Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Nachvollziehbare KI

Erklären, für wen, was und wofür

Gefördert vom

Bundesministerium
für Forschung, Technologie
und Raumfahrt

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN

WHITEPAPER

Samek, W., Schmid, U. et al.
AG Technologische Wegbereiter
und Data Science

Inhalt

Empfohlene Zitierweise
Samek, W., Schmid, U. et al. (2025):
Nachvollziehbare KI: Erklären, für wen, was und wofür.
DOI: https://doi.org/10.48669/pls_2025-2

Zusammenfassung

KI-Systeme mit hoher Komplexität sind häufig sogenannte Black Boxes. Wie sie zu ihren Ergebnissen kommen, ist oft schwer nachvollziehbar. In vielen Anwendungsgebieten, wie der medizinischen Diagnostik, der Kreditbewilligung im Finanzwesen oder der Qualitätskontrolle in der Produktion, ist die Nachvollziehbarkeit jedoch entscheidend, um die Ergebnisse einordnen und hinterfragen zu können. So erhalten Entwickelnde Informationen, um KI-Systeme zu verbessern, und Verbrauchende können in Erfahrung bringen, welche Faktoren für die Kreditentscheidung einer Bank einschlägig waren. Die Beispiele zeigen, dass Erklärungen von KI-Entscheidungen für verschiedene Zielgruppen verständlich sein müssen.

Vor diesem Hintergrund widmen sich Expertinnen und Experten der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme im vorliegenden Whitepaper dem Thema der „Nachvollziehbaren KI" entlang folgender Fragestellung: Wem soll was, wie und wozu erklärt werden?

Zunächst werden zentrale Begriffe zum Themengebiet bestimmt und eingeordnet, um dann auf zwei zentrale Zielsetzungen von erklärbarer KI (XAI) näher einzugehen: (1) zur Vertrauenswürdigkeit von KI-Systemen beizutragen und (2) ihre Qualität zu erhalten und zu verbessern. Anschließend wird dargelegt, dass eine adäquate Form der Erklärung entscheidend ist, um den verschiedenen Zielgruppen und ihren Bedürfnissen gerecht zu werden. Mit welchen XAI-Methoden dies erfolgen kann, wird in einem Überblick aufgezeigt. Dabei geht das Paper auch genauer auf neuere Trends wie XAI für generative KI und interaktive erklärbare KI ein.

Anhand von sieben verschiedenen Personae – von der KI-Spezialistin bis hin zum einfachen Anwender – wird illustriert, wie unterschiedlich die Konstellation aus individuellen Charakteristika wie etwa Vorwissen und Zielsetzung sowie Formen und Methoden der Erklärung ausfallen kann.

Um die weitere Entwicklung dieser KI-Technologie voranzutreiben und ihr Potenzial noch besser auszuschöpfen, werden allgemeine sowie speziell auf die Zielgruppenorientierung ausgerichtete Gestaltungsoptionen vorgeschlagen. So sollte die Forschung etablierte Methoden verbessern und XAI-Methoden für neue Arten von KI (weiter-)entwickeln. Möglich sind Instrumente zur Inspizier- und Kontrollierbarkeit großer KI-Modelle oder Standardwerkzeugkästen für eine Modellkorrektur ohne erneutes Training. In der Lehre sollte XAI im Sinne eines Engineering-Tools als Teil von AI Engineering stärker in KI- und Data-Science-Studiengängen verankert werden. Unternehmen könnten vermehrt auf XAI setzen, um beispielsweise etwa interne Kommunikationshürden abzubauen und sich von Wettbewerbern abzugrenzen.

1 Einleitung

Mit der Verbreitung von Chatbots (wie ChatGPT, Copilot und Mistral's Le Chat) im Zuge der rasanten Entwicklung großer Sprachmodelle und multimodaler Modelle in den letzten zwei Jahren ist KI für viele Menschen im Alltag greifbar geworden. Viele dieser Anwendungen basieren auf KI-Modellen, die auf Basis mehrschichtiger künstlicher neuronaler Netze trainiert wurden (Deep Learning). Wie und warum solche Anwendungen zu ihren Ergebnissen kommen, also warum eine Anweisung an einen Chatbot genau zu der Wortfolge führt, die er ausgibt, oder warum ein Bildgenerator auf eine Anfrage hin genau dieses und kein anderes Bild erstellt, bleibt in der Regel undurchsichtig. Diese Herausforderung betrifft allgemein KI-Systeme mit hoher Komplexität, ob sie nun darauf abzielen, Inhalte zu generieren, Vorhersagen zu treffen oder Klassifikationen auszugeben.[1] Aufgrund dieser Unklarheit, was zwischen der Modelleingabe und der Modellausgabe passiert, werden solche Modelle auch als Black-Box-Modelle bezeichnet (Erklärbox 1). Diese mangelnde Transparenz kann zu Vertrauensverlust seitens der Nutzenden in KI-Technologien führen und auch die Weiterentwicklung und Verbesserung von KI-Systemen hemmen.

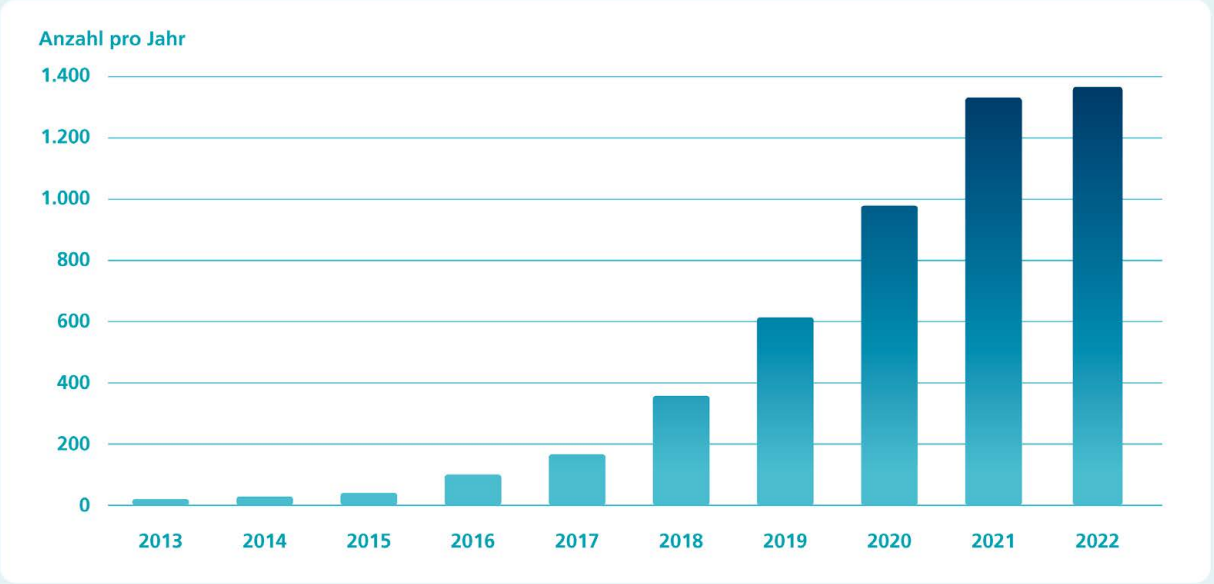
Forschungsfeld Erklärbare KI

Diesen Herausforderungen stellt sich insbesondere Erklärbare KI (englisch explainable artificial intelligence; kurz: XAI) als Forschungsrichtung, die sich seit 2016 bereits in diesem Bereich ausdifferenziert und zu einem produktiven Forschungsgebiet entwickelt (Abbildung 1). Dies hat zu einer Vielfalt an Methoden und Werkzeugen geführt, um die Herausforderungen von Erklärbarkeit zu adressieren.[2] Seit 2021 hat das Forschungsfeld einige weitreichende Fortschritte gemacht, wodurch XAI-Methoden heute die Erkennung von Verzerrungen in KI-Modellen erleichtern und auch konkrete Modellverbesserungen ermöglichen. Darüber hinaus kann XAI auch eingesetzt werden, um wissenschaftliche Entdeckungen in den Domänenwissenschaften voranzutreiben. Erklärbare KI ist zudem ein bedeutender Strang des Forschungsgebiets der menschenzentrierten KI (Human-Centered AI, HCAI), aus dessen Perspektive Erklärbarkeit in Benutzendenschnittstellen eingebettet ist und stets in konkreten Kontexten und Schnittstellen stattfindet (siehe Abschnitt: Form der Erklärung). Es ist ein Gebiet, das sich der Aufgabe annimmt, KI-Systeme zu entwickeln, die menschliche Fähigkeiten erweitern oder ergänzen, anstatt sie zu ersetzen. Menschliche Kontrolle soll bewahrt werden, sodass sichergestellt wird, dass KI-Systeme menschliche Bedürfnisse auf verantwortungsvolle Weise bedienen (siehe Abschnitt: Vertrauen gewinnen).

1 Als KI-System ist im Rahmen des Whitepapers ein Software- oder Hardwaresystem zu verstehen, das ein oder mehrere KI-Modelle nutzt, um zum Beispiel Vorhersagen und Klassifikationen zu erstellen, und das auch über Komponenten für Erklärbarkeit und Interaktivität etc. verfügen kann. Die Erklärbarkeit bezieht sich dabei auf die Modelle des KI-Systems.

2 Für einen Überblick über Methoden wie SHAP, LRP und Lime und ihre jeweiligen Vor- und Nachteile siehe Kraus et al. 2021 und Abschnitt 4 und einen Überblick über Theorien, Algorithmen und Anwendungen bei Samek et al. 2019.

Abbildung 1: **XAI-Publikationen (Anzahl pro Jahr)**



Quelle: Jacovi (2023).

Im Forschungsfeld XAI haben sich zwei Communities gebildet: Die erste konzentriert sich auf Werkzeuge zur Daten- und Modellexploration und Modellverbesserung, die zweite beschäftigt sich mehr mit XAI im Kontext ethischer Kriterien für verantwortungsvolle KI. Die Perspektive der zweiten Community ist weiter verbreitet als die erste (Biecek & Samek 2024, S. 5). Insbesondere wird mit dieser Perspektive in der Literatur häufig eine nicht-technische Zielgruppe für Erklärungen adressiert. Die Arbeit beider Communities ist von großer praktischer Relevanz, wie beispielsweise eine Gegenüberstellung der Perspektive von Forschung und Praxis entlang des Zyklus des Maschinellen Lernens zeigt (Decker et al. 2023). Aus der Gegenüberstellung geht unter anderem hervor, dass offenbar in der Praxis die Vielfalt der XAI-Werkzeugkiste aus der Forschung noch nicht ausgeschöpft wird und zugleich Praktiker nach Werkzeugen nachfragen, die so noch nicht vorhanden sind.

Gesellschaftliche Relevanz

Transparenz im Sinne von Nachvollziehbarkeit ist besonders dann relevant, wenn maschinelles Lernen (ML) in gesellschaftlichen Bereichen eingesetzt werden soll, in denen Vertrauen in Personen, Institutionen oder Technologien besonders wichtig ist (Tabelle 1), wie beispielsweise im Gesundheitswesen (Budde et al. 2023, Müller-Quade et al. 2020), in der Justiz (Rostalski et al. 2024) oder auch in der Mobilität, um Sicherheitsnachweise zu untermauern (Bahlmann 2024). Auch in der Industrie kann es unter dem Blickwinkel der Qualitätssicherung sinnvoll sein, auf XAI zu setzen, sowie ethisch und rechtlich geboten sein, ML-Modelle erklärbar zu gestalten – insbesondere, wenn es um die Sicherheit von Menschen geht (Gramelt et al. 2024).

Erklärbare KI kann dazu beitragen, Werte wie Transparenz und Verlässlichkeit zu fördern, die auch in der europäischen Regulierung oder in internationalen Standards angestrebt werden. So besteht nach der Datenschutzgrundverordnung (DSGVO) ein Recht auf Erläuterung des Verfahrens und der Grundsätze, die bei einer automatisierten Verarbeitung der personenbezogenen Daten einer betroffenen Person zur Anwendung kamen, um auf der Grundlage dieser Daten zu einem bestimmten Ergebnis zu gelangen. Die relevanten

Informationen müssen dabei in präziser, transparenter, verständlicher und leicht zugänglicher Form übermittelt werden (DSGVO, Art. 15 h sowie EuGH, Urteil vom 27. Februar 2025 – C203/22 [ECLI:EU:C:2025:117], Rn. 58). In der Norm der International Organisation for Standardisation (ISO) zum KI-Managementsystem von 2023 wird Erklärbarkeit unter dem Begriff der verantwortungsvollen KI bereits als Zielkategorie genannt (ISO/IEC 42001, B.9.3). In der europäischen KI-Verordnung, dem AI Act, stehen hingegen die Kriterien Transparenz (inkl. Dokumentation) und menschliche Aufsicht im Vordergrund (Panigutti et al. 2023). „Erklärbare KI" als Begriff findet sich dort allerdings nicht. Dort heißt es in den Transparenzanforderungen für Hochrisikoanwendungen: „Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass ihr Betrieb hinreichend transparent ist, damit die Betreiber die Ausgaben eines Systems angemessen interpretieren und verwenden können." (Art. 13 Abs. 1 KI-VO). Aktuell wird darüber hinaus der KI-Verhaltenskodex für Allzweck-KI (engl. General-Purpose AI Code of Practice) ausgearbeitet (AI Office 2025). In einem Entwurf des Kodex wird der Mangel an Erklärbarkeit und Transparenz als Faktor gelistet, der das systemische Risiko, das von solchen KI-Modellen ausgeht, potenziell beeinflussen kann (siehe Appendix 1.4.3, ebenda). XAI-Methoden können zu einer solchen Transparenz einen Beitrag leisten.

