern auch dass sie schnell auf die relevanten Informationen zuzugreifen können, ohne von überflüssigen Details überwältigt zu werden. Die Erklärungen sollten so weit vereinfacht werden, ohne die essenziellen Informationen zu verlieren (Coroama & Groza, 2022, S. 405; Kulesza et al., 2015, S. 128; Sokol & Flach, 2019, S. 61).

Anwendungsbeispiel: Um die Nutzerfreundlichkeit der Erklärungen zu erhöhen, filtert GradeAI Solutions bekannte oder offensichtliche Fakten automatisch aus der Erklärung heraus. So erhält der Nutzer nur für ihn relevante Informationen, und wird nicht durch übeflüssige Details abgelenkt. So kann man bspw. in der Erklärung zur Rückmeldung einer Klausur offensichtliche oder bereits gut gemeisterte Punkte auslassen, wie z.B. das korrekte Zitieren.

ST8 *Überraschende Erklärungen (Engl. Novelty):* Zeigen die Erklärungen unerwartete oder überraschende Merkmale auf? Erklärungen sollten den Nutzern nicht nur Informationen liefern, die zu erwarten sind. Stattdessen sollten die Erklärungen überraschende oder ungewöhnliche Charakteristiken bieten, um die Aufmerksamkeit

der Nutzer auf interessante Aspekte zu lenken. Dieses Kriterium erfordert jedoch zwingend die Berücksichtigung des Hintergrundwissens des individuellen Nutzers, um auch wirklich eine überraschende und neuartige Erklärung bieten zu können (Sokol & Flach, 2019, S. 61).

Anwendungsbeispiel: GradeAI Solutions legt Wert darauf, dass die Erklärung zu einer falschen Lösung des Lernenden nicht nur generische und offensichtiche Gründe enthält, sondern auch bspw. Fehlermuster zeigt, die bereits in mehreren Klausuren vorgefallen sind. Das könnte bspw. eine Formulierungsschwäche sein, die seit längerem besteht. Somit hat der individuelle Schüler einen wichtigen Ansatzpunkt zur Verbesserung zukünftiger Klausuren.

ST9 ***Kontextaulisierte Erklärungen (Engl. Contextfullness):*** Liefert das KI-System Erklärungen, die in einem umfassenden Kontext eingebettet sind, sodass relevante Bedingungen, Einschränkungen und Faktoren berückstichtigt werden? Mithilfe von Erklärungen in einem größeren Zusammenhang können Nutzer nicht nur die unmittelbaren Entscheidungsgrundlagen nachvollziehen, sondern auch in Beziehung zu ähnlichen Fällen oder allgemeinen Muster setzen. Dadurch hat man ein tieferes Verständnis der zugrunde liegenden Kriterien, sieht systemische Zusammenhänge und kann Erklärungen besser im eigenen Anwendungsfall einordnen (Sokol & Flach, 2019, S. 60).

Anwendungsbeispiel: Lehrende erhalten gegebenfalls neben der Punktevergabe eine umfassende kontextualisierte Erklärung, die nicht nur die genauen Gründe für die Bewertung der einzelnen Aufgaben enthält, sondern auch darstellt, wie ähnliche Arbeiten in derselben Jahrgangsstufe oder im gleichen Themengebiet bewertet wurden.

ST10 ***Handlungsorientierte Erklärungen (Engl. Actionability):*** Zeigen die Erklärungen den Nutzern konkrete Handlungsoptionen auf, um gewünschte Ergebnisse zu erzielen? Eine handlungsorientierte Erklärung sollte klare Handlungsempfehlungen enthalten, anhand dessen der Nutzer den Entscheidungsprozess des Modells positiv beeinflussen kann. Dadurch wird die Erklärbarkeit nicht nur ein Mittel zur reinen Transparenz, sondern auch zur proaktiven Unterstützung der Nutzer. Ein Beispiel hierfür sind kontrafaktische Aussagen ("Wenn du X änderst, erreichst du Y") (Singh et al., 2024).

Anwendungsbeispiel: Bei der Rückmeldung zur Bewertung eines Schülers weist die Erklärung von GradeAI Solutions bspw. darauf hin, dass dem Text an Argumentationskraft mangelt. Die Erklärung geht jedoch weiter und gibt der Lehrkraft eine klare Handlungsempfehlung. Diese könnte lauten: „Die Argumentationsfähigkeit des Schülers könnte durch gezielte Übungen zur Strukturierung von Texten verbessert werden“.

4.3 Technische Dimension: Fokus auf die technologischen Grundlagen der Erklärbarkeit

Hier werden die technischen Aspekte der Erklärbarkeit betrachtet. Die technische Prüfung gewährleistet, dass die bereitgestellten Erklärungen nicht nur nachvollziehbar, sondern auch konsistent und robust sind.

T1 ***Herkunft der Erklärungen (Engl. Provenance):*** Legt das KI-System offen klar, was die Herkunft der Erkärung ist bzw. welche Informationen zur Erstellung der bereitgestellten Erklärungen verwendet wurden? Die Herkunft stammt meistens aus diesen drei Hauptquellen: modellbasiert, datenbasiert oder einer Kombination aus beiden. Modellbasierte Erklärungen basieren entweder auf einer Interaktion mit dem Black-Box-Modell oder nutzen die internen Parameter eines White-Box-Modells, wie bspw. Gewichtungen oder spezifische Entscheidungsmechanismen (Krause et al., 2016). Datenbasierte Erklärungen hingegen stützen sich auf relevante Datenpunkte, bspw. aus einem oder mehreren Trainings-, Evaluations- und Validierungsdatensätzen. Bei einer Erklärungen mit einer Kombination beider Quellen werden sowohl die Modellparameter als auch bestimmte Datenpunkte berücksichtigt. Ein KI-System sollte idealerweise jede Erklärung mit einer „Erklärbarkeitsspur“ begleiten, die dokumentiert, welche Trainingsdaten oder Modellparameter für eine Entscheidung besonders einflussreich waren (Koh & Liang, 2017). Diese Transparenz ist von Bedeutung, um Verzerrungen zu erkennen und die Verlässlichkeit der Erklärungen zu gewährleisten. Auditoren prüfen dabei, ob die Herkunft der Erklärung klar dokumentiert und verständlich präsentiert wird, sowie ob Nutzer die Quellen und Mechanismen nachvollziehen können (Sokol & Flach, 2019, S. 59).

Anwendungsbeispiel: Eine modellbasierte Erklärung könnte darauf hinweisen, dass 50% der Bewertung auf inhaltlicher Richtigkeit und 30% auf sprachlicher Qualität basiert. Eine datenbasierte Erklärung könnte sich auf den Vergleich mit drei ähnlichen Arbeiten stützen. Kombinierte Erklärungen würden diese beiden Ansätze zusammenführen, indem sie sowohl die Gewichtung der Kriterien als auch die Referenzarbeiten erläutern. Der jeweilige Prüfer geht der Frage nach, ob die Herkunft der Erklärung offen kommuniziert wird und ob die sie den Lernenden und Lehrenden ermöglichen, die Bewertungsschritte von GradeAI Solutions zu verstehen.

T2 ***Kompromiss von Erklärbarkeit und Leistung (Engl. Explainability Trade-off):*** Wurde die Entscheidung für oder gegen Erklärbarkeit in Relation zur Leistung ausreichend dokumentiert und begründet? Manche KI-Modelle können ohne signifikante Verluste in ihrer Leistung erklärt werden, während andere Systeme mit einem Verlust in der Vorhersagegenauigkeit oder Effizienz einhergehen, um eine höhere Erklärbarkeit zu gewährleisten (Moser, 2023, S. 33; Sokol & Flach, 2019, S. 63). Auditoren müssen hierbei prüfen, ob ein Kompromiss zwischen einer verbesserten

Erklärbarkeit und mehr Präzision vertretbar ist, abhängig von den Anforderungen der Nutzer und dem Anwendungszweck des Systems. Ebenfalls sollte geprüft werden, ob bei der Entscheidung zugunsten höherer Leistung Schritte zur Verbesserung der Erklärbarkeit des Modells vorgesehen oder bereits umgesetzt wurden (Rosenfeld, 2021, S. 47; Langer, Baum, et al., 2021; Sokol & Flach, 2019, S. 59).

Anwendungsbeispiel: Hat man sich für ein präzises neuronales Netzwerk entschieden, erhält der Nutzer zwar eine genaue Note, dafür jedoch schwerer verständliche Erklärungen. Alternativ könnte ein Entscheidungsbaum eingesetzt werden, der weniger genau aber besser nachvollziehbar ist. Das Team von GradeAI Solutions hat sich bereits für das neuronale Netzwerk entschieden, da Präzision bei der Bewertung von Lernergebnissen bzw. Klausuren oberste Priorität hat. Es sollte geprüft werden, ob der gewählte Kompromiss vertretbar ist und die eingeschränkte Erklärbarkeit transparent kommuniziert wird. Im Falle einer Entscheidung zugunsten der Präzision sollte geprüft werden, ob Maßnahmen zur Verbesserung der Erklärbarkeit des Modells geplant oder implementiert wurden, bspw. durch eine bessere Visualisierung der Bewertungen mithilfe von LIME.

T3 ***Schutz vor Informationlecks in Erklärungen (Engl. Information Leakage):*** Geben die Erklärungen des KI-Systems potenziell sensible Informationen über die Struktur oder Parameter des Modells oder die verwendeten Trainingsdaten preis? Das umfasst sowohl den Schutz von präzisen Schwellenwerten und Support-Vektoren als auch die Vermeidung der Offenlegung spezifischer Trainingsdatenpunkte und ist wichtig, um potenziellen Missbrauch oder Manipulationen des KI-Systems zu vermeiden. Sind geeignete Maßnahmen zur Begrenzung solcher Risiken implementiert, um Sicherheit und Datenschutz zu gewährleisten? Darunter zählt bspw. die Anonymsierung von Datenpunkten, die Verallgemeinerung von Schwellenwerten oder auch die Verwendung von ungenauen Beschreibungen (Sokol & Flach, 2019, S. 61).

Anwendungsbeispiel: Wird bspw. eine Matheprüfung von GradeAI Solutions bewertet, wäre die folgende Erklärung problematisch, da sie einen genauen Schwellenwert bzw. den genauen Bewertungsmaßstab offenlegt: “Die Antwort wurde als falsch bewertet, da die Genauigkeit unter 70% liegt”. Eine solche Erklärung könnte es dem Lernenden ermöglichen, das KI-System gezielt zu manipulieren oder Schwächen auszunutzen, indem bspw. gezielt versucht wird, die 70% zu erreichen, da eine höhere Leistung zum Bestehen nicht erforderlich ist. Stattdessen sollte die Erklärung weniger angreifbar formuliert sein, etwa so: “Die Antwort wurde mit einer Genauigkeit von 68 % bewertet, die unter den Mindestanforderungen liegt. Verbesserungen in den Rechenschritten könnten das Ergebnis verbessern”. So wird kein kritischer Schwellenwert preisgegeben. Auditoren sollten somit überprüfen, ob Erklärungen sensible Informationen vermeiden, gleichzeitig aber verständlich und hilfreich bleiben.

T4 ***Rahenwerk für die Datenverwaltung (Engl. Data-Governance-Framework):*** Ist ein Data-Governance-Framework implementiert, das den gesamten Lebenszyklus der verwendeten Daten berücksichtigt? Die Leistungsfähigkeit von KI-Modellen und Methoden wie LIME hängt maßgeblich von der Qualität der verwendeten Daten ab und ist somit die Grundlage für verständliche, zuverlässige und faire Erklärungen (FROX AG, 2024). Ein Data-Governance-Framework sollte klare Rollen und Verantwortlichkeiten für die Datenbestände definieren, bspw. durch die Benennung von Data-Ownern. Außerdem sollten Prozesse etabliert sein, um Datenqualitätsprobleme wie fehlende Werte, Duplikate oder inkonsistente Formate zu identifizieren und zu beheben. Die Diversität der Daten ist ebenfalls zu gewährleisten, damit alle Nutzergruppen berücksichtigt werden und die Erkärungen fair sind. Daten mit Qualitätsproblemen sollten entweder vollständig ausgeschlossen oder nach ihrer Korrektur erneut in den Datenbestand aufgenommen werden. Somit lassen sich mögliche Verzerrungen verhindern. Weiterhin ist es entscheidend, dass sämtliche Datentransformationsschritte von der Bearbeitung der Rohdaten bis zur Erstellung der Merkmale dokumentiert und begründet werden, um eine Transparenz sicherzustellen (Andrae, 2021, S. 43).

Anwendungsbeispiel: Ein Auditor überprüft, ob GradeAI Solutions den gesamten Lebenszyklus angemessen managt. So stellt ein Auditor bspw. fest, ob alle Schritte wie die Bereinigung der Daten oder die Transformation in ein einheitliches Format dokumentiert und verständlich begründet wurden, um unter anderem festzustellen, ob die Daten verzerrungensfrei sind. Wenn eine Prüfungsantwort aufgrund fehlender Daten nicht bewertet werden kann, sollte das für Lehrende und Lernende klar erkennbar und begründet sein.

T5 ***Wahrheitsgetreue Wiedergabe der Modellogik (Engl. Soundness):*** Spiegeln die bereitgestellten Erklärungen die interne Funktionsweise und Entscheidungslogik des zugrunde liegenden Modells korrekt und vollständig wieder? Eine hohe wahrheitsgetreue Wiedergabe der Modellogik stellt sicher, dass die Nutzer darauf vertrauen können, dass sie nicht nur verständliche, sondern auch wahrheitsgemäße Erklärungen bekommen. Dieses Kriterium ist besonders bei Post-Hoc-Erklärungen wichtig, da hier die Gefahr besteht, dass separate Erklärungsmodelle wie LIME die ursprüngliche Modelllogik des KI-Modells ungenau reproduzieren. Es wird oft gemessen durch den Vergleich der Outputs des Hauptmodells und des Erklärungsmodells, z.B. durch eine Berechnung der durchschnittlichen Rangkorrelation zwischen den beiden Modellen oder durch statische Kennzahlen wie Korrelation oder anderen Leistungsmetriken. Kurzgefasst zielt dieses Kriterium darauf ab, nichts als die Wahrheit zu sagen, und ist eins der der Maße für die Erklärbarkeitstreue gemäß Kulesza et al. (2015) (Kulesza et al., 2015, S. 127; Sokol & Flach, 2019, S. 60; Zhou et al., 2024, S. 11).