Tabelle 1: **XAI aus Sicht verschiedener Domänen**

Domäne	Perspektive auf XAI (Auszug)
Medizin Müller-Quade et al. 2020, S. 26 & 40 	„Ergebnisse, die mit Hilfe eines KI-Systems ermittelt wurden, müssen nachvollziehbar und erklärbar sein. Dazu zählt vor allem, dass eine Ärztin oder ein Arzt die vorgeschlagenen Analyseergebnisse kritisch hinterfragen können muss. Wenn Medizinerinnen und Mediziner Nachfragen stellen, muss das KI-System in der Lage sein, qualifizierte und interpretierbare zutreffende Gründe für die ausgegebene Analyse darzulegen. Die behandelnden Ärztinnen und Ärzte dürfen das vorgeschlagene Ergebnis nicht unreflektiert als Ergebnis übernehmen. Vielmehr sollen sie dieses mit in die Behandlungsentscheidung einbeziehen – als einen Aspekt, der ihr Wissen und ihre Erfahrungen ergänzt. […] Die Schwierigkeit: So wünschenswert eine maximale Nachvollziehbarkeit einerseits erscheint, könnte sie andererseits zu einer Informationsüberflutung führen. Weder den Patientinnen und Patienten noch dem medizinischen Personal wäre dadurch geholfen."
Justiz Rostalski et al. 2024 	„Wenn komplexe KI-Systeme zur Unterstützung richterlicher Entscheidungen zum Einsatz kommen, deren Entscheidungsfindung weder für die Richterinnen und Richter noch die Betroffenen nachvollziehbar ist („Black-Box"-Problematik), stellt dies ein gravierendes Problem für das Vertrauen in das Justizsystem dar. […] Je mehr KI Einfluss auf tatsächliche Urteile oder Empfehlungen eines Gerichts hat, desto bedeutender wird die Anforderung an Transparenz und Erklärbarkeit der KI-Beiträge. Ebenso müssen den Betroffenen die für die Entscheidungsempfehlung maßgeblichen Gründe offen zugänglich sein. Dies trifft sowohl auf die Faktoren, die ein KI-System für eine Empfehlung oder Entscheidung verwendet, und deren Gewichtung zu als auch auf die Gründe von Anwältinnen und Anwälten oder Richterinnen und Richtern, der Empfehlung eines KI-Systems (nicht) zu folgen."
Mobilität Bahlmann et al. 2024 	„Die Ziele können dabei je nach Phase des Produktzyklus unterschiedlich sein: Während der Entwicklung und Zulassung, um beispielsweise Fehler im System leichter entdecken zu können, oder auch um einen beispielsweise im Straßenverkehr geforderten Sicherheitsnachweis angemessen zu untermauern. In dieser Phase ist XAI insbesondere für Unternehmen und deren Entwicklerinnen und Entwickler sowie Institutionen und Behörden für die Prüfung des Verhaltens relevant. Während der Anwendung, um sinnvolles Feedback zu geben und das momentane Verhalten interaktiv und nachvollziehbar zu gestalten. Hier sollten die Erklärungen hauptsächlich der Nachvollziehbarkeit für Nutzerinnen und Nutzer dienen, beispielsweise im Falle eines Eingreifens von Assistenzsystemen. Nach der Anwendung, um eine Fehlfunktion analysieren zu können, beispielsweise zum Zweck einer iterativen Verbesserungsschleife oder zur Klärung einer Haftungsfrage im Falle eines Verkehrsunfalls. Hier ist XAI primär für Institutionen und Behörden wichtig, um Risikosituationen und Unfälle nachvollziehen zu können."
Industrie Gramelt et al. 2024 	„Ein typischer Anwendungsfall für die Erkennung von Anomalien in der Industrie ist die Qualitätssicherung in der Produktionslinie. Um den Prozess der Erkennung von Defekten zu beschleunigen, ist es hilfreich, Endnutzenden eine Begründung für die Klassifizierungsentscheidung zu geben, was im Falle von Bildern durch die Hervorhebung der für die Anomalieerkennung verantwortlichen Pixel erreicht werden kann. Dies erhöht auch das Vertrauen in das Modell selbst. Darüber hinaus ist es nicht nur ein ethisches, sondern auch ein rechtliches Gebot, KI-Systeme erklärbar zu machen, insbesondere in Bereichen, in denen die Sicherheit von Menschen auf dem Spiel steht." [Übers. d. Red.]

Quelle: Eigene Zusammenstellung.

Wirtschaftliche Relevanz

Die Unternehmensberatung McKinsey (2022) verweist in einer Studie darauf, dass Unternehmen, die das digitale Vertrauen[3] der Verbrauchenden stärken, indem sie zum Beispiel KI erklärbar machen, eine höhere Chance haben, ihren Gewinn signifikant zu steigern (ebenda). XAI kann laut der Studie die Produktivität erhöhen, indem beispielsweise Fehler in KI-Anwendungen schneller erkannt oder die Anwendungen verbessert werden. Sie kann dazu beitragen, das Vertrauen von Verbrauchenden, Behörden und der Öffentlichkeit in KI-Anwendungen zu stärken und Risiken für Unternehmen, beispielsweise durch Regulierung, zu verringern. Darüber hinaus kann das Wissen über den Grund hinter KI-gestützten Vorhersagen wichtige Erkenntnisse für die Geschäftsentwicklung liefern (etwa bei Vorhersagen zur Kundenabwanderung) oder dabei helfen zu überprüfen, ob die Vorhersagen mit den Unternehmenszielen übereinstimmen. XAI ist aber auch wichtig für das Change-Management bei der Einführung von KI-Systemen in Unternehmen (Stowasser et al. 2020): Es gilt als wichtiges Kriterium für die menschengerechte Gestaltung der Arbeitswelt im Hinblick auf die Mensch-Maschine-Interaktion (Huchler et al. 2020).

Ziele des Whitepapers

Vor dem Hintergrund dieser gesellschaftlichen und wirtschaftlichen Relevanz möchte dieses Whitepaper im Wesentlichen drei Beiträge leisten:

1. Entlang von „W-Fragen“ soll herausgearbeitet werden, dass XAI nicht als Selbstzweck oder nach dem Grundsatz eines „one-size-fits-all“ eingesetzt werden sollte. Es wird ein reflektierter Einsatz von XAI empfohlen, der die Frage adressiert, wem, was, wozu und wie erklärt werden soll. Das heißt, wie Erklärungsadressat bzw. Erklärungszielgruppe (wem), Erklärungsgegenstand (was), Erklärungsziel (wozu) und Erklärungsumsetzung (wie) im jeweiligen Fall ideal zu verknüpfen sind.

Zentrale Fragestellungen zu XAI

3 „Digitales Vertrauen“ wird definiert als Zuversicht, dass ein Unternehmen Verbraucherdaten schützt, wirksame Cybersicherheit gewährleistet, durch vertrauenswürdige KI-gestützte Produkte und Dienstleistungen anbietet und Transparenz in Bezug auf KI und Datennutzung schafft (vgl. Boehm et al. 2022).

2. Dabei werden wesentliche Aspekte von XAI eingeordnet (inklusiver neuer Trends sowie Entwicklungen und Grenzen, wie zum Beispiel XAI für große Sprach- und Allzweckmodelle sowie interaktive XAI).

3. Dadurch soll eine kalibrierte Erwartungshaltung erreicht werden. Die Erwartung einer vollständigen und umfassenden Erklärung zu komplexen KI-Systemen ist nicht realistisch und auch nicht in jedem Fall notwendig. Selbst der Mensch scheitert oft daran, umfassende, gute und sinnvolle Erklärungen für Entscheidungen zu liefern. Daher sollte man dies auch nicht von KI-Systemen erwarten, sondern XAI mit einer realistischen Erwartungshaltung entwickeln und anwenden. XAI kann letztlich nur Aspekte des Modellverhaltens erklären. Diese Kalibrierung ist auch deshalb sinnvoll, weil die Erwartungen an technische Maßnahmen zur Förderung von Transparenz und Erklärbarkeit zwischen Forschenden und Entwickelnden einerseits und der Gesellschaft andererseits sehr unterschiedlich sein können (Gyevnar, Ferguson & Schafer 2023). Auch unter Forschenden gibt es mitunter unterschiedliche Vorstellungen darüber, welche Rolle die XAI-Methoden einnehmen sollen sowie welche Ziele damit verfolgt werden sollen beziehungsweise können (Biecek & Samek 2024).

Im Rahmen dieses Whitepapers wird XAI nicht im naturwissenschaftlichen Verständnis der Trias „Beschreiben, Erklären, Vorhersagen" behandelt, da in diesem Kontext keine Vorhersagen möglich sind, sondern nur Ex-post-Erklärungen. Stattdessen nimmt das Papier eine pragmatisch-utilitaristische Perspektive ein und fragt: Wie kann uns XAI helfen, Aspekte rund um KI-Systeme zu erhellen und KI-Systeme zu verbessern?

Erklärbox 1

White-Box-Modelle und Black-Box-Modelle

„**White-Box-Modelle** […] erlauben das grundsätzliche Nachvollziehen ihrer algorithmischen Zusammenhänge; sie sind somit selbsterklärend in Bezug auf ihre Wirkmechanismen und die von ihnen getroffenen Entscheidungen." (Kraus et al. 2021, S. 3)

Beispiele: Entscheidungsbäume, Regression

„Bei **Black-Box-Modellen** wie neuronalen Netzen ist es aufgrund ihrer Verflechtung und Vielschichtigkeit in der Regel nicht mehr möglich, die innere Funktionsweise des Modells nachzuvollziehen." (Kraus et al. 2021, S. 3)

Beispiele: Transformer, Convolutional Neural Networks

Allerdings können auch White-Box-Modelle Erklärungen benötigen, zum Beispiel, wenn es sich um komplexe Systeme handelt. Weiterhin sind auch Regressionen und Entscheidungsbäume oftmals schwer zu interpretieren, wenn eine größere Zahl an Variablen einbezogen wird. Daraus folgt, dass XAI-Methoden als Werkzeuge dienen können, die sowohl wertvolle Einsichten in „transparente" als auch in „intransparente" Modelle geben können (Atzmueller et al. 2024, Biecek & Samek 2024).

2 Klärung von Begriffen und ihres Verhältnisses

Im Themenkomplex XAI werden viele unterschiedliche Begriffe verwendet, die in der öffentlichen Diskussion nicht immer klar definiert sind und mitunter unsystematisch betrachtet werden. Es ist daher notwendig, sich zunächst Klarheit über Begriffe wie Transparenz, Nachvollziehbarkeit, Interpretierbarkeit und Erklärbarkeit zu verschaffen und vor allem die Zusammenhänge zwischen diesen Begriffen zu klären. Dabei ist zu beachten, dass in der Forschung die Diskussion zu Definition, Abgrenzung und Systematisierung der Begriffe kontinuierlich andauert und daher Begriffe teilweise unterschiedlich definiert und voneinander abgegrenzt werden.

Unterschiede im Verständnis von Transparenz

Transparenz[4] kann in einem eher engen technischen Sinne verstanden werden. Transparenz wird in dieser Perspektive auf technische Lösungen bezogen, um KI-Systeme für bestimmte Zielgruppen nachvollziehbar zu machen oder auch als Mittel zur Verbesserung von KI-Modellen und -Anwendungen. Die Diskussion zu Transparenz in der Forschung geht allerdings weit darüber hinaus. So werden unter anderem Vorschläge für Berichtssystematiken über maschinelles Lernen (Reporting) gemacht und zur Einstufung von Lernmethoden (AI Labeling) oder auch zu Messsystematiken, um die Energieeffizienz eines KI-Systems visuell kommunizieren zu können (Morik et al. 2022, Fischer et al. 2022, 2024, Fischer, Liebig & Morik 2024). Open Source ist ebenfalls als Thema der Transparenz zu nennen.

Verhältnis von Nachvollziehbarkeit, Transparenz, Interpretierbarkeit und Erklärbarkeit

Transparenz im Sinne der Nachvollziehbarkeit eines KI-Systems ist eine Systemeigenschaft (auch als comprehensible AI bezeichnet), die das „Innenleben" des Systems – die Funktionsweise, die internen Schlussfolgerungen und Verarbeitungsprozesse – für relevante Nutzendengruppen nachvollziehbar macht und ihnen die Systemein- und -ausgaben (z. B. Systementscheidungen) adäquat erklären kann. Sowohl Erklärbarkeit als auch Interpretierbarkeit können zu einer nachvollziehbaren KI beitragen (Schramm, Wehner & Schmid 2023, Schwalbe & Finzel 2024, Poretschkin et al. 2021).