Anwendungsbeispiel: Auditoren haben zu prüfen, ob Erklärungen von GradeAI Solutions „unsound“ sind. Es kann passieren, dass eine Antwort als „falsch“ mit dieser Erklärung markiert wird: „Die Antwort war fehlerhaft, weil die Begründung unzureichend

war“. Wenn das zugrunde liegende Modell jedoch die Antwort aufgrund einer falschen Rechenoperation abgelehnt hat, ist die Erklärung „unsound“. Lehrende könnten dadurch zu falschen Schlüssen über die Bewertung kommen und Vertrauen in das System verlieren.

T6 ***Vollständigkeit der Erklärungen (Engl. Completeness)*****:** Liefert die Erklärung alle wesentlichen Informationen über die Daten und die Funktionsweise des zugrunde liegenden Modells? Vollständige Erklärungen bieten Nutzern einen umfassenden Überblick und ermöglichen mehr Vertrauen in das System, indem sie darüber informiert werden, welche Informationen dem KI-System zur Verfügung stehen und wie es diese nutzt. Die Vollständigkeit einer Erklärung zielt kurzgefasst darauf ab, die ganze Wahrheit zu sagen, und ist ein weiteres Maß für die Erklärbarkeitstreue gemäß Kulesza et al. (2015) (Chatti et al., 2022, S. 225; Kulesza et al., 2015, S. 127; Sokol & Flach, 2019, S. 60; Zhou et al., 2024, S. 11). Dieses Kriterium steht in enger Verbindung zu Kriterium ST7 (Prägnanz der Erklärungen) und erfordert eine sorgfältige Berücksichtigung beider Aspekte, da sie sich teilweise gegenseitig beeinflussen.

Anwendungsbeispiel: Die Erklärungen von GradeAI Solutions gehen über den Einzelfall hinaus. Sie erläutern den Lehrenden nicht nur die Bewertungslogik für eine spezifische Klausur, sondern zeigen auch, wie ähnliche Arbeiten aus verschiedenen Klassenstufen mit ähnlichen Kriterien bewertet wurden und wie diese einen Einfluss auf diese Bewertung haben. Eine Lehrkraft könnte bspw. sehen, dass das KI-System bestimmte grammatikalische Fehler strenger bewertet im Vergleich zu anderen Aspekten wie Argumentationsstruktur oder Schreibstil, da die Parallelklasse hierbei große Lücken aufweist. Durch diese umfassende Darstellung aller relevanten Informationen versteht die Lehrkraft, wie das System die verschiedenen Kriterien anwendet und wie einzelne Entscheidungen im größeren Bewertungsprozess eingebettet sind.

T7 Empfindlichkeit/Stabilität der Erklärungen (Engl. Sensitivity/Stability): Wie empfindlich reagiert die Erklärung des KI-Systems auf minimale Veränderungen der Eingabedaten? Eine robuste Erklärung sollte konsistent bleiben, selbst im Falle von geringfügigen und irrelevanten Änderungen an den Eingabewerten. Mithilfe dieses Kriteriums kann man beser beurteilen, ob die Erklärungen zuverlässig und stabil sind oder ob sie durch insignifikante Störungen verzert werden (Coroama & Groza, 2022; Rosenfeld, 2021).

Anwendungsbeispiel: Sollte ein Lernender ein irrelevantes Merkmal wie bspw. den Formatierungsstil ändern, sollte die neu generierte Erklärung von GradeAI Solutions konsistent bleiben. Falls die Erklärung bei solchen unwesentlichen Änderungen stark variiert, deutet das auf eine hohe Empfindlichkeit und damit auf eine potenziell instabile Erklärung hin. Bleibt die Erklärung jedoch unverändert oder zeigt nur marginale Anpassungen, wird das als positive Robustheit bewertet.

T8 Monotonie der Merkmalserklärungen (Engl. Monotonicity): Spiegeln die Erklärungen die relative Wichtigkeit der Eingabemerkmale korrekt wider? Die Merkmalsbeiträge (wie Eingabefaktoren) sollten monoton zur Modelleistung beitragen, wenn Merkmale nacheinander entsprechend ihrer Wichtigkeit hinzugefügt werden. Ein wichtiges Merkmal sollte immer mehr oder gleich viel zur Leistung beitragen als ein weniger wichtiges Merkmal, ansonsten liegt eine nicht-monotone Reihenfolge der Merkmalszuordnung vor (Coroama & Groza, 2022, S. 405).

Anwendungsbeispiel: Wenn ein Lernender nachvollziehen möchte, wie seine Leistung bzw. Bewertung zustande gekommen ist, kann GradeAI Solutions die Bedeutung einzelner Faktoren wie z.B. Antwortgenauigkeit (wichtigstes Merkmal), Schwierigkeitsgrad der Aufgabe und Bearbeitungszeit (am wenigsten wichtig) für die Gesamtbewertung in einer schrittweisen Erklärung bereitstellen. Hierfür fügt das System nacheinander die erklärungsrelevanten Merkmale in der Reihenfolge ihrer Wichtigkeit hinzu und berechnet für jedes Hinzufügen die zugehörige Modellleistung. Die Monotonie der Erklärung stellt sicher, dass die Modellleistung mit jedem hinzugefügten wichtigeren Merkmal logisch und konsistent steigt. Sollte die Reihenfolge der erklärten Wichtigkeit inkorrekt sein oder unerwartete Veränderungen wie ein Leistungsabfall auftreten, deutet das auf eine nicht-monotone Reihenfolge der Merkmalszuordnung hin und der Lernende könnte eine falsche Einschätzung der Bedeutung der Merkmale erhalten.

5 Schlussbetrachtung

Mit der rasanten Entwicklung im Bereich der KI gehen erhebliche Risiken einher, insbesondere wenn KI-Systeme in sensiblen Anwendunsbereichen intransparente Enscheidungen treffen.

Um solchen Problemen entgegenzuwirken, ist ein menschenzentrierter Ansatz erforderlich, der darauf abzielt, KI-Systeme so zu gestalten, dass Nutzer dem System vertrauen können. Hierbei ist die Erklärbarkeit von durch KI-Systemen generierten Enscheidungen von großer Bedeutung, die nicht nur Vertrauen schafft, sondern auch unter anderem eine Überprüfung auf Fehler in algorithmenbasierten Entscheidungen und das Einhalten von rechtlichen Bedingungen ermöglicht.

Eine Erklärung sollte immer dann verfügbar sein, wenn ein KI-System eine Entscheidung trifft, die das Leben eines betroffenen Menschen beeinträchtigen kann. Art. 6 KI-VO als auch Anh. III der KI-VO heben hierbei klar hervor, welche konkreten Anwendungsbereiche eine Transparenz bzw. Erklärbarkeit wünschen lassen und rechtlich voraussetzen.

Eines dieser Bereiche, die eine hinreichende Transparenz der KI-Systeme voraussetzt, ist die allgemeine oder berufliche Bildung gemäß Anh. III Abs. 3 der KI-VO und Art. 13 KI-VO. So handelt Kapitel drei dieser Bachelorarbeit rund um die Erklärbarkeit von KI-Systemen im Bildungswesen. Dies umfasst, wo bereits KI im Bildungswesen zum Einsatz kommt und in Zukunft noch Potenzial bietet, warum die Erklärbarkeit im Bildungskontext notwendig ist, wer die Empfänger der Erklärungen sind und wie deren Informationsgehalt ausgestaltet sein sollte sowie potenzielle Risiken im Zusammenhang mit Erklärbarer KI-Systeme im Bildungsbereich.

Zur Beantwortung der Forschungsfrage, anhand welcher Kriterien Auditoren die Erklärbarkeit von KI-Systemen überprüfen können und wie sich das entwickelte Prüfrahmenwerk im Kontext eines Anwendungsfalls im Bildungswesen veranschaulichen lässt, wurde ein Rahmenwerk mit sozialen, sozio-technischen und technischen Kriterien zusammengestellt. Diese Kriterien wie bspw. Stabilität oder Herkunft der Erklärungen bieten Auditoren eine strukturierte Grundlage, um KI-Systeme auf ihre Erklärbarkeit zu prüfen. Zur praktischen Demonstration wird das Anwendungsbeispiel aus Kapitel 3.5 herangezogen.

Entäuschend bei der Konzeption des Prüfrahmenwerkes war es, dass die KI-Verordnung mit Art. 13 zwar eine Transparenzpflicht mit sich bringt, jedoch nicht genau verstehen lässt, wie “Interpretierbarkeit” genau zu definieren ist und ob die europäische Regierungsbehörde die Begriffe der Transparenz und Erklärbarkeit gleichsetzt oder einen Unterschied dazwischen sieht (siehe Kapitel 3.2.5). Rechtlich gesehen muss beim Audit somit nicht geprüft werden, ob das KI-System erklärbar ist, sondern lediglich ob eine Transparenz vorliegt.

Obwohl eine ausführliche Liste mit Kriterien vorliegt, dient das Prüfrahmenwerk lediglich als mögliche Grundlage, Orientierungshilfe oder Inspiration, um mangelhafte Erklärbarkeit und damit verbundene potenzielle Risiken zu prüfen. Die Prüfer müssen je nach Anwendungsfall entscheiden, welche der ausgeführten Kriterien sie womöglich auslassen oder verfeinern wollen und ob sie neue Kriterien ergänzen wollen.

Dieses Prüfrahmenwerk könnte möglicherweise eine Diskussion darüber anstoßen, wie multidisziplinäre Erklärbarkeits-Auditierungsprozesse in der Praxis realisiert werden können, und als Schritt in Richtung der Etablierung von Qualitätsstandards (z.B. DIN-Normen) beitragen.

Anhang A

Eine vollständige Übersicht aller Kriterien sowie der dazugehörigen Audit-Fragen zur Messung der Erklärbarkeit von KI-Systemen, gegliedert nach den drei Dimensionen (Soziale, Sozio-technische und Technische Dimension)