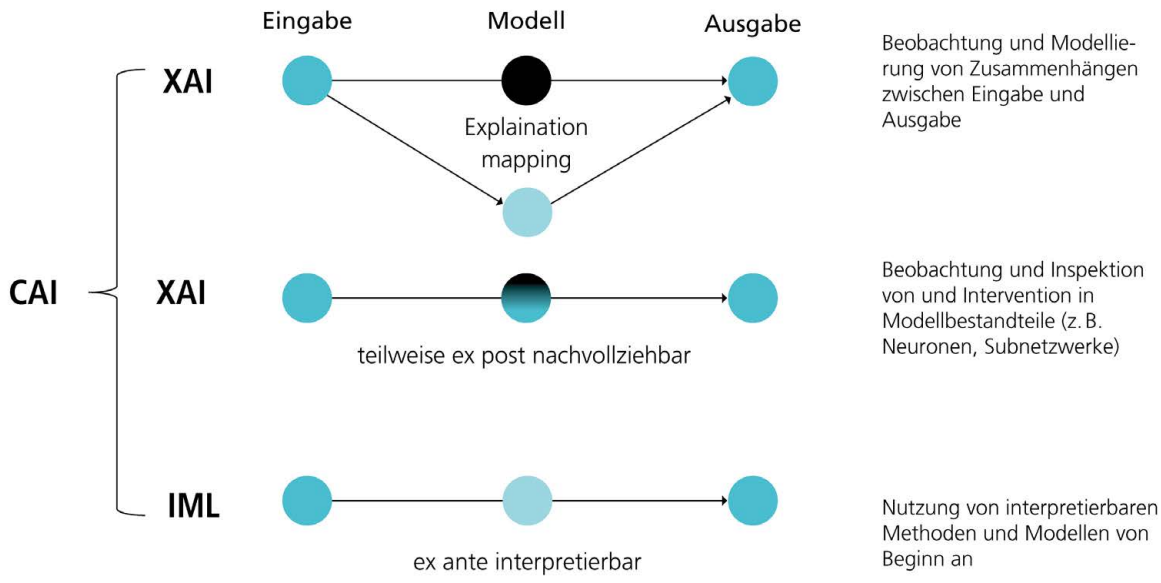
Erklärbarkeit zielt darauf ab, das Verhalten eines KI-Modells nachvollziehbar zu machen, also zu erklären, wie die Modellergebnisse (z. B. Vorhersagen oder Klassifikationen) zustande gekommen sind, und den Nutzendengruppen aktiv Begründungen dafür zur Verfügung zu stellen (Kraus et al. 2021, Poretschkin et al. 2021, Schramm, Wehner & Schmid 2023). Diese XAI-Methode liegt hierbei oft außerhalb des Modells (Abbildung 2, oberer Pfad), modellinterne Prozesse von Black-Box-Modellen werden dann nicht erklärt. Stattdessen bilden solche Methoden die Eingabe eines KI-Modells auf dessen Ausgabe ab. Erklärungen, die auf diese Weise erstellt werden, sind entsprechend nicht immer getreu dem Verhalten des Modells (siehe Kapitel 5: Evaluierung) (ebenda). Allerdings sind XAI-Methoden nach dem oberen Pfad der Abbildung 2 bei (kommerziellen) KI-Systemen, deren Modell selbst nicht zugänglich ist und die deshalb nur als Black-Box-Modelle verwendbar sind, die einzige Möglichkeit, Erklärbarkeit zu erreichen.

4 XAI wird aber auch in manchen Studien attestiert, Transparenz eher als Systemeigenschaft als zu erreichenden Selbstzweck zu betrachten (Gyevnar, Ferguson & Schafer 2023). Das Verständnis des Gesetzgebers unterscheidet sich deutlich von einer solchen Sichtweise. Es handelt sich um ein breites und umfassendes Verständnis von Transparenz als Mittel, um zu umfassenderen Werten wie Rechenschaftspflicht, Menschenrechten und nachhaltiger Innovation beizutragen (ebenda). Dies kann zu Diskrepanzen zwischen gesellschaftlicher Wünschbarkeit und technischer Machbarkeit führen.

Neuere XAI-Methoden zielen allerdings auch auf die Erklärung von Zuständen und Prozessen von KI-Modellen ab, die auf neuronalen Netzwerken beruhen (mechanistische Interpretierbarkeit, siehe Abschnitt 4), sodass das Innenleben der Netzwerke beobachtet und inspiziert werden kann oder auch sogar darin interveniert werden kann (Abbildung 2, mittlerer Pfad).

Die Interpretierbarkeit bezieht sich auf das Modell selbst. Ist ein Modell interpretierbar, kann es als White-Box-Modell bezeichnet werden (Abbildung 2, unterer Pfad). Das Modell folgt standardmäßig einem für Menschen nachvollziehbaren Entscheidungsprozess (Schramm, Wehner & Schmid 2023). Die algorithmischen Mechanismen zur Generierung des Modells sind nachvollziehbar, der verwendete Lernprozess als Ganzes ist somit transparent (Kraus et al. 2021, Poretschkin et al. 2021). Ein interpretierbares Modell sollte in der Lage sein, selbst eine Erklärung für eine Klassifikation oder Vorhersage zu liefern, und diese Erklärung sollte getreu dem Entscheidungsfindungsprozess und dem Verhalten des Modells sein (Schramm, Wehner & Schmid 2023).

Abbildung 2: **Konzept der Comprehensible AI (nachvollziehbaren KI)**



CAI = Comprehensible Artificial Intelligence, d. h. Nachvollziehbare KI; **XAI** = Explainable AI;
IML = Interpretable Machine Learning, d. h. interpretierbares maschinelles Lernen.

Quelle: Erweiterte Darstellung nach Schramm, Wehner & Schmid (2023).

3 Gegenstände der Erklärung

Vereinfacht lassen sich drei Ebenen im Prozess des maschinellen Lernens unterscheiden: Eingabe-, Modell- und Ausgabe-Ebene (Abbildung 2). Hier wiederum lassen sich zwei Phasen unterscheiden, die Trainingsphase, in der auf Basis der Eingabe – eines Datensatzes – ein Modell gelernt wird, und die Inferenzphase, in der auf Basis des gelernten Modells eine Ausgabe berechnet wird. Jede der Ebenen und Phasen kann Gegenstand einer Erklärung durch XAI-Methoden sein.

Eingabeebene – Daten

Die Grundlage jedes aktuellen KI-Modells sind große Datenmengen. Sie dienen als Trainingsmaterial für die Lernalgorithmen, die das KI-Modell erzeugen. Dabei kann es sich beispielsweise um Text-, Bild- oder Videodaten handeln. So nutzen Modelle wie GPT-4 große Mengen von Textdaten aus dem Internet (u. a. der Common-Crawl-Datensatz, Common Crawl Foundation 2025). Als Trainingsdaten für KI-Modelle werden auch tabellarische Daten, Zeitreihen, Sensordaten und vieles mehr verwendet. Darüber hinaus können unterschiedliche Datentypen zur Erstellung multimodaler Modelle verwendet werden. Da die Qualität, Zusammensetzung und Aufbereitung des Datensatzes sowie dessen Kuratierungen nach bestimmten Kriterien (z. B. Fairness) einen Einfluss auf das KI-Modell und damit auf das Modellverhalten haben, sind Werkzeuge wünschenswert, die generell das Verständnis des Datensatzes erhöhen und mit denen Verzerrungen, Fehler oder Ausreißer in den Datensätzen entdeckt werden können. XAI-Methoden können helfen, Rückschlüsse auf die Qualität der Datenbasis zu ziehen, indem untersucht wird, wie stark welche Merkmale (z. B. Wörter, Pixel) über den gesamten Datensatz zur Modellausgabe beitragen (globale Erklärung), wobei der Fokus auf den Merkmalen liegt, und was diese über den Datensatz aussagen. Auch die Entdeckung von Bestandteilen oder Einträgen in einem Trainingsdatensatz, die sehr viel Einfluss auf eine Vorhersage haben, kann ein Ziel sein, um die Daten und ihre Qualität besser zu verstehen (Training Data Attribution).

Modellebene – Modellkomponenten

Beim Training eines KI-Modells werden mit Hilfe von Lernalgorithmen (z. B. neuronale Netze mit mehreren Netzwerkebenen, sogenanntes Deep Learning) Muster, Beziehungen und Abhängigkeiten innerhalb des Datensatzes erlernt. Auf diese Weise entsteht ein Modell, das Eingabedaten auf eine Ausgabe abbildet. Auf Modellebene ist daher von Interesse, was tatsächlich gelernt wurde und welche Eigenschaften eines Modells auf welche Weise zu einer Klassifikation oder Vorhersage beitragen. Beim Deep Learning kann die Rolle der verschiedenen Netzwerkschichten von Interesse sein oder das Aktivitätsmuster der „Neuronen", um auf ihre Rolle bei der Klassifikation oder Vorhersage zu schließen. Mögliche Fragestellungen sind in diesem Kontext (vgl. hierzu Reuker et al. 2023): Welche Konzepte werden durch einzelne Neuronen eines künstlichen neuronalen Netzwerks repräsentiert? Welche spezifischen Funktionen haben bestimmte Bestandteile – zum Beispiel Sub- und Teilnetzwerke – solcher Netzwerke und ihre Interaktion? Sobald das Modell erstellt ist und sich im Betrieb befindet, kann XAI auch eine Rolle bei der Modellüberwachung spielen. Hier stellt sich zum Beispiel die Frage, worauf fehlerhafte Vorhersagen zurückzuführen sind, die gegebenenfalls während des laufenden Betriebs auftreten können.

Ausgabeebene – Vorhersagen und Klassifikationen

Bei der Ausgabeebene steht die Inferenz im Mittelpunkt, also wie aus dem trainierten KI-Modell und neuen Daten, wie zum Beispiel ein Prompt für einen Chatbot oder ein Bild, dessen Inhalt bestimmt werden soll, eine Vorhersage oder Klassifikation berechnet wird. Es geht darum, zu erklären, wie eine konkrete Modellausgabe zustande gekommen ist. Dies kann die Klassifikation eines Objekts auf einem Bild, die Antwort eines Chatbots auf eine Anfrage oder ein Vorschlag für eine medizinische Diagnose sein. Dabei können zum einen einzelne konkrete Ausgaben des KI-Modells im Mittelpunkt stehen (lokale Erklärung). Zum anderen ist auf dieser Ebene auch die Erklärung des Modellverhaltens im Allgemeinen beziehungsweise der gelernten Zusammenhänge im Allgemeinen von Interesse (globale Erklärung). Auch hier kann von Interesse sein, welches Merkmal wie zu einer Vorhersage oder Klassifikation beigetragen hat, der Fokus liegt dann aber auf der Ausgabe.

Neuere Methoden wie SemanticLens (Dreyer et al. 2025) betrachten die drei Ebenen nicht als separate Elemente einer Erklärung, sondern verfolgen einen ganzheitlichen Ansatz. Sie ermöglichen es, die Entscheidungen eines KI-Systems sowohl in Bezug auf die Komponenten der KI (z. B. einzelne Neuronen) als auch auf die Eingabedaten zu erklären und diese beiden Ebenen miteinander in Beziehung zu setzen. Auf diese Weise lässt sich die Rolle und Funktionsweise aller Komponenten von KI systematisch auswerten und validieren.

4 Ziele der Erklärung

Es ist wichtig, sich über realistische Ziele im Klaren zu sein und damit zu klären, wann erklärbare KI sinnvoll ist und wann nicht. Erklärbare KI sollte kein Selbstzweck sein, denn in vielen Fällen ist Erklärbarkeit nicht unbedingt notwendig. Menschen haben oft Vertrauen in andere Menschen, Tiere und technische Systeme, wenn diese verlässlich und erwartungskonsistent sind (vgl. für Mobilität auch Bahlmann et al. 2024, S. 17). Ein Beispiel hierfür sind Blindenführhunde: Wie der Hund einen Zebrastreifen oder einen freien Übergang erkennt, werden wir nicht im Detail wissen. Aber wir können darauf vertrauen, dass die Hundetrainerinnen und -trainer den Hund gut auf seine Aufgabe vorbereitet haben und der Hund daher seine Aufgabe zuverlässig erfüllt. Vor diesem Hintergrund ist es wichtig, sich darüber im Klaren zu sein, „wozu" XAI eingesetzt werden soll. Um sich an dieser Frage zu orientieren, lohnt es sich, zwei allgemeinere Ziele von XAI, die sich teilweise überschneiden, näher zu betrachten:

1. Nutzung von XAI-Methoden, um den Dimensionen vertrauenswürdiger KI gerecht zu werden. Damit wird das allgemeine Ziel verfolgt, Vertrauen in KI zu stärken.

2. Nutzung von XAI-Methoden als Ingenieurswerkzeug, um die Qualität von KI-Modellen zu verbessern.

Naturgemäß werden unterschiedliche Zielgruppen von XAI jeweils eigene Gründe haben, wozu sie XAI einsetzen wollen (siehe Abschnitt 5): beispielsweise, um etwas über die Daten und eine konkrete Problemstellung zu lernen (siehe Identifikation von Antibiotika, Persona 2) oder um eine monetär oder existentiell wichtige Entscheidung nachvollziehbar zu machen (z. B. bei Investitionsentscheidungen oder medizinischen Diagnosen). Dennoch werden die Zielsetzungen aus den Bereichen „XAI für vertrauenswürdige KI" und „XAI als Ingenieurswerkzeug" in weiten Teilen ein Mittel zum Zweck der Realisierung zielgruppenspezifischer Ziele darstellen, sei es die Kundenzufriedenheit, Wissensgenerierung oder die Entscheidungsunterstützung.

Vertrauen gewinnen: Mit XAI zu vertrauenswürdiger KI

Nachvollziehbare KI kann auf verschiedene Weise zur Vertrauenswürdigkeit von KI-Systemen beitragen. Vertrauenswürdigkeit wird als mehrdimensionales Konzept verstanden, das in der Regel mehrere ethische und technische Prinzipien umfasst. Weltweit wurden Hunderte von Katalogen und Richtlinien für ethische KI entwickelt. Besonders häufig werden dabei Prinzipien wie Transparenz, Sicherheit, Vertrauenswürdigkeit, Autonomie und Privatsphäre beziehungsweise Datenschutz genannt (vgl. Kluge Corrêa et al. 2023). Solche Kategorien finden sich auch in Normen der International Organisation for Standardisation (ISO) unter dem Begriff „responsible AI" (ISO/IEC 42001, B.9.3) oder im Prüfkatalog des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme (IAIS), der sechs Prüfdimensionen für vertrauenswürdige KI definiert: Autonomie und Kontrolle, Fairness, Transparenz, Verlässlichkeit, Sicherheit, Datenschutz (Poretschkin et al. 2021). Zu diesen sechs Dimensionen kann XAI auf unterschiedliche Weise beitragen und damit auch zum Ziel einer verantwortungsvollen Gestaltung von KI-Systemen.