Dimension	Kriterium	Audit-Frage	Quellen
Soziale Dimension	Funktionale Erklärbarkeit (Engl. Functionality)	Ist das KI-System so konzipiert, dass es von den Nutzern verstanden werden kann, oder bietet es Mechanismen wie Erklärbare KI-Modelle, um seine Entscheidungsprozesse nachträglich verständlich zu machen?	(Langer et al., 2021, S. 165)
	Vertrauensvolle/Treue Erklärbarkeit (Engl. Trust/Fidelity)	Werden Informationen bereitgestellt, die die Entscheidungsprozesse des KI-Systems genau und wahrheitsgemäß beschreiben?	(Hoffman et al., 2018, S. 19; Langer, Baum, et al., 2021, S. 165)
	Zufriedenheit mit Erklärungen (Engl. Satisfaction)	Sind die Erklärungen für den Nutzer subjektiv zufriedenstellend?	(Hoffman et al., 2018, S. 5)
	Verständlichkeit der Erklärungen (Engl. Comprehensibility)	Helfen die Erklärungen den Nutzern, die Entscheidungsprozesse des Systems besser nachzuvollziehen? Wie viel Aufwand ist nötig, damit Menschen die Erklärungen verstehen?	(Coroama & Groza, 2022, S. 405; Langer, Oster, et al., 2021, S. 14)
	Nichtdiskriminierung, Ethische Verantwortung und Rechenschaftspflicht (Engl. Non-Discrimination, Ethical Responsibility and Accountability)	Ermöglicht die Erklärbarkeit des KI-Systems eine verantwortungsvolle Entscheidungsfindung und eine klare Zuweisung von Verantwortung für getroffene Entscheidungen?	(Noerr PartGmbB, 2024; Andrae, 2021, S. 41; Langer, Baum, et al., 2021, S. 166)
	Einhaltung der Anforderungen der DSGVO im Kontext der Erklärbarkeit von KI-Systemen (Engl. Compliance with the requirements of the GDPR)	Entspricht die Datenverarbeitung den datenschutzrechtlichen Vorgaben?	(DSGVO, 2016, Art. 5-6, Art. 12-14, Art. 22)
	Einhaltung der Transparenzanforderungen der KI-VO im Kontext der Erklärbarkeit von KI-Systemen (Art. 13 KI-VO) (Engl. Compliance with the requirements of the AI Act)	Ist das Hochrisiko-KI-System hinreichend transparent gestaltet?	(KI-VO, 2024, Art. 13)
Sozio-technische Dimension	Klärung der Verantwortlichen durch Erklärbarkeit (KI-Haftungsrichtlinie) (Engl. Liability)	Ist eine fundamentale Erklärbarkeit gegeben, mit welcher sich die Verantwortlichen entlang der gesamten Wertschöpfungskette des KI-Systems klären lassen?	(Noerr PartGmbB, 2024; Schreiber & Dimov, 2024)
	Erklärungsdomäne (Engl. Explanation Domain)	Werden die bereitgestellten Erklärungen entweder im ursprünglichen Format der Modellparameter und Datenmerkmale (Originaldomäne) bereitgestellt, oder in einer eingewandelten, verständlicheren Form (Transformationsdomäne)?	(Sokol & Flach, 2019, S. 58)
	Kausale und Handlungsorientierte Erklärungen (Engl. Causality vs. Actionability)	Wird dem Nutzer des KI-Systems deutlich kommuniziert, ob kausale oder handlungsorientierte Erklärungen verwendet werden, um Missverständnisse und falsche Schlussfolgerungen zu vermeiden und sicherzustellen, dass keine kausale Verbindung irrtümlich angenommen wird?	(Sokol & Flach, 2019, S. 59)
	Anpassungsfähigkeit der Erklärungen (Engl. Adaptability)	Sind die Erklärungen an die unterschiedlichen Bedürfnisse und das Hintergrundwissen der jeweiligen Stakeholder/Nutzer angepasst?	(Langer, Baum, et al., 2021, S. 165; Schneider & Handali, 2019; Walton, 2007)
	Erklärungsmedium (Engl. Explanatory Medium)	Bietet das KI-System Erklärungen in einem Medium an, das für die Zielgruppe geeignet ist und die Art der Information verständlich vermittelt, ohne die Nutzer zu überfordern oder wichtige Details auszulassen?	(Sokol & Flach, 2019, S. 58)
	Interaktive Kommunikation mit Erklärungen (Engl. Interactivity)	Können Nutzer mit den bereitgestellten Erklärungen interagieren? Kann das KI-System auf spezifische Nachfragen eingehen und bei Bedarf zusätzliche Informationen oder eine tiefere Analyse bereitstellen?	(Sokol & Flach, 2019, S. 58)
	Förderung von Neugier durch Erklärungen (Engl. Curiosity)	Fördert das KI-System durch seine Erklärungen die Neugier der Nutzer, indem es Wissenslücken aufzeigt und zur aktiven Interaktion anregt?	(Hoffman et al., 2018, S. 16)
	Prägnanz der Erklärungen (Engl. Conciseness/Simplicity)	Sind die Erklärungen so selektiv und prägnant wie möglich gestaltet, um den Nutzer nicht mit unnötigen Informationen zu überlasten?	(Coroama & Groza, 2022, S. 405; Sokol & Flach, 2019, S. 61)
	Überraschende Erklärungen (Engl. Novelty)	Zeigen die Erklärungen unerwartete oder überraschende Merkmale auf?	(Sokol & Flach, 2019, S. 61)
	Kontextualisierte Erklärungen (Engl. Contextfullness)	Liefert das KI-System Erklärungen, die in einem umfassenden Kontext eingebettet sind, sodass relevante Bedingungen, Einschränkungen und Faktoren berücksichtigt werden?	(Sokol & Flach, 2019, S. 60)
	Handlungsorientierte Erklärungen (Engl. Actionability)	Zeigen die Erklärungen den Nutzern konkrete Handlungsoptionen auf, um gewünschte Ergebnisse zu erzielen?	(Singh et al., 2024)
Technische Dimension	Herkunft der Erklärungen (Engl. Provenance)	Legt das KI-System offen klar, was die Herkunft der Erklärungen ist bzw. welche Informationen zur Erstellung der bereitgestellten Erklärungen verwendet wurden?	(Krause et al., 2016; Kim & Lang, 2017; Sokol & Flach, 2019, S. 59)
	Kompromiss von Erklärbarkeit und Leistung (Engl. Explainability Trade-off)	Wurde die Entscheidung für oder gegen Erklärbarkeit in Relation zur Leistung ausreichend dokumentiert und begründet?	(Moser, 2023, S. 33; Rosenfeld, 2021; Langer, Baum, et al., 2021; Sokol & Flach, 2019, S. 60)
	Schutz vor Informationlecks in Erklärungen (Engl. Information Leakage)	Gaben die Erklärungen des KI-Systems potenziell sensible Informationen über die Struktur oder Parameter des Modells oder die verwendeten Trainingsdaten preis? Sind geeignete Maßnahmen zur Begrenzung solcher Risiken implementiert, um Sicherheit und Datenschutz zu gewährleisten?	(Sokol & Flach, 2019, S. 61)
	Rahmenwerk für die Datenverwaltung (Engl. Data-Governance-Framework)	Ist ein Data-Governance-Framework implementiert, das den gesamten Lebenszyklus der verwendeten Daten berücksichtigt?	(FROX-AG, 2024; Andrae, 2021, S. 43)
	Wahrheitsgetreue Wiedergabe der Modelllogik (Engl. Soundness)	Spiegeln die bereitgestellten Erklärungen die interne Funktionsweise und Entscheidungslogik des zugrunde liegenden Modells korrekt und vollständig wieder?	(Kulesza et al., 2015, S. 127; Sokol & Flach, 2019, S. 60; Zhou et al., 2024, S. 11)
	Vollständigkeit der Erklärungen (Engl. Completeness)	Liefert die Erklärung alle wesentlichen Informationen über die Daten und die Funktionsweise des zugrunde liegenden Models?	(Chatti et al., 2022, S. 225; Kulesza et al., 2015, S. 127; Sokol & Flach, 2019, S. 60; Zhou et al., 2024, S. 11)
	Empfindlichkeit/Stabilität der Erklärungen (Engl. Sensitivity/Stability)	Wie empfindlich reagiert die Erklärung des KI-Systems auf minimale Veränderungen der Eingabedaten?	(Coroama & Groza, 2022; Rosenfeld, 2021)
	Monotonie der Merkmalserklärungen (Engl. Monotonicity)	Spiegeln die Erklärungen die relative Wichtigkeit der Eingabemerkmale korrekt wider?	(Coroama & Groza, 2022, S. 405)

Eigene Abbildung

Literaturverzeichnis

Alpaydın, E. (2022). *Maschinelles Lernen* (3. Aufl.). De Gruyter Oldenbourg. https://doi.org/10.1515/9783110740196

Andrae, K. (2021, Januar 18). *Erklärbare Künstliche Intelligenz: Tools und Methoden für erklärbare KI-Modelle*. https://phaidra.fhstp.ac.at/detail/o:4486

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI* (No. arXiv:1910.10045). arXiv. https://doi.org/10.48550/arXiv.1910.10045

Authority of the House of Lords. (2018). *AI in the UK: ready, willing and able*. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Avid Center. (2023). *Khanmigo as an AI Personal Tutor and Assistant*. AVID Open Access. https://avidopenaccess.org/resource/khanmigo-as-an-ai-personal-tutor-and-assistant/

Baranzke, Dr. H. (2021, Oktober 20). *„Hans, das denkende Pferd von Elberfeld“* (Wuppertal). http://www.uni-wuppertal.de/de/news/detail/hans-das-denkende-pferd-von-elberfeld/

Barilla, G. (2024, März 1). *A Closer Look at AI: Data Mining Techniques*. Akkio. https://www.akkio.com/post/ai-data-mining

Bronner, Dr. P. (2024, Oktober 24). *ChatGPT & KI im Unterricht*. https://www.patrickbronner.de/KI

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Chatti, M. A., Guesmi, M., Vorgerd, L., Ngo, T., Joarder, S., Ain, Q. U., & Muslim, A. (2022). Is More Always Better? The Effects of Personal Characteristics and Level of Detail on the Perception of Explanations in a Recommender System. *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 254–264. https://doi.org/10.1145/3503252.3531304

Chaushi, A., Selimi, B., Chaushi, A., & Apostolova, M. (2024). *Explainable Artificial Intelligence in Education: A Comprehensive Review* (2. Aufl., Bd. 2153). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63787-2

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2020). *Toward Personalized XAI: A Case Study in Intelligent Tutoring Systems* (No. arXiv:1912.04464). arXiv. https://doi.org/10.48550/arXiv.1912.04464

Coroama, L., & Groza, A. (2022). Evaluation Metrics in Explainable Artificial Intelligence (XAI). In *Advanced Research in Technologies, Information, Innovation and Sustainability* (S. 401–413). https://link.springer.com/content/pdf/10.1007/978-3-031-20319-0.pdf

D16-Digitalminister. (2022). *Positionierung der Länder gegenüber der geplanten KI-Verordnung der Europäischen Union*. https://im.baden-wuerttemberg.de/fileadmin/redaktion/m-im/intern/dateien/pdf/20221212_Positionierung_der_Laender_gegen%C3%BCber_der_geplanten_KI-Verordnung_der_Europaeischen_Union.pdf

Döbel, I., Leis, Dr. M., Vogelsang, M., Neustroev, D., Petzka, Dr. H., Riemer, A., Rüping, Dr. S., Voss, Dr. A., Wegele, M., & Welz, Dr. J. (2018). *Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung* (S. 56). Fraunhofer-Gesellschaft. https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf

DSGVO. (2016). *Datenschutz-Grundverordnung (DSGVO)*. https://dsgvo-gesetz.de/kapitel-3/

Duden, „erklären“. (o. J.). Abgerufen 20. Oktober 2024, von https://www.duden.de/rechtschreibung/erklaeren

Ebers, M., Hoch-Loy, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). Der Entwurf für eine EU-KI-Verordnung: Richtige Richtung mit Optimierungsbedarf. *Recht Digital*, *1*(11), Article 11.

Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., & Van Gerven, M. (Hrsg.). (2018). *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98131-4

EU AI Act. (2024). *Abschnitt 2: Anforderungen an KI-Systeme mit hohem Risiko | EU-Gesetz über künstliche Intelligenz*. https://artificialintelligenceact.eu/de/section/3-2/

Europäische Kommission. (2024, September 25). *KI-Gesetz | Gestaltung der digitalen Zukunft Europas*. https://digital-strategy.ec.europa.eu/de/policies/regulatory-framework-ai

Fetzer, J. H. (1990). What is Artificial Intelligence? In J. H. Fetzer (Hrsg.), *Artificial Intelligence: Its Scope and Limits* (S. 3–27). Springer Netherlands. https://doi.org/10.1007/978-94-009-1900-6_1

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data* (1. Aufl.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973000

Fraunhofer IOSB. (2023, September 10). *Adaptive Lernsysteme*. https://www.iosb.fraunhofer.de/de/kompetenzen/systemtechnik/interoperabilitaet-assistenzsysteme/forschungsthemen/adaptive-lernsysteme.html

FROX AG. (2024, Juli 12). Erfolgsfaktoren für KI im ITSM mit ServiceNow: Daten, Ethik und Mitarbeitende. *FROX*. https://www.frox.ch/newsroom/blog-artikel/erfolgsfaktoren-ki-im-itsm/

Future of Life Institute. (2024). *Artikel 50: Transparenzverpflichtungen für Anbieter und Betreiber bestimmter KI-Systeme | EU-Gesetz über künstliche Intelligenz*. https://artificialintelligenceact.eu/de/article/50/

Generaldirektion Kommunikation. (2024). *KI-Verordnung tritt in Kraft—Europäische Kommission*. https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_de

Gillis, A. S., & Moore, J. (2023, März). *Knowledge-based systems (KBSes)*. Knowledge-Based Systems (KBSes). https://www.techtarget.com/searchcio/definition/knowledge-based-systems-KBS

Giosmin, E. (2023, September). *Interpretability vs explainability: Understanding the Differences and Importance in the World of Artificial Intelligence*. https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence/

Grammarly Inc. (2024). *How Grammarly Works | Grammarly*. https://www.grammarly.com/how-grammarly-works

Grand View Research. (2024). *Artificial Intelligence Market Size & Outlook*. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market

Gunning, D., & Aha, D. W. (2021). DARPA's Explainable Artificial Intelligence Program. *Applied AI Letters*. https://doi.org/10.1002/ail2.61

Harwardt, M., & Köhler, M. (2023). Künstliche Intelligenz entlang der Customer Journey. In *Künstliche Intelligenz entlang der Customer Journey: Einsatzpotenziale*

von KI im E-Commerce (S. 21–29). Springer Fachmedien. https://doi.org/10.1007/978-3-658-39109-6_3

Hau, D. (2023, Februar 24). *Auswirkungen von Künstlicher Intelligenz auf den Bildungsbereich*. https://www.science.lu/de/ki-der-bildung/auswirkungen-von-kuenstlicher-intelligenz-auf-den-bildungsbereich

HEG-KI. (2019). *Ethik-Leitlinien für eine vertrauenswürdige KI*. https://op.europa.eu/de/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Heynike, F. (2024, Juli 16). *AI Act: Die EU möchte die Risiken von KI in den Griff bekommen*. KPMG-Law. https://kpmg-law.de/der-ai-act-kommt-eu-moechte-risiken-von-ki-in-den-griff-bekommen/

Hoffman, R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. *ArXiv*. https://www.semanticscholar.org/paper/Metrics-for-Explainable-AI%3A-Challenges-and-Hoffman-Mueller/be711f681580d3a02c8bc4c4dab0c7a043f4e1d2

Höllmer, J. (2023, August 17). Moodle Chatbot—KI als persönlicher Uni-Dozent. *scieneers*. https://www.scieneers.de/moodle-chatbot/

Humanwissenschaftilche Fakutltät der Universität Potsdam. (2023, Mai). *KI-Toolbox—Für innovativen und motivierenden Unterricht*. https://www.uni-pots-dam.de/fileadmin/projects/schulnetzwerkdigital/Dokumente/Dokumente_DB/2023_KI_ToolBox.pdf

Hwang, G.-J., Chu, H.-C., & Yin, C. (2017). Objectives, methodologies and research issues of learning analytics. *Interactive Learning Environments*, *25*(2), 143–146. https://doi.org/10.1080/10494820.2017.1287338

Jochum, K. (2020, August 20). *Prüfungsnoten: Algorithmus verstärkt soziale Ungleichheiten*. https://www.inside-it.ch/post/pruefungsnoten-algorithmus-verstaerkt-soziale-ungleichheiten-20200820, https://www.inside-it.ch/post/pruefungsnoten-algorithmus-verstaerkt-soziale-ungleichheiten-20200820

John Wiley & Sons, Inc. (2024). *Knewton Adaptive Learning and How it's Unique*. https://support.knewton.com/s/article/Knewton-Adaptive-Learning-and-How-it-s-Unique