Autonomie und Kontrolle

Autonomie bezieht sich hier sowohl auf das KI-System als auch auf den Menschen. Auf Systemebene werden häufig verschiedene Autonomiegrade unterschieden, bei denen die Kontrollmöglichkeiten des Menschen

variieren. So kommt dem Menschen bei höheren Autonomiegraden in der Regel eine überwachende oder letztentscheidende Rolle zu. In diesem Fall schlägt das KI-System dem Menschen gegebenenfalls Optionen vor oder bittet um die Freigabe einer Aktion (Beyerer et al. 2021). Die Erklärbarkeit dieser Systemausgabe ist daher entscheidend, um als Mensch eine informierte und kritisch reflektierte Entscheidung treffen zu können, aber auch, um eine effektive Aufsicht über ein KI-System mit einem höheren Grad an Autonomie ausüben zu können. Dies betrifft zugleich die Autonomie des Menschen im Sinne von Handlungsfähigkeit. Diese Handlungsfähigkeit wird in vielen Fällen erst durch die Nachvollziehbarkeit der Modellausgabe hergestellt.

Fairness

Fairness bezieht sich unter anderem auf die Abwesenheit von ungerechtfertigter Diskriminierung von Nutzenden. XAI kann dabei helfen, diskriminierende Verzerrungen in KI-Modellen aufzudecken und nachzuweisen. Damit schafft sie auch die Grundlage für die Korrektur solcher Verzerrungen. Hier besteht ein Zusammenhang zur Autonomie. Wenn Nutzende die Ausgabe eines KI-Systems anhand einer Erklärung auf Bias hin überprüfen können, erhalten sie Informationen, die ein Handeln erst möglich machen. Dadurch sind sie in der Lage, auch eine Veränderung herbeizuführen. XAI kann im gesamten Lebenszyklus von KI eine Rolle spielen, um dem Ziel der Fairness gerecht zu werden (vgl. Deck et al. 2024). Fairness ist allerdings kein absolutes Konzept, sondern abhängig beispielsweise von bestimmten kulturellen oder auch Unternehmenskontexten. Entsprechend können hier Ansätze des erklärenden interaktiven maschinellen Lernens hilfreich sein (siehe Abschnitt: Fokus: interaktive erklärbare Künstliche Intelligenz), aktuelle Biases in einem Modell sichtbar und korrigierbar machen (Heidrich et al. 2023).

Transparenz

Transparenz ist ein zentrales Anliegen von XAI. Durch die Offenlegung des „Innenlebens" von Black-Box-Modellen und die Nachvollziehbarkeit von Modelleingaben und -ausgaben werden die Modelle verständlicher und die Ergebnisse nachvollziehbarer.

Verlässlichkeit

Die Verlässlichkeit bezieht sich unter anderem auf die Fehlerhaftigkeit von Vorhersagen, die Robustheit des Modells sowie den Modell- beziehungsweise Konzeptdrift (Poretschkin et al. 2021, S. 22). Bei der Überprüfung der Zuverlässigkeit von Systemausgaben kann XAI hilfreich sein, um zum Beispiel falsche Korrelationen, die gelernt wurden, aufzudecken („Kluger Hans"-Effekt) (Lapuschkin et al. 2019). Darüber hinaus kann XAI eine Lösung für das Model Monitoring darstellen, um Modeldrifts zu identifizieren (Koebler et al. 2023). Bei XAI kann wiederum die Zuverlässigkeit und Robustheit der Erklärung selbst zum Thema werden (siehe Abschnitt 4, Bewertung).

Sicherheit

XAI kann zur funktionalen Sicherheit (Safety) von KI-Systemen beitragen, indem es Forschenden und Entwickelnden Werkzeuge an die Hand gibt, um Fehler und Schwachstellen in KI-Modellen zu erkennen und zu beheben (Biecek & Samek 2024). Sie kann dabei helfen herauszufinden, warum sich ein KI-Modell auf eine bestimmte Weise verhält und wie dieses Modellverhalten gegebenenfalls verhindert werden kann. Bei großen Sprachmodellen können XAI-Methoden beispielsweise Kombinationen von Neuronen identifizieren, die bestimmte Themen (auch Features oder Konzepte genannt) hervorrufen, wenn sie aktiviert werden. So können Teile des künstlichen neuronalen Netzes entdeckt werden, die ein Sprachmodell gefährlich machen

können (z. B. Konzepte über biologische Waffen, vgl. Levy 2024). Darüber hinaus kann eine „erklärbare KI [...], die ihre Entscheidungen nachvollziehbar macht, [...] eine unerwünschte Manipulation des KI-Algorithmus offenlegen" (Houdeau 2022).

Datenschutz

Diese Dimension betrifft den Schutz personenbezogener Daten, aber auch den Schutz von Geschäftsgeheimnissen. XAI kann zur Einhaltung verschiedener Datenschutzprinzipien beitragen. Da XAI unter anderem die wichtigsten Einflussfaktoren und Merkmale aufzeigen kann, die zu einer Modellausgabe geführt haben, könnte es dazu beitragen, die Datenerhebung und -verarbeitung zu reduzieren (Prinzip der Datensparsamkeit). Darüber hinaus könnte XAI dabei helfen, Proxy-Attribute zu identifizieren, die die Ausgabe von KI-Systemen mit bestimmten Datenkategorien (z. B. sexuelle Orientierung, Religion etc.) verknüpfen, und damit das Bewusstsein für solche Verknüpfungen schärfen (für weitere Informationen zum Verhältnis von XAI und Datenschutz siehe: European Data Protection Supervisor 2023).

Qualität verbessern: XAI als Ingenieurswerkzeug

XAI kann als Werkzeug von KI-Ingenieurinnen und -Ingenieuren eingesetzt werden, um die Qualität eines KI-Systems zu verschiedenen Zeitpunkten im KI-Lebenszyklus zu verbessern oder zu erhalten (Erklären, um zu revidieren, engl. explain to revise, explain to correct, debugging, vgl. Anders et al. 2022, Weber et al. 2023, Pahde et al. 2023, Schmid 2024).

Datenbasis verbessern

Da mit Hilfe von XAI-Methoden der Einfluss spezifischer Datenpunkte im Trainingsdatensatz auf Vorhersagen, Modellleistung und Modellparameter evaluiert werden kann, können umgekehrt auch Rückschlüsse auf die Datenbasis gezogen werden. Damit können verschiedene Herausforderungen der Datenqualität und -erfassung adressiert werden, wie beispielsweise die Bewertung der Datenqualität oder die Erkennung fehlerhafter Datenlabels. Darüber hinaus kann die Auswahl von Teilmengen eines Datensatzes oder die Beschaffung zusätzlicher Daten durch Informationen aus der Anwendung von XAI-Methoden unterstützt werden (Decker et al. 2023). Es können weiterhin unterrepräsentierte oder fehlerhafte Trainingsdaten identifiziert werden, um den Lernprozess zu verfeinern und damit eine höhere Genauigkeit der Vorhersagen und Klassifikationen zu erreichen.

Modellqualität erhalten und verbessern

Die Verbesserung der Modellqualität wirkt sich auch auf Dimensionen wie Zuverlässigkeit und Fairness des Konzepts der vertrauenswürdigen KI aus. Um die Qualität von Modellen zu erhalten und zu verbessern, kann XAI eingesetzt werden, um zu überprüfen, ob sich ein Modell so verhält, wie es sollte (Verifikation). Treten Auffälligkeiten auf, kann XAI helfen, Verzerrungen, Limitationen des Modells oder unerwünschtes Modellverhalten zu erkennen und zu beseitigen (Erklärbox 2). Wenn das Modell in Betrieb ist, kann XAI Teil der Modellüberwachung werden, um zu erkennen, wann die Qualität des Modells mit der Zeit abnimmt (Koebler et al. 2023). Auf diese Weise können Maßnahmen ergriffen werden, um das KI-Modell im Laufe der Zeit instand zu halten.

Erklärbox 2

Zyklus Aufdecken und Korrigieren

(1) Identifikation von Modellschwächen mit XAI-Methoden (z. B. Attributions-basiert oder Konzept-basiert, siehe Abschnitt 4), wie etwa falsch gelernte Zusammenhänge

(2) Lokalisierung von Fehlern bzw. Artefakten in Datensatzeinträgen oder in den gelernten Repräsentationen des Modells

(3) Korrektur des Modells, z. B. durch

- Anpassung des Datensatzes
- Anpassung der Verlustfunktion
- Löschen oder Verstärken relevanter Subnetzwerke eines KI-Modells
- Feedback des Menschen im Rahmen von interaktiven und erklärbaren KI-Systemen (z. B. reinforcement learning with human feedback, active learning etc.)

(4) Evaluierung, inwiefern das Modellverhalten noch durch lokalisierte Artefakte beeinflusst wird oder durch die Korrektur relevante Informationen für das Modell verloren gingen.

Quelle: Zusammenstellung basierend auf Samek (2025) erweitert um den Aspekt des menschlichen Feedbacks.

Parsimonie (Sparsamkeitsprinzip), Modell-Kompression und Ressourceneffizienz

Mit Hilfe von XAI können redundante oder unnötige Komponenten eines KI-Modells identifiziert werden und diejenigen, die den größten Einfluss auf die Modellausgabe haben (siehe z. B. Yeom et al. 2021, Vakilzadeh Hatefi et al. 2024). Beispielsweise können Verbindungen zwischen unbedeutenden Konzepten und bestimmten Neuronen eines neuronalen Netzes entdeckt werden, sodass diese Neuronen aus dem Netz entfernt werden können. Auf diese Weise können weniger komplexe Modelle erstellt werden, die gegebenenfalls auch mit weniger Ressourcen betrieben werden können. XAI kann somit zur Steigerung der Ressourceneffizienz von KI beitragen.

Domänenanpassung und Optimierung der Nutzenden-System-Interaktion

Im Allgemeinen können XAI-Methoden die Interaktion zwischen Nutzenden und dem KI-System optimieren, indem sie Einblicke in die Art und Weise geben, wie ein System funktioniert und wie bestimmte Ausgaben zustande kommen (Langer et al. 2021). So können personalisierte Assistenzsysteme für Endnutzende allgemein und Domänenfachleute entstehen (Göbel et al. 2022). Nutzende wie Domänenexpertinnen und -experten werden das Verhalten des Modells mit ihrer Intuition oder ihrem Erfahrungswissen abgleichen, sodass Diskrepanzen zwischen dem KI-Modell und den Domänenspezifika erkannt und minimiert werden können (vgl. Bhatt et al. 2020, S. 650). XAI-Methoden können hier unterstützen, indem sie beispielsweise Domänenexpertinnen und -experten den Abgleich mit dem Erfahrungswissen durch eine Erklärung erleichtern und so wiederum präziseres Feedback für die Anpassung des KI-Systems geben können. XAI kann einen Beitrag leisten, Hürden im Austausch zwischen Domänenexpertinnen und -experten, Modellerstellenden, Produktentwickelnden und Endverbrauchenden abzubauen (Decker et al. 2023) und auf diese Weise die Qualität von KI-Anwendungen zu verbessern.

5 Umsetzung der Erklärung

Grundsätzlich ist zu prüfen, ob für einen bestimmten Zweck ein einfaches White-Box-Modell sinnvoller ist als die Verwendung eines komplexen Black-Box-Modells. Wenn ein intransparentes Modell verwendet wird, das einer Erklärung bedarf, müssen sich Modellerstellende, KI-Forschende und -Entwickelnde unter anderem mit den Fragen auseinandersetzen, wie eine Erklärung gestaltet sein sollte (Form), wie XAI implementiert werden sollte (Methode) und wie die Korrektheit und der Nutzen einer Erklärung bewertet und verbessert werden kann (Evaluation). Im Anschluss an die Erläuterungen zu Form, Methode und Evaluation werden zwei Trends mit besonderer Aktualität näher dargestellt mit ihren jeweils spezifischen Herausforderungen – XAI für generative KI und interaktive erklärbare KI.

Form der Erklärung

Ziel sollte es sein, eine kontext- und zielgruppengerechte Erklärungsform anzubieten. Je nach den Bedürfnissen der Zielgruppe sollten Interpretationsspielräume erweitert oder reduziert werden, um so die Aufmerksamkeit je nach Zielgruppenprofil, Kontext und Interesse zu lenken. Erklärungen sollten alle relevanten Informationen enthalten, die für die Ausgabe eines KI-Systems von Bedeutung sind, und dabei so kompakt wie möglich und im jeweiligen Fall sinnvoll sein (Bekkemoen 2024). Grundsätzlich gilt: Während Modellerstellende (inklusive KI-Forschende) und gegebenenfalls in geringerem Maße auch Domänenexpertinnen und -experten mit abstrakteren und komplexeren Formen der Erklärung umgehen können, könnten Endverbrauchende von diesen leicht überfordert werden. Daher sind in diesem Fall niedrigschwellige und leichter zugängliche Formen der Erklärung erforderlich (vgl. kognitiver Aufwand der Erklärung, engl. „cognitive load", Abbildung 3, S. 25). Dies gilt auch für die Wahl der Modalitäten, da je nach Situation und Kontext eine andere Modalität vorteilhaft sein kann. So kann bei zeitkritischen Entscheidungen die visuelle Hervorhebung sinnvoll sein und bei Entscheidungen mit hoher Tragweite der Einsatz unterschiedlicher Modalitäten. Wenn es aber um komplexe relationale Abhängigkeiten geht, können verbale Erklärungen hilfreich sein (Schmid & Wrede 2022).

Es geht auch darum, das Vertrauen der Nutzenden zu kalibrieren: Weder zu viel noch zu wenig Vertrauen in das KI-System ist sinnvoll, insbesondere wenn es sich um entscheidungsunterstützende Systeme in Anwendungsbereichen handelt, in denen Entscheidungen mit hoher Tragweite getroffen werden (z. B. medizinische Diagnostik, Justiz etc.). Während KI-Forschende oder Modellerstellende mit facettenreichen Erklärungen Modelle explorieren und daraus lernen wollen, müssen Entscheiderinnen und Entscheider oder auch Endverbrauchende vor einer Informationsüberflutung bewahrt werden.

Auf der Ebene der Erklärungsform kann auch die Zusammenfassung oder vereinfachte Navigation zwischen verschiedenen Erklärungsvarianten ein Ziel sein oder die Vereinfachung einer komplexen und umfangreichen Erklärung, die eine hohe kognitive Belastung erzeugt.