Johnston, J. (2024, August 17). *The History of AI: From Rules-based Algorithms to Generative Models*. Lantern. https://lanternstudios.com/insights/blog/the-history-of-ai-from-rules-based-algorithms-to-generative-models/

Kahoot! (o. J.). *What is Kahoot! | How to play Kahoot!* Kahoot! Abgerufen 25. Oktober 2024, von https://kahoot.com/what-is-kahoot/

Kahoot! (2024, Oktober 19). *How to use Kahoot! AI tools*. Kahoot! Help & Resource Center. https://support.kahoot.com/hc/en-us/articles/17152945038355-How-to-use-Kahoot-AI-tools

Khan Academy. (2024a). *Khanmigo for learners: Always-available tutor, powered by AI*. https://khanmigo.ai/learners

Khan Academy. (2024b). *Khanmigo for teachers: Free, AI-powered teacher assistant*. https://khanmigo.ai/teachers

Khosravi, P., Liang, Y., Choi, Y., & Broeck, G. V. den. (2019). *What to Expect of Classifiers? Reasoning about Logistic Regression with Missing Features* (No. arXiv:1903.01620). arXiv. https://doi.org/10.48550/arXiv.1903.01620

KI-VO. (2024). *KI-Verordnung (KI-VO)*. https://ai-act-law.eu/de/

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. *Proceedings of the 34th International Conference on Machine Learning*, 1885–1894. https://proceedings.mlr.press/v70/koh17a.html

Kolkman, D. (2020, August 26). “F**k the algorithm”?: What the world can learn from the UK’s A-level grading fiasco. *Impact of Social Sciences*. https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

KPMG. (2023). *Trust in artificial intelligence*. KPMG. https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-in-artificial-intelligence.html

Kramer, S. (2019). *Educational Data Mining: Möglichkeiten und Unmöglichkeiten*. https://www.lob.uni-mainz.de/files/2019/10/LOB-Tagung_2019_Vortrag_Kramer.pdf

Kraus, Dr. T., Ganschow, L., Eisenträger, M., & Wischmann, Dr. S. (2021). *Erklärbare KI: Anforderungen, Anwendungsfälle und Lösungen*. Technologieprogramm KI-Innovationswettbewerb des Bundesministeriums für Wirtschaft und Energie. https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.pdf?__blob=publicationFile

Krause, J., Perer, A., & Bertini, E. (2016). *Using Visual Analytics to Interpret Predictive Machine Learning Models* (No. arXiv:1606.05685). arXiv. https://doi.org/10.48550/arXiv.1606.05685

Krieg, Prof. A. (2023). Proof of Concept – Zur Anbindung von Moodle an ChatGPT. *Hochschulforum Digitalisierung*. https://hochschulforumdigitalisierung.de/proof-of-concept-zur-anbindung-von-moodle-an-chatgpt/

Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*(3), 633.

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *Proceedings of the*

20th International Conference on Intelligent User Interfaces, 126–137. https://doi.org/10.1145/2678025.2701399

Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. (2021). Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 164–168. https://doi.org/10.1109/REW53955.2021.00030

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). *What Do We Want From Explainable Artificial Intelligence (XAI)? -- A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research* (No. arXiv:2102.07817). arXiv. https://doi.org/10.48550/arXiv.2102.07817

Laudien, A. (2024, September 30). *KI-Aufsicht: Sind die Datenschutzbehörden jetzt raus?* Dr. Datenschutz. https://www.dr-datenschutz.de/ki-aufsicht-sind-die-datenschutzbehoerden-jetzt-raus/

Lipton, Z. C. (2018). *The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.* https://doi.org/10.1145/3236386.3241340

Martini, M. (2019). *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz*. Springer. https://doi.org/10.1007/978-3-662-59010-2

Merriam-Webster, Inc. (2024, Oktober 6). *Definition of ARTIFICIAL INTELLIGENCE*. https://www.merriam-webster.com/dictionary/artificial+intelligence

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen. (2023). *Umgang mit textgenerierenden KI-Systemen* (S. 12). Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen.

Mollick, E. (2024). *Using AI Tutors to Flip Your Classroom*. https://www.edutopia.org/article/using-ai-tutors-flipped-classroom/

Moser, M. (2023). *Die Erklärbarkeit und Erklärungsbedürftigkeit algorithmenbasierter Entscheidungen: Unionsrechtliche Entwicklungen im Bereich der "explainable AI“*. http://diglib.uibk.ac.at/ulbtirolhs/8498954

Nicoletti, L., & Bass, D. (2023, September 6). *Humans Are Biased. Generative AI Is Even Worse*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Noerr PartGmbB. (2024, September 20). *KI-Haftungsrichtlinie: Studie des Europäischen Parlaments zur Haftung für KI*. https://www.noerr.com/de/insights/ki-haftungsrichtlinie-studie-des-eu-parlaments-zur-haftung-fuer-ki

Norddeutscher Rundfunk. (2023, Mai 27). *Abitur in Hamburg: Künstliche Intelligenz zum Schummeln genutzt*. https://www.ndr.de/nachrichten/hamburg/Abitur-in-Hamburg-Kuenstliche-Intelligenz-zum-Schummeln-genutzt,abitur540.html

Páez, N. (2023). *An algorithm to combat school dropout in Argentina | The UNESCO Courier*. An Algorithm to Combat School Dropout in Argentina. https://courier.unesco.org/en/articles/algorithm-combat-school-dropout-argentina

Rachha, A., & Seyam, M. (2023). *Explainable AI In Education: Current Trends, Challenges, And Opportunities*. 232–239. https://doi.org/10.1109/SoutheastCon51012.2023.10115140

Raza, M. (2024, Juli 23). *Explainable vs. Interpretable Artificial Intelligence*. Splunk. https://www.splunk.com/en_us/blog/learn/explainability-vs-interpretability.html

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *„Why Should I Trust You?“: Explaining the Predictions of Any Classifier* (No. arXiv:1602.04938). arXiv. https://doi.org/10.48550/arXiv.1602.04938

Richter, A., Gačić, T., Kölmel, B., & Waidelich, L. (2019). Künstliche Intelligenz und potenzielle Anwendungsfelder im Marketing. In Deutscher Dialogmarketing Verband e.V. (Hrsg.), *Dialogmarketing Perspektiven 2018/2019: Tagungsband 13. Wissenschaftlicher interdisziplinärer Kongress für Dialogmarketing* (S. 31–52). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25583-1_2

Rieder, G., & Simon, J. (2017). Big Data: A New Empiricism and its Epistemic and Socio-Political Consequences. In W. Pietsch, J. Wernecke, & M. Ott (Hrsg.), *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data* (S. 85–105). Springer Fachmedien. https://doi.org/10.1007/978-3-658-12153-2_4

Rosenfeld, A. (2021, April). Better Metrics for Evaluating Explainable Artificial Intelligence: Blue Sky Ideas Track. *ResearchGate*. https://www.researchgate.net/publication/349111351_Better_Metrics_for_Evaluating_Explainable_Artificial_Intelligence_Blue_Sky_Ideas_Track