Aus den verschiedenen Überlegungen zur Form der Erklärung ergibt sich, dass dem User Interface Design (UI-Design) und damit auch der interaktiven Visualisierung (Brath et al. 2023) eine wichtige Rolle zukommt, um den jeweiligen Zielgruppen einen adäquaten und zielgerichteten Zugang zu den Erklärungen zu ermöglichen. So könnten UI-Designs entwickelt werden, die den Nutzenden nicht nur eine einzige Erklärung anbieten, sondern die Erklärungen im Idealfall als interaktiven und iterativen Prozess abbilden. Das User Interface könnte in diesem Fall zum Beispiel die Navigation zwischen verschiedenen Erklärungen, die unterschiedliche Facetten des KI-Modells ausleuchten, ermöglichen. Es könnte weiterhin den Wechsel zwischen verschiede-

nen Erklärungstiefen erleichtern, die Personalisierung von Erklärformen zulassen oder die Aufmerksamkeit auf situativ relevante Informationen lenken, insbesondere für zeitkritische Anwendungen in der Entscheidungsunterstützung.

Methode der Erklärung

Das „Wie" der Erklärung hängt vor allem auch von der Methode ab, die eingesetzt wird. Im Folgenden wird zunächst ein kurzer Überblick über gängige Typen und Methoden für XAI gegeben (Tabelle 2), um im Anschluss auf neuere Trends im Forschungsfeld XAI einzugehen. Die XAI-Typen werden jeweils kurz erläutert und beispielhaft einschlägige Methoden ausgewiesen.

Tabelle 2: **Überblick klassischer Typen und Methoden**

Art	Kurzerläuterung	Methoden (Beispiele)
Relevanz-, Feature-, Attributions- und Perturbations-basiert	Basiert auf der Beobachtung von Paaren von Modellein- und -ausgaben (1) Relevanz: Analyse der Relevanz von Informationen (Features, Wörter, Pixel etc.) in der Eingabe auf die Ausgabe (Erklärbox 3) (2) Aktivitätsmaximierung: Eingaben finden oder generieren, die bestimmte Modellbestandteile maximal aktivieren (3) Perturbation: Systematische Variation bzw. Störung von Eingabe(-bestandteilen) (Perturbation), um Einfluss auf Ausgabe(-bestandteile) zu messen	Local Interpretable Model-Agnostic Explanations (LIME), Layer-wise Relevance Propagation (LRP), Shapley Additive Explanation (SHAP). Activation Maximisation, Sensitivitätsanalyse (Perturbation)
Surrogat-basiert	Approximation eines Black-Box-Modells durch ein interpretierbares White-Box-Modell	Surrogatmodelle basierend auf regelbasierten Systemen oder Entscheidungsbäumen, Local Interpretable Model-Agnostic Explanations (LIME)
Wissens-basiert bzw. basierend auf hybrider KI	Vorhandenes menschliches Wissen wird genutzt, um z. B. durch Deep Learning erstellte KI-Modelle erklärbarer zu machen oder um XAI-Methoden zu verbessern	XAI wird durch Wissen in Form von Wissensgraphen oder -basen, logischen Regeln oder algebraischen Formeln informiert
Beispiel-basiert	Erklärungen werden durch Beispiele (aus dem Datensatz oder einer künstlich erzeugten Klasse) repräsentiert oder kommuniziert (Erklärbox 3)	Counterfactuals, Prototypen, Influential Datapoints

Quelle: Zusammenstellung basierend auf Kraus et al. (2021) und Beckh et al. (2023). Die verschiedenen Methoden werden dort ausführlich mit Vor- und Nachteilen und ihrer Anwendungsreichweite, etwa bei verschiedenen Datentypen, dargestellt.

In den vergangenen Jahren ist die Forschung in verschiedenen Richtungen vorangegangen. Eine Auswahl von Methodentrends ist in der Tabelle 3 zusammengestellt. Es handelt sich zum Teil um Ansätze, die in jüngerer Zeit weiterentwickelt wurden und bereits näher an der Anwendung sind, wie etwa interaktive erklärbare KI, und zum anderen um neue Ansätze der XAI-Forschung, wie mechanistische Interpretierbarkeit.

Tabelle 3: **Überblick über Methodentrends**

Art	Kurzerläuterung	Methoden (Auswahl)
Attention-basiert / Perturbations-basiert	Diese Form von XAI setzt am Aufmerksamkeitsfilter an, der z. B. beim Deep Learning genutzt wird, um bei einer Sequenz (z. B. Wort- oder Pixelreihen) die Bedeutung von Einheiten der Sequenz relativ zu anderen Einheiten einer Sequenz zu ermitteln.	Attention manipulation
Alternative Netzwerkarchitekturen	Ex-ante-Nutzung alternativer künstlicher neuronaler Netzwerkarchitekturen, die mehr Nachvollziehbarkeit versprechen.	Kolmogorov-Arnold Networks
Argumentations-basiert	Erklärung von Modellausgaben durch Argumente, die für und gegen diese Modellausgabe sprechen, bzw. durch Abwägen von Belegen für und wider eine Modellausgabe.	Argumentation frameworks
Konzept-basiert	Zuordnung von Konzepten (bei Bildern z. B. Auge, Finger o. Ä.) zu Modelleingaben und Messung der Bedeutung der vordefinierten Konzepte für die Modellvorhersagen (Erklärbox 3).	Concept Relevance Propagation (CRP), Prototypical Concept Explanation (PCX)
Interaktive erklärbare XAI; erklärende Dialoge	Nutzende erhalten Erklärungen und können Nachfragen stellen und ggf. weitere Erklärungen anfordern. Die interaktive Komponente ermöglicht ggf. auch die Anpassung des KI-Modells oder eines Surrogat-Modells durch Feedback (Erklärbox 3 für illustrative Verbindung von generierter Erklärung und Feedback).	Counterfactuals unterstützen das menschliche Feedback, das wiederum das KI-Modell verfeinert, Explanatory Interactive Machine Learning mit CAIPI, Co-Adaptive Guidance
Mechanistische Interpretierbarkeit	Identifikation von Zusammenhängen zwischen gelernten Repräsentationen (auch Konzepte oder Features genannt) und Kombinationen aktivierter Neuronen, in einem künstlichen neuronalen Netzwerk; Identifikation von funktionalen Subnetzen und Schaltkreisen (neural circuits)[5]; ganzheitliche Erklärung der Beziehungen zwischen Repräsentation, Modellverhalten und Daten.	Dictionary learning, sparse auto-encoder, circuits, SemanticLens

Quelle: Auswahl basierend auf Achtibat et al. (2023), Beckh et al. (2023), Deiseroth et al. (2023), Longo et al. (2024), Teso & Kersting (2019), Liu et al. (2024), Nandis (2024), Schmid und Wrede (2022) und Sperrle et al. (2021).

5 Features können als ein Cluster von künstlichen Neuronen betrachtet werden, die auf ein Konzept zurückzuführen sind. Circuits sind Gruppen von Features, die miteinander verbunden sind und für bestimmte Aufgaben im Neuronalen Netzwerk zuständig sind.

Erklärbox 3

Illustration Relevanz-, Beispiel- und Konzept-basierter Erklärungen

Relevanz-basiert

Die XAI-Komponente des KI-Systems gibt Nutzenden eine Erklärung der Ausgabe auf Basis der für die Klassifikation relevanten Pixel. Nutzende erkennen, dass hier ein Teil der Pixel irrelevant ist.

Beispiel-basiert

Die XAI-Komponente des KI-Systems gibt Nutzenden ein Referenzbeispiel. Nutzende erkennen, dass das Referenzbeispiel falsch klassifiziert ist.

Konzept-basiert

Die XAI-Komponente des KI-Systems liefert Nutzenden Erklärung(en) auf der Basis von Konzepten, wie etwa „schwarze Flügelfarbe", bzw. auf Basis der Verknüpfung von Konzepten. Nutzende erkennen mögliche Fehlklassifikationen oder Fehler in der Verknüpfung der Konzepte.

Quelle: Darstellung basiert auf Teso et al. (2023).

Evaluierung von Erklärungen

Eine Erklärung muss korrekt sein, um den verschiedenen Zielgruppen Informationen zu liefern, auf die sie sich verlassen können. Da XAI-Methoden bei Black-Box-Modellen die Modelleingaben (Trainingsdaten) auf die Modellausgaben (z. B. Klassifikationen, Vorhersagen) abbilden, kann es jedoch vorkommen, dass eine Erklärung nicht oder nicht vollständig korrekt ist. Daher wird in der XAI-Forschung an Methoden zur Bewertung von Erklärungen geforscht, wobei verschiedene Kriterien wie Modelltreue (Faithfulness) und Robustheit gemessen werden (Hedström et al. 2023):

a) Die Modelltreue misst dabei, inwiefern die Erklärung und das tatsächliche Modellverhalten übereinstimmen.

b) Die Robustheit misst die Stabilität von Erklärungen bei leichten Störungen der Eingaben unter der Annahme, dass die Modellausgabe ähnlich bleibt.

Um Erklärungen neuronaler Netzwerke zu evaluieren, werden Metriken für die Bewertung von Erklärungen entwickelt und als Programmbibliotheken bereitgestellt (z. B. das Quantus-Toolkit oder das explAIner Framework, siehe Hedström et al. 2023, Spinner et al. 2019).

Fokus: XAI für generative KI

Ein Großteil der XAI-Forschung fokussiert bisher auf KI-Modelle, die für Aufgaben der Klassifikation vorgesehen sind.[6] Mit den Fortschritten in der Entwicklung generativer KI-Modelle, der Verbreitung dieser Technologie über einfach zu bedienende Weboberflächen mit zum Teil über 100 Millionen von aktiven Nutzenden und dem Einsatz in verschiedensten Sektoren und Applikationen, gewinnt die Frage der Erklärbarkeit solcher Modelle an Bedeutung, das heißt die Erklärung ihrer Vorhersagen, die Inspizierbarkeit ihrer internen Repräsentationen und die Verifizier- und Kontrollierbarkeit ihres Verhaltens. Die Nutzendengruppe ist entsprechend ihrer Größe sehr vielfältig und damit sind die adressatengerechten Anpassungen entsprechend komplex. Die Erklärbarkeit generativer Modelle ist vor allem auch deshalb relevant, weil ihre Ausgaben, wie die anderer KI-Modelle auf Basis von Machine Learning, nicht zuverlässig korrekt sind und sogar irreführend sein können (im Fall von generativer KI, speziell Sprachmodellen, äußert sich dies u. a. in sog. Halluzinationen). Erklärungen durch das KI-Modell selbst, wenn man etwa den Chatbot fragt, wie er zu einem Ergebnis kommt, sind demselben Problem unterworfen. Auch Fragen der Sicherheit, etwa wenn ein Sprachmodell Anleitungen für die Herstellung von Biowaffen oder Sprengkörpern ausgibt, oder Fragen der Rechenschaftspflicht sind ohne Anwendung von XAI vermutlich effektiv nicht zu lösen. Schließlich ist ein Verständnis des Modells mittels XAI auch sinnvoll, um die Verwendung von generativen Modellen durch die Nutzenden zu unterstützen (Prompt Engineering) und so bessere Ergebnisse zu erzielen. Zudem kann ein besseres Verständnis grundsätzlich auch zur langfristigen Entwicklung besserer Modellarchitekturen führen.

Generative KI stellte die XAI-Forschung jedoch vor neue Herausforderungen, denn durch die schiere Größe der KI-Modelle (Sprachmodelle, multimodale Modelle) und Trainingsdatensätze sowie die Komplexität der KI-Systeme und der Modellausgaben sind viele klassische XAI-Methoden (z. B. Perturbationen) nicht mehr anwendbar oder stoßen auch bei einer Anpassung der Methoden für generative Modelle an Grenzen. Es zeichnet sich

6 Siehe Samek (2025) sowie Schneider (2024) für eine tiefergehende Auseinandersetzung mit den Themen dieses Abschnitts.

ab, dass mit Konzept-basierten Ansätzen, mechanistischer Interpretierbarkeit sowie holistischen Ansätzen (z. B. SemanticLens) neue Wege jenseits der weit verbreiteten Attributions-basierten Ansätze beschritten werden müssen, wobei sich wegen der Größe der Modelle Fragen zur Skalierbar- und Automatisierbarkeit stellen. Schließlich ist bei generativer KI der Aspekt der Kontrolle gegebenenfalls wichtiger als der Aspekt der Erklärung. Das heißt: Es müssen Ansätze entwickelt werden, die es Menschen erlauben, Modellbestandteile zu identifizieren, die im Zusammenhang mit (un)erwünschtem Modellverhalten stehen, um KI-Modelle kontrollieren und adaptieren zu können (sog. „Model Steering"), ohne dass ein erneutes Modelltraining nötig wird. Vor dem Hintergrund, dass aktuelle große KI-Modelle oft nachtrainiert und aufwendig angepasst werden müssen, sind solche XAI-basierten Methoden zur Modellkorrektur besonders erstrebenswert.

Fokus: Interaktive erklärbare Künstliche Intelligenz (interaktive XAI)

Interaktive XAI bezieht die Endnutzenden in den maschinellen Lernprozess mit ein. Den Nutzenden werden die Modellvorhersagen und Klassifikationen nachvollziehbar erklärt und sie haben ihrerseits die Möglichkeit, korrigierend in den Lernprozess einzugreifen, sei es auf Modell- oder auf Datensatzebene (Schmid 2024). So zeigen anwendungsnahe Studien zu interaktiver XAI, dass beispielsweise Modelle zur Krebserkennung oder zur Anomalieerkennung in der industriellen Qualitätskontrolle verbessert werden können (Plattform Lernendes System 2023, Schmid & Finzel 2020, Gramelt, Höfer & Schmid 2024). Es konnte zudem demonstriert werden, dass durch interaktive XAI die Vorhersage- und Erklärungskraft von KI-Modellen erhöht und damit mehr Vertrauen beziehungsweise begründetes Misstrauen in KI-Systeme aufgebaut werden kann (Teso und Kersting 2019, Gramelt, Höfer & Schmid 2024). Darüber hinaus spielt interaktives XAI, wie bereits gezeigt (siehe Abschnitt 3), unter anderem bei der Kontrolle von Systemen mit höheren Autonomiegraden und bei der Anpassung von KI-Systemen an Domänen und Individuen eine wichtige Rolle. So kann durch ein geeignetes User-Experience-Design den Nutzenden die Möglichkeit gegeben werden, durch Interaktivität im Dialog mit dem KI-System verschiedene Ebenen und Formen der Erläuterung abzurufen (Finzel et al. 2021).

Interaktive XAI sieht sich jedoch mit verschiedenen Herausforderungen konfrontiert (Teso et al. 2023, S. 12):

- Menschliche Eigenschaften: Es können Unklarheiten über den semantischen Gehalt von Interaktionsobjekten (Wörter, Bilder, Symbole etc.) auftreten, zum Beispiel, weil die Bedeutung generell mehrdeutig ist oder sich von Mensch zu Mensch unterscheidet. Auch die menschliche Wahrnehmung von Beispielen (siehe Beispiel-basierte XAI) kann eine starke subjektive Komponente aufweisen. Gerade in der Interaktion mit KI-Systemen, in der es nicht nur um das Verstehen, sondern auch um ein sinnvolles Feedback für das System geht, können solche menschlichen Eigenschaften eine Herausforderung darstellen.

- Mangelnde Modelltreue einer Erklärung kann zu einem Problem in der Interaktion werden, wenn etwa die Erklärung den Nutzenden einen vermeintlichen Klassifikationsfehler nahelegt und dieser ein entsprechendes Feedback an das System gibt.

- Die Kommunikation zwischen Menschen findet meist auf der Basis von abstrakten, aber aussagekräftigen Konzepten statt. Wie KI-Systeme dazu befähigt werden können, auf Basis solcher Konzepte zu kommunizieren, ist weiterhin eine bedeutende Forschungsfrage, gerade in der Forschungsrichtung zu Konzept-basierter XAI.

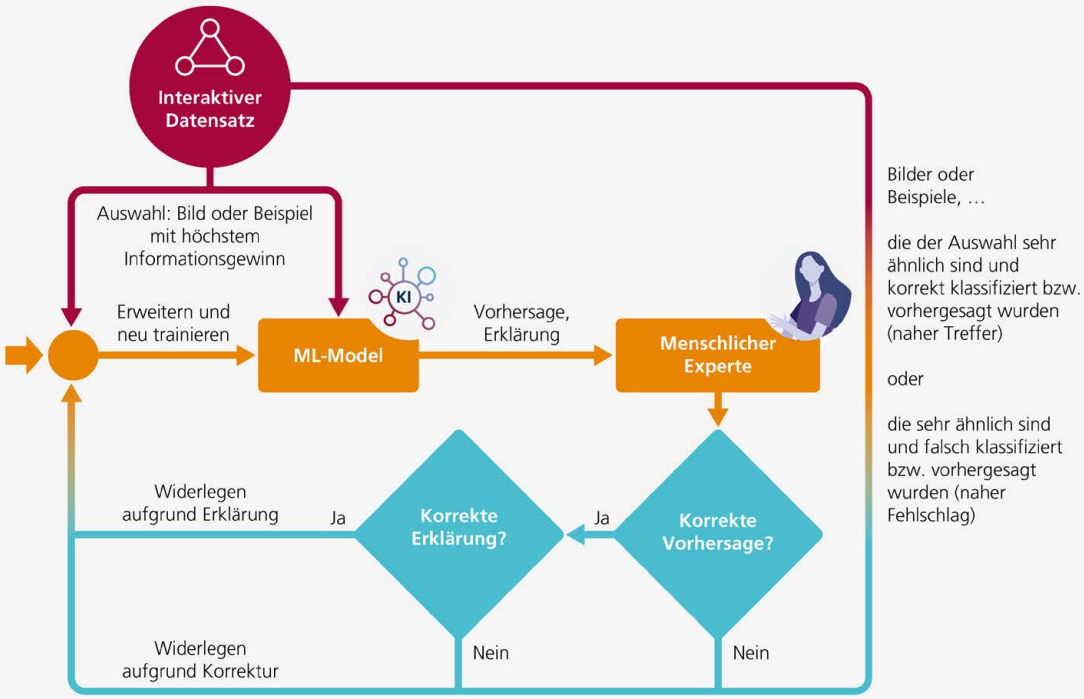
- Die Beziehung zwischen Erklärungen und Vertrauen auf Seiten der Nutzenden ist nicht zwingend eindeutig, so kann beispielsweise eine Erklärung, die nicht sofort verständlich ist, zu ungerechtfertigtem Misstrauen gegenüber dem System führen. Gerade die Beziehung zwischen interaktiver XAI und (Nutzenden-)Vertrauen ist noch nicht genügend erforscht.[7]

- Wenn Nutzende in der Interaktion korrigierende oder annotierende Informationen an das KI-System weitergeben, kann dies „Rauschen" (z. B. ungenaues Feedback, Flüchtigkeitsfehler etc.) in den Lernprozess einführen und zu Aufwand und Kosten auf Seiten der Nutzenden führen. Der Umgang mit diesen Phänomenen bedarf weiterer Forschung.

- Die Evaluierung von erklärbarer interaktiver XAI stellt eine besondere Herausforderung dar. Es fehlt vor allem an standardisierten Benchmarks und Metriken.

Erklärbox 4

Prozess interaktiver XAI – ein Beispiel



Quelle: Gramelt, Höfer und Schmid (2024).

7 Siehe hierzu auch das BMBF-geförderte Projekt Ethyde.

6 Zielgruppen der Erklärung und ihre Charakteristika

Die Erwartungen und Anforderungen an die Erklärung von KI-Systemen werden je nach Zielgruppe unterschiedlich sein (Tabelle 2, Abbildung 3 und Bahlmann et al. 2024). Dies hängt maßgeblich von Faktoren wie dem Vorwissen der Betroffenen zu KI oder dem Anwendungsgebiet, aber auch von Zeitdruck, Motivation und Interessen ab. Gute Erklärungen werden daher auf die jeweilige Zielgruppe zugeschnitten beziehungsweise anpassbar sein müssen. Grundsätzlich lassen sich drei Zielgruppen unterscheiden (Tabelle 2).

Bei der ersten Gruppe steht die Validierung von Modellen im Vordergrund. XAI soll Nutzende, wie Modellentwickelnde und KI-Forschende, vor allem während des Modelltrainings unterstützen. Diese Gruppe möchte XAI nutzen, um die Datenbasis zu untersuchen, Modelle zu explorieren und Fehler zu beseitigen (Debugging). Die Erklärungen sollen modellgetreu sein und die Nutzenden in die Lage versetzen, Modelle zu verbessern.

Eine weitere Zielgruppe stellt menschliche Werte in den Vordergrund. Ihr geht es um eine vertrauenswürdige KI, die rechtliche Vorgaben einhält und ethische Werte berücksichtigt, sodass Vertrauen in die Modellausgaben gerechtfertigt ist. Erklärungsbedarf besteht vor allem in der Inferenzphase während des aktiven Betriebs der KI-Modelle. Zielgruppe sind vor allem KI-Laien und Endverbrauchende (z. B. Patientinnen und Patienten). Für sie müssen Erklärungen leicht zugänglich und verständlich sein. Wenn KI-Ausgaben für Personen dieser Zielgruppe weitreichende Konsequenzen haben, sollten die Erklärungen ihre Entscheidungs- und Handlungsfähigkeit unterstützen.

Eine dritte Zielgruppe liegt eher zwischen den beiden ersten. Für sie stehen sowohl menschliche Werte als auch die Validierung im Vordergrund. Ihnen geht es einerseits um die Verbesserung von KI-Modellen mit XAI-Methoden und andererseits um die Überprüfung von Kriterien für vertrauenswürdige KI beziehungsweise deren Einhaltung. Produktentwickelnde, für die Vertrauenswürdigkeit gegebenenfalls auch ein Alleinstellungsmerkmal des Produkts ist, gehören ebenso zu dieser Gruppe wie Zertifizierende und Auditierende, die sowohl wertbasierte Kategorien als auch technische Leistungskriterien prüfen. Domänenforschende gehören zu dieser Gruppe, für die sowohl die Verbesserung domänenspezifischer Modelle als auch die Erklärung von Modellausgaben in der Inferenzphase von Bedeutung sind. Die Erklärungen sollten diese Zielgruppe in ihrer jeweiligen spezifischen Tätigkeit unterstützen und befähigen.

In Erweiterung dieser grundsätzlichen Unterscheidung wird im Folgenden der Zusammenhang zwischen dem Wozu, Wann, Wer und Wie von XAI-Erklärungen anhand von illustrativen und fiktiven Personae und Beispielen aus der Literatur noch einmal aus unterschiedlichen Perspektiven dargestellt.

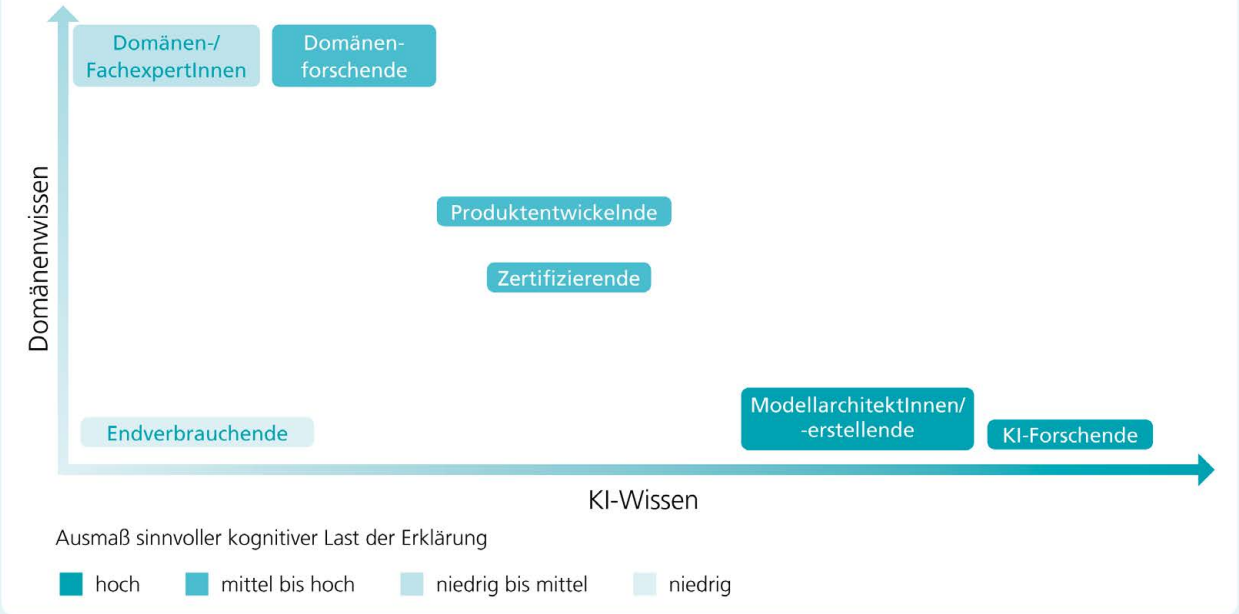
Tabelle 4: **Übersicht XAI-Zielgruppen und ihre Charakteristika**

W-Fragen	Validierungsorientiert	Wertorientiert	Mischtypus
Wozu werden Erklärungen benötigt?	Untersuchung von Datensätzen, Modellexploration, Debugging, Modellverstehen	Verantwortungsvolle Modelle, Rechtskonformität, Vertrauen in Modellausgaben, Ethik	Beides
Wann werden die Erklärungen benötigt?	Eher in der Trainingsphase	Eher in der Inferenzphase	Beides
Wer benötigt die Erklärung (Auswahl)?*	Modellerstellende, KI-Forschende	KI-Laien, Endverbrauchende	Produktentwickelnde, Zertifizierende und Auditoren, DomänenexpertInnen, Domänenforschende
Welche Art der Erklärung benötigt die Zielgruppe?	Modellgetreue vielseitige und komplementäre Erklärungen, die das Modellverständnis und die Entscheidungs- und Handlungsfähigkeit fördern.	Einfache, zugängliche, verständliche und benutzerfreundliche Erklärungen, ggf. bei weitreichenden Folgen der Modellausgabe Förderung der Entscheidungs- und Handlungsfähigkeit.	XAI sollte spezifisch bei der Tätigkeit unterstützen und befähigen.

*Weitere Rollen, die in den Mischtypus fallen können, sind u. a. Ingenieurinnen und Ingenieure, Maschinenbedienende in der Produktion, Produktionsleitende, Produktdesignerinnen und -designer, Entscheiderinnen und Entscheider sowie Managerinnen und Manager.

Quelle: Basierend auf Biecek & Samek (2024) mit Erweiterung um den „Mischtypus" sowie Brath et al. (2023).