Ruhland, C., Schnuecker, A., Shegupta, U., Seegerer, S., & Meissner, R. (2023). *Learning Analytics für Lehrkräfte*. https://doi.org/10.35542/osf.io/dvkgx

Samek, W., & Müller, K.-R. (2019). *Towards Explainable Artificial Intelligence* (No. arXiv:1909.12072). arXiv. https://doi.org/10.48550/arXiv.1909.12072

Schmid, Dr. U., Blanc, Dr. B., Hochbauer, M., & Wellert, L. (2023). *Schule und KI - Ein praxisorientierter Leitfaden* (S. 66). mmb Institut GmbH. https://www.telekom-stiftung.de/sites/default/files/files/Leitfaden-Schule-und-KI.pdf

Schmid, Dr. U., Blanc, Dr. B., Toepel, M., Pinkwart, Prof. Dr. N., & Drachsler, Prof. Dr. H. (2021). *KI@Bildung: Lehren und Lernen in der Schule mit Werkzeugen Künstlicher Intelligenz* (S. 58). mmb Institut – Gesellschaft für Medien- und

Kompetenzforschung mbH. https://www.telekom-stif-tung.de/sites/default/files/files/media/publications/KI%20Bildung%20Schlussbericht.pdf

Schneider, J., & Handali, J. P. (2019, Mai). (PDF) PERSONALIZED EXPLANATION FOR MACHINE LEARNING: A CONCEPTUALIZATION. *ResearchGate*. https://www.researchgate.net/publication/333092516_PERSONALIZED_EXPLANATION_FOR_MACHINE_LEARNING_A_CONCEPTUALIZATION

Schreiber, M., & Dimov, V. (2024, November 14). *KI und Haftung in der Praxis: Ein Überblick*. HÄRTING Rechtsanwälte. https://haerting.de/wissen/ki-und-haftung-in-der-praxis-ein-ueberblick/

Seward, C. (2015, Dezember 15). *Accelerating Warehouse Operations with Neural Networks*. https://engineering.zalando.com/posts/2015/12/../../../posts/2015/12/accelerating-warehouse-operations-with-neural-networks.html

Singh, R., Miller, T., Sonenberg, L., Velloso, E., Vetere, F., Howe, P., & Dourish, P. (2024). *An Actionability Assessment Tool for Explainable AI* (No. arXiv:2407.09516). arXiv. https://doi.org/10.48550/arXiv.2407.09516

Sokol, K., & Flach, P. (2019). *Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches* (No. arXiv:1912.05100). arXiv. https://doi.org/10.48550/arXiv.1912.05100

Sperl, A. (2024, Februar 15). *KI-Tools für das wissenschaftliche Arbeiten*. https://www.fernuni-hagen.de/zli/blog/ki-tools-fuer-das-wissenschaftliche-arbeiten/

Stanford Education. (2021, November 20). *Chatbot as a teaching tool*. IT Teaching Resources. https://teachingresources.stanford.edu/resources/chatbot-as-a-teaching-tool/

Staube, B., Menkemann, S., Kurz, M., & Gülcan, C. (2012, Dezember 19). *Vor- und Nachteile von Lernspielen*. https://www.online-lernprogramme.de/lernprogramme-fuer-kinder/vor-und-nachteile-von-lernspielen/

Theodotou, Dr. M. (2024, Dezember 1). *2024 Trends: AI-Powered Learning Analytics*. https://elearningindustry.com/2024-trends-ai-powered-learning-analytics

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems* (No. arXiv:1806.07552). arXiv. https://doi.org/10.48550/arXiv.1806.07552

Turnitin, LLC. (2024). *Gradescope | Moderne Benotungs- & Bewertungsplattform | Turnitin*. https://www.turnitin.de/products/similarity/

TutorSpace GmbH. (2024, August 26). *Alena—Deine persönliche Lernhelferin*. Alena. https://lerne-mit-alena.de/

Verch, Prof. Dr. U. (2024, Juli). *Anspruchsvolle Regeln für vertrauenswürdige Künstliche Intelligenz*. HochschulefürAngewandteWissenschaftenHamburg. https://journals.sub.uni-hamburg.de/hup3/apimagazin/article/download/213/231/2005

Vodafone Stiftung Deutschland gGmbH. (2023). *Aufbruch ins Unbekannte: Schule in Zeiten von künstlicher Intelligenz und ChatGPT* (S. 24). Vodafone Stiftung Deutschland gGmbH.

Walton, D. (2007, Juli 22). Dialogical Models of Explanation. *ResearchGate*. https://www.researchgate.net/publication/221622830_Dialogical_Models_of_Explanation

Weik, H. (2021, April 11). Maschinelle Lernverfahren erklärbar machen. *interaktiv online*. https://interaktiv.ipa.fraunhofer.de/kuenstliche-intelligenz-und-maschinelles-sehen/maschinelle-lernverfahren-erklaerbar-machen/

Wenig, V. (2024, Februar 18). *Der Lehrer aus dem Internet: Mit Künstlicher Intelligenz zum Mathe-Abitur*. https://www.berliner-zeitung.de/mensch-metropole/ki-lehrerin-alena-mit-kuenstlicher-intelligenz-zum-mathe-abitur-li.2186976

Xaqt, Inc. (2023). *Mastering Rule-Based Systems*. Mastering Rule-Based Systems: Implementation, Benefits, and Best Practices. https://www.xaqt.com/blog/mastering-rule-based-systems/

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zhou, J., Gandomi, A., & Holzinger, A. (2024). (PDF) Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *ResearchGate*. https://doi.org/10.3390/electronics10050593

Žlahtič, B., Završnik, J., Blažun Vošner, H., & Kokol, P. (2024). Transferring Black-Box Decision Making to a White-Box Model. *Electronics*, *13*(10), Article 10. https://doi.org/10.3390/electronics13101895

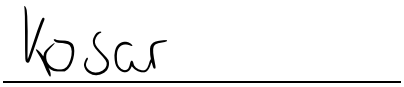
Zweig, K. A. (2018). Wo Maschinen irren können: Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung. *Arbeitspapier Algorithmenethik*. https://doi.org/10.11586/2018006

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht. Es wurde ChatGPT zur Unterstützung des Schreibprozesses verwendet. Die Nutzung umfasste die Überarbeitung der Grammatik, des Schreibstils und der Rechtschreibung. Es wird ausdrücklich darauf hingewiesen, dass die endgültige Verantwortung für die inhaltliche Richtigkeit, die kritische Reflexion und die Interpretation der Ergebnisse beim Autor dieser Arbeit liegt. Die KI diente lediglich als Werkzeug und nicht als Ersatz für das kritische und analytische Denken des Autors. Die vorgelegte Arbeit hat weder in der gegenwärtigen noch in einer anderen Fassung schon einem anderen Fachbereich der Hochschule Ruhr West oder einer anderen wissenschaftlichen Hochschule vorgelegen.

Düsseldorf, 12.12.2024

Ort, Datum



Unterschrift