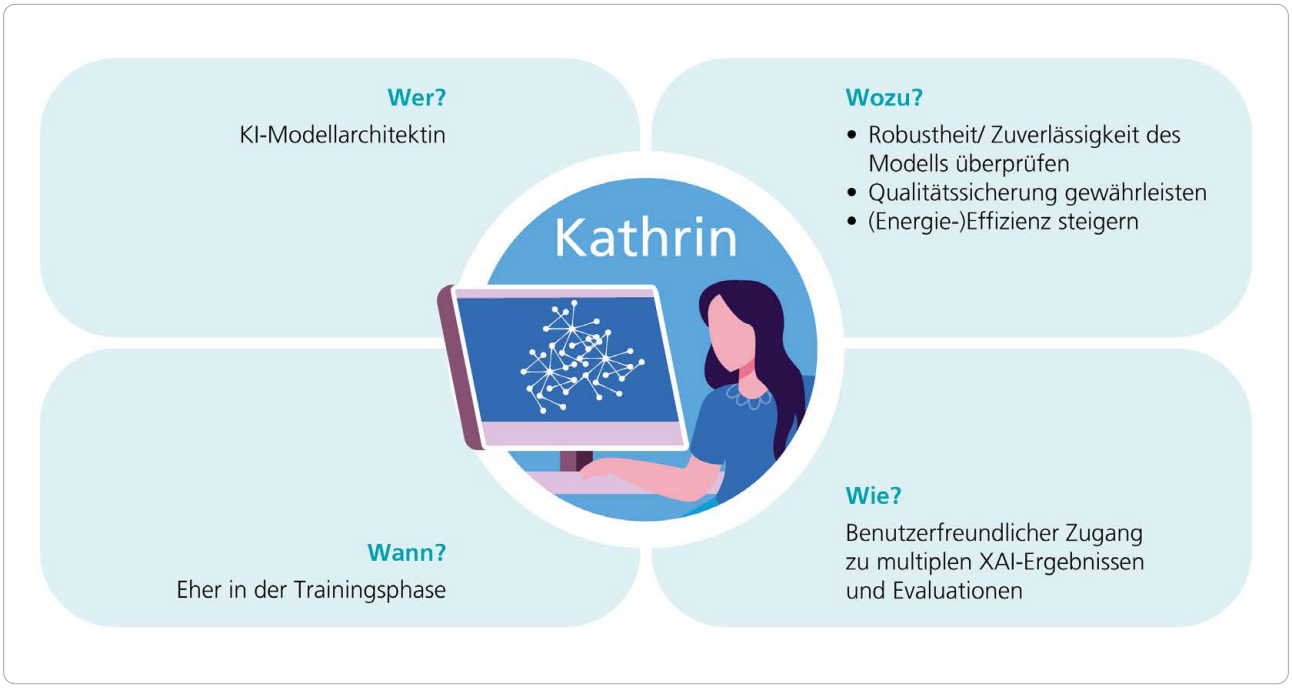
Abbildung 3: **Zielgruppen (Auswahl) nach Domänen- und KI-Wissen**



Quelle: Überblick zu verschiedenen Zielgruppen abgeleitet aus den Personas.

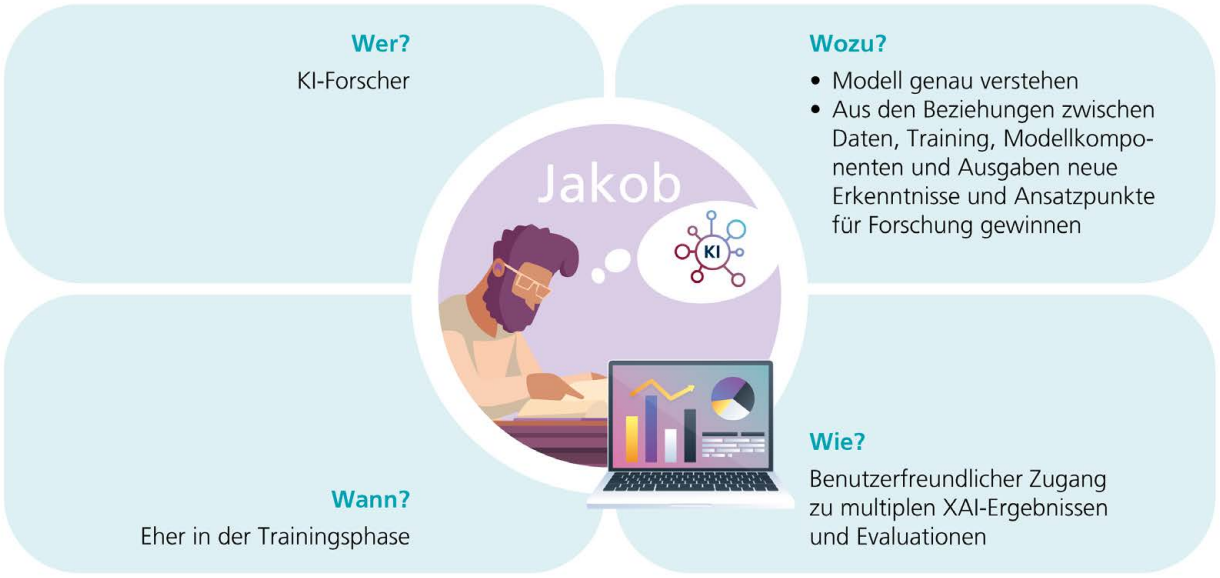
Personae 1: Kathrin – Modellarchitektin und Jakob – KI-Forscher

Kathrin ist KI-Modellarchitektin in einem großen Unternehmen, sie hat ein einschlägiges Studium absolviert und einige Jahre Berufserfahrung gesammelt. Sie verfügt daher über ein hohes Maß an technischem Wissen und fortgeschrittenen Modellierungstechniken. Obwohl sie im Rahmen ihrer Arbeit immer wieder mit verschiedenen Anwendungsdomänen konfrontiert wurde, ist sie keine Domänenexpertin mit tiefgreifendem Erfahrungswissen. Ihre Motivation ist es, die bestmöglichen KI-Modelle zu erstellen, das heißt Modelle, die auf möglichst effiziente Weise genau das tun, was sie sollen. Sie verwendet XAI-Methoden als eine Reihe von Werkzeugen, um Modelle zu verstehen und zu überarbeiten, um Modelle beispielsweise vor dem Hintergrund der Qualitätssicherung in Unternehmen zu verbessern. Zu diesem Zweck verwendet sie eine Benutzeroberfläche, die es ihr ermöglicht, durch die verschiedenen, komplementären Ergebnisse der XAI-Methoden sowie durch die Ergebnisse der Evaluierung zu navigieren.



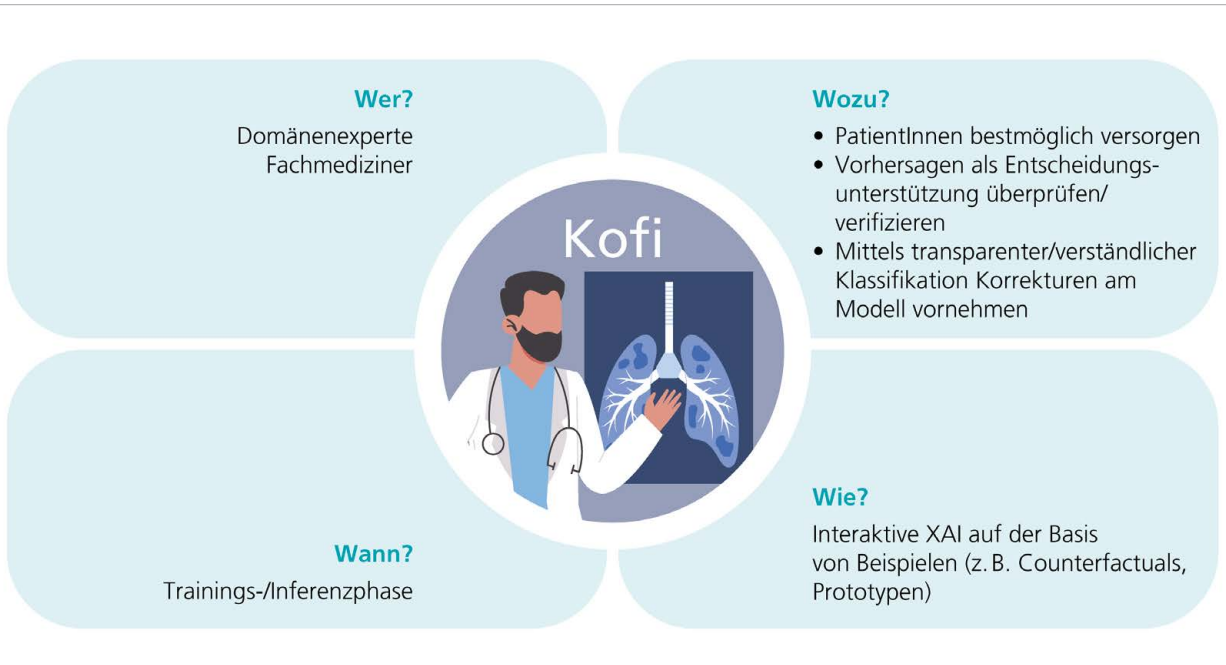
Auf diese Weise kann Kathrin das Modell erkunden und sein Verhalten aus verschiedenen, sich teilweise ergänzenden oder sogar widersprechenden Perspektiven beobachten. Sie nutzt diese Schnittstelle, um Fehler und Verzerrungen im Datensatz aufzuspüren, um herauszufinden, wie bestimmte Modellausgaben entstanden sind und auf welche Komponenten des Modells (z. B. aktivierte Neuronen eines tiefen neuronalen Netzes) und des Datensatzes sie zurückzuführen sind. Sie identifiziert unbedeutende Teile des Modells, die für den Zweck des Modells nicht relevant sind, und erhöht so die (Energie-)Effizienz. Sie verwendet XAI in erster Linie als Ingenieurswerkzeug. Mittels Model Monitoring, zu dem auch XAI-Methoden beitragen können, versucht Kathrin Informationen darüber zu sammeln, ob die Robustheit und Zuverlässigkeit des Modells auch über die Zeit gegeben ist oder ob gegebenenfalls Maßnahmen ergriffen werden müssen. So können zum Beispiel auch Reputationsschäden für Organisationen verhindert werden.

Jakob, ein KI-Forscher, weist ähnliche Charakteristika auf. Allerdings hat er sich im Laufe der Jahre ein noch tieferes Wissen über KI angeeignet. Seine Motivation liegt eher darin, das Modell sehr genau zu verstehen und aus den Beziehungen zwischen Daten, Training, Modellkomponenten und Ausgaben, die durch die Modellexploration sichtbar werden, neue Erkenntnisse und Ansatzpunkte für seine Forschung zu gewinnen.



Personae 2: Kofi – Domänenexperte und Aliya – Domänenforscherin

Kofi ist als Facharzt an einer Universitätsklinik ein Domänenexperte und verfügt über ein ausgeprägtes Fach- und Erfahrungswissen über die Zusammenhänge, Aufgaben, Datentypen und Datensätze in der Medizin.

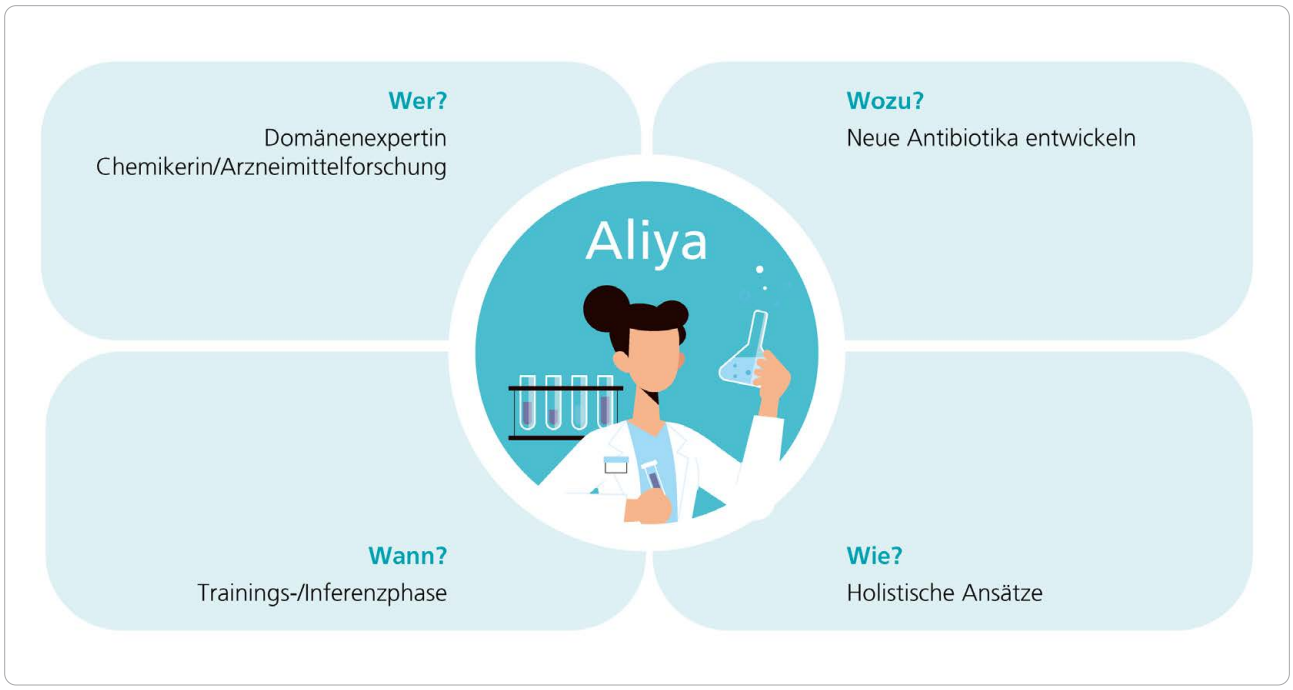


Seine Motivation ist es, Menschen in der ihm zur Verfügung stehenden Zeit bestmöglich medizinisch zu versorgen. Er hat noch kein tieferes Wissen über KI und KI-Methoden, ist aber sehr motiviert, sich mit KI zu beschäftigen, wenn sie in seinem Fachgebiet eingesetzt wird, einen echten Mehrwert für die Patientinnen und Patienten verspricht und grundsätzlich vertrauenswürdig ist. Kofi wünscht sich daher intuitiv zugängliche, schnell erfassbare, quantifizierbare und nachvollziehbare Erklärungen der Modellausgaben, die es ihm einerseits ermöglichen, auf Basis der KI-Vorschläge zu handeln, andererseits aber auch, die Vorschläge kritisch zu hinterfragen, ohne ihn mit zu vielen Informationen zu überfordern. Im besten Fall kann er durch eine Feedbackschleife das KI-System auf Basis seines Fachwissens verbessern und so die Diskrepanz zwischen Modellen und Domänenwissen minimieren. Insgesamt wünscht er sich dafür eine einfache und intuitive Benutzeroberfläche, damit die Ergebnisse auch unter Zeitdruck schnell eingeordnet werden können. Interaktive erklärbare KI auf der Basis von Beispielen (z. B. Counterfactuals, Prototypen) kann hier eine Lösung darstellen.

Tabelle 5: **Beispiele XAI in der Domäne Medizin**

Quelle	Beispiel	Zielsetzung	Geeignete Methoden
van Aken et al. (2022) und Mullenbach et al. (2018)	Vorhersagen von Sprachmodellen in der Medizin	Überprüfung und Verifikation von Vorhersagen, um diese schnell als Entscheidungsunterstützung nutzen zu können.	prototypische Netzwerke, Diagnose-spezifische Attention
Schmid und Finzel (2020)	Klassifikationen von Tumorgewebe	Klassifikation transparent und verständlich machen, um auch interaktiv Korrekturen am Modell vornehmen zu können (Modell verbessern).	Hybride KI, Surrogat-Modell, interaktive Komponente

Aliya ist Chemikerin. Als Domänenforscherin verfügt sie über ein sehr tiefes und breites Wissen in ihrem Fachgebiet und insbesondere in ihrem Spezialgebiet der Arzneimittelforschung. Durch ihr Studium und ihre Forschungsarbeit ist sie mit Statistik und statistischer Modellierung in ihrem Fachgebiet vertraut.

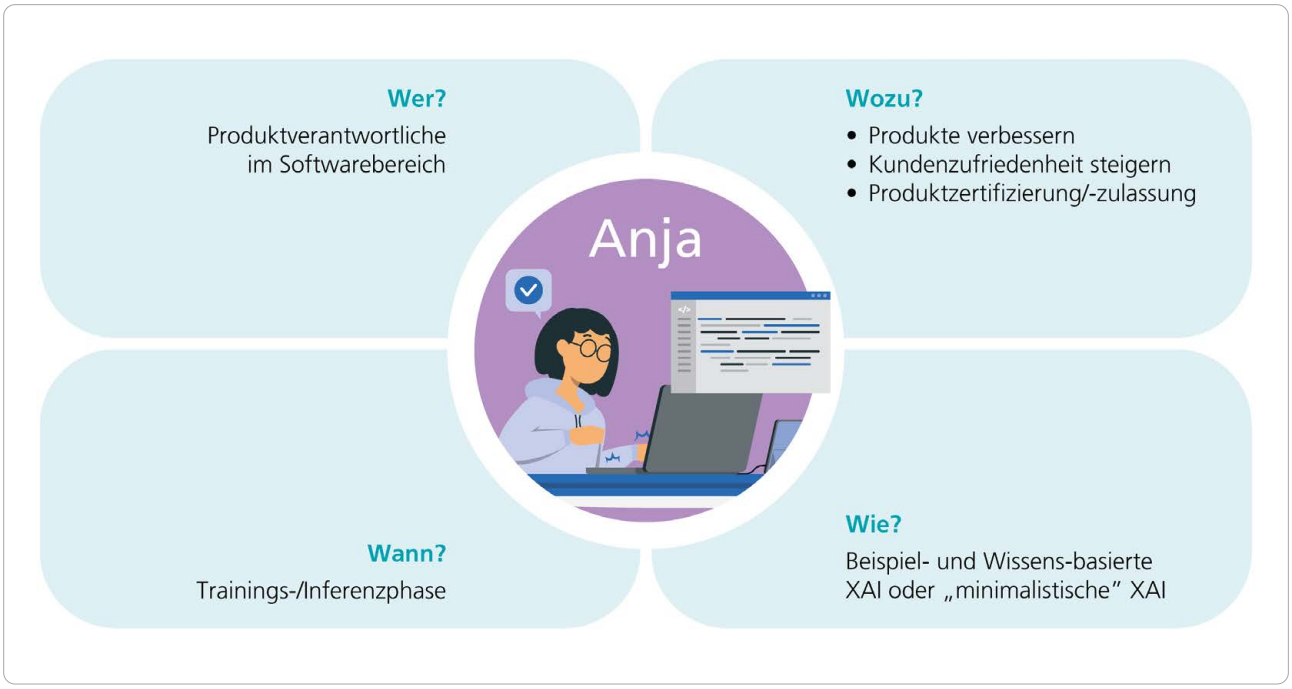


Sie interessiert sich besonders für Antibiotika und möchte den Stand der Forschung erweitern und Wege finden, neue Antibiotika zu finden. Als Domänenforscherin investiert sie gerne mehr Zeit in die Analyse von Informationen, wenn es sich lohnt. So können die Erklärungen der XAI-Methoden auch umfangreicher und komplexer sein, um daraus Erkenntnisse für die Forschung abzuleiten. Sie ist an XAI interessiert, weil Erklärungen helfen können, einen Untersuchungsgegenstand zu präzisieren oder Ansatzpunkte für neue Hypothesen und Untersuchungen von Kausalzusammenhängen zu finden, mögliche Erklärungen für Phänomene zu identifizieren, aber auch, um gegebenenfalls Modelle in ihrem Bereich zu verbessern. Gerade die verborgenen Zusammenhänge in den Daten, die KI-Modelle für ihre Vorhersagen nutzen, können durch XAI erhellt werden. Da sie als Forscherin von Neugier getrieben ist, möchte sie insbesondere bei neuronalen Netzen verstehen, was dahintersteckt, und findet es spannend, dass heute an Netzwerkarchitekturen geforscht wird, die von vornherein interpretierbarer sind und für die an Beispielen aus der Mathematik und Physik gezeigt wurde, dass durch sie Gesetzmäßigkeiten (wieder-)entdeckt werden können, sowie an Methoden, um die Funktion von Neuronengruppen nachvollziehen zu können (z. B. durch holistische Ansätze).

Tabelle 6: **Beispiel XAI in der Domänenforschung**

Quelle	Beispiel	Zielsetzung	Geeignete Methoden
Wong et al. (2024)	Modellvorhersagen in der Antibiotika-Forschung	Identifizierung und Entwicklung neuer Antibiotika unter anderem durch Akkumulation von Wissen über die Hintergründe zu den Vorhersagen eines Modells.	Monte Carlo tree search Dieser Suchalgorithmus ermöglicht es dem Modell, nicht nur eine Schätzung der antimikrobiellen Aktivität eines jeden Moleküls zu erstellen, sondern auch eine Vorhersage darüber, welche Substrukturen des Moleküls wahrscheinlich für diese Aktivität verantwortlich sind.

Persona 3: Anja – Produktverantwortliche



Anja ist Produktverantwortliche in einem Unternehmen, das Anwendungen mit KI-Funktionalität anbietet. Sie hat Wirtschaftsinformatik studiert und kennt daher sowohl die betriebswirtschaftlichen Aspekte von Softwareanwendungen als auch die technische Seite der Softwareentwicklung. Bereits während ihres Studiums hat sie sich mit KI-Anwendungen beschäftigt. Vor diesem Hintergrund ist sie hinsichtlich ihres KI-Wissens zwischen den Modellerstellenden, den KI-Forschenden und den Endanwendenden anzusiedeln und übersetzt die Bedürfnisse der Anwendenden in Anforderungen an das Modell. Durch ihre Berufserfahrung verfügt Anja über ein fundiertes Wissen über das Marktumfeld zukünftiger KI-Anwendungen, die Bedürfnisse und Wünsche der Konsumentinnen und Konsumenten sowie über ein grundlegendes Wissen über die Spezifika der Anwendungsdomäne. Dazu gehören auch Benutzerfreundlichkeit und Qualitätssicherung (z. B. funktionale Sicherheit und Zuverlässigkeit). Auch in der Produktentwicklung setzt sie XAI-Methoden ein, um einerseits die Interaktion zwischen Nutzenden und dem KI-System zu optimieren und andererseits die Integration des Modells in die Anwendungsdomäne zu verbessern und damit die Kundenzufriedenheit zu erhöhen. Passende XAI-Methoden wären für Anja unter anderem Beispiel- und Wissens-basierte XAI oder auch „minimalistische" XAI, die auf zugrundeliegende, verwendete Daten verweisen kann. Da sie die Produktzertifizierung und -zulassung im Hinterkopf behalten muss, tauscht sie sich mit unternehmensinternen Auditierenden aus, die ebenfalls XAI nutzen. Während die Erklärungen in der internen Entwicklungsphase komplex ausfallen können, sollten sie für die Endnutzenden einfach zugänglich sein. Als Produktentwicklerin wünscht sich Anja Erklärungen, die sie und ihr Team in ihrem spezifischen Tätigkeitsprofil unterstützen.

Erklärbox 5

XAI-Werkzeuge in Unternehmen

Ressourcen

Christoph Molnar (2024): Interpretable Machine Learning *A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/

Weitere:
Onlinekurs erklärbare Künstliche Intelligenz des KI-Campus (derzeit in Überarbeitung)

Geeignete Methoden in der Praxis

In Unternehmen werden beispielsweise klassische XAI-Methoden, wie LIME und SHAP, eingesetzt, aber auch neuere Herangehensweise wie Partial Dependency Plots und Layer-wise Relevance Propagation.

Persona 4: Akira – Zertifizierender und Auditor

Akira ist ein Zertifizierer für KI-Anwendungen in einer unabhängigen KI-Prüfstelle. Er verfügt über ein fundiertes Verständnis der Kriterien für vertrauenswürdige KI, der Standards (ISO/IEC 42001 etc.) und regulatorischen Anforderungen sowie deren Prüfung (Tabelle 1). Er hat Informatik studiert, verfügt über ausgeprägte analytische Fähigkeiten und Berufserfahrung in der Entwicklung von KI und im Risikomanagement. Aufgrund seiner Berufserfahrung und der Tatsache, dass er in seinem Beruf mit KI-Anwendungen aus verschiedenen Domänen konfrontiert wird, hat er auch ein solides Domänenwissen. Sein Wissen über KI bewegt sich zwischen den Modellerstellenden und den KI-Forschenden auf der einen Seite und Endnutzerinnen und -nutzern auf der anderen Seite. Er hat generell ein starkes Interesse an KI und seine Motivation, sich mit KI zu beschäftigen, ist besonders hoch, wenn die Prüfkriterien seiner Prüfstelle betroffen sind. Daher interessiert er sich für Prüfungskataloge. Dabei interessieren ihn sowohl die Erklärbarkeit der Daten und des Modells als auch der Modellausgaben. Da Erklärbarkeit als Teil des Kriteriums Transparenz ebenfalls Teil des Prüfkatalogs ist, benötigt er auch Evaluationswerte für die jeweilige XAI-Komponente der KI-Anwendung. Insgesamt wünscht er sich eine Zertifizierungsanwendung mit einer einfach zu bedienenden Oberfläche, die genau auf die Bedürfnisse der Prüfstelle und deren Kriterienkatalog zugeschnitten ist, sodass seine Entscheidungsfähigkeit bei der Prüfung verschiedener Anwendungen unterstützt wird und die Prüfung effizient durchgeführt werden kann. Als Zertifizierer wünscht sich Akira Erklärungen, die ihn und sein Team in ihrem spezifischen Tätigkeitsprofil unterstützen.

Persona 5: Tobias – Endverbrauchender

Tobias ist ein Durchschnittsbürger. Wie rund 50 Prozent der Bevölkerung schätzt er sein Wissen über KI weder hoch noch niedrig ein (3 auf einer Skala von 1 bis 5, vgl. CAIS 2024). Im Vergleich zu Modellerstellenden, Forschenden, Zertifizierenden sowie Produktentwickelnden ist sein Wissen über KI jedoch als eher gering einzustufen. Tobias hat in der Vergangenheit gelegentlich KI-Anwendungen genutzt, wie etwa Texte mit DeepL übersetzt und mit ChatGPT experimentiert. Da in diesen Anwendungsfällen aus seiner Sicht nicht viel auf dem Spiel steht, ist er nicht unbedingt an einer Erklärung der Modellausgabe interessiert. Als er einen Kredit aufnehmen will, erfährt er, dass die Bank ein KI-System zur Bonitätsprüfung einsetzt. Tobias steht dem Einsatz von KI in Finanzinstituten neutral gegenüber (Männer geben im Durchschnitt eine 3 auf einer Skala von 1 bis 5 an, CAIS 2024). Als sein Kreditantrag abgelehnt wird, weil das KI-System der Bank das Kreditausfallrisiko als zu hoch einschätzte, will er genau wissen, wie das Ergebnis der KI zustande kam, und bittet um eine Erklärung. Daraufhin erhält er von der Bank eine Erklärung mit einem Counterfactual zur Erläuterung der KI-Ergebnisse, sodass er nachvollziehen kann, welche Faktoren ausschlaggebend sind und wie sich diese verändern müssen, um den Kredit bewilligt zu bekommen. Trotz des negativen Ergebnisses hat er eine Information erhalten, die es ihm ermöglicht zu handeln, da er nun weiß, an welchen Faktoren er arbeiten kann. Mit Hilfe der XAI-Erklärung kann Tobias auch überprüfen, ob seine persönlichen Daten, die in das KI-Modell eingeflossen waren, korrekt sind. Für Durchschnittsnutzende wie Tobias sind Beispiel-basierte Erklärungen wie Counterfactuals und Prototypen hilfreich, aber auch konzeptbasierte Erklärungen oder Saliency Maps (zum Beispiel für Bilder), da sie Erklärungen liefern, die in der Regel einen niedrigschwelligen Zugang ermöglichen und leicht verständlich sind.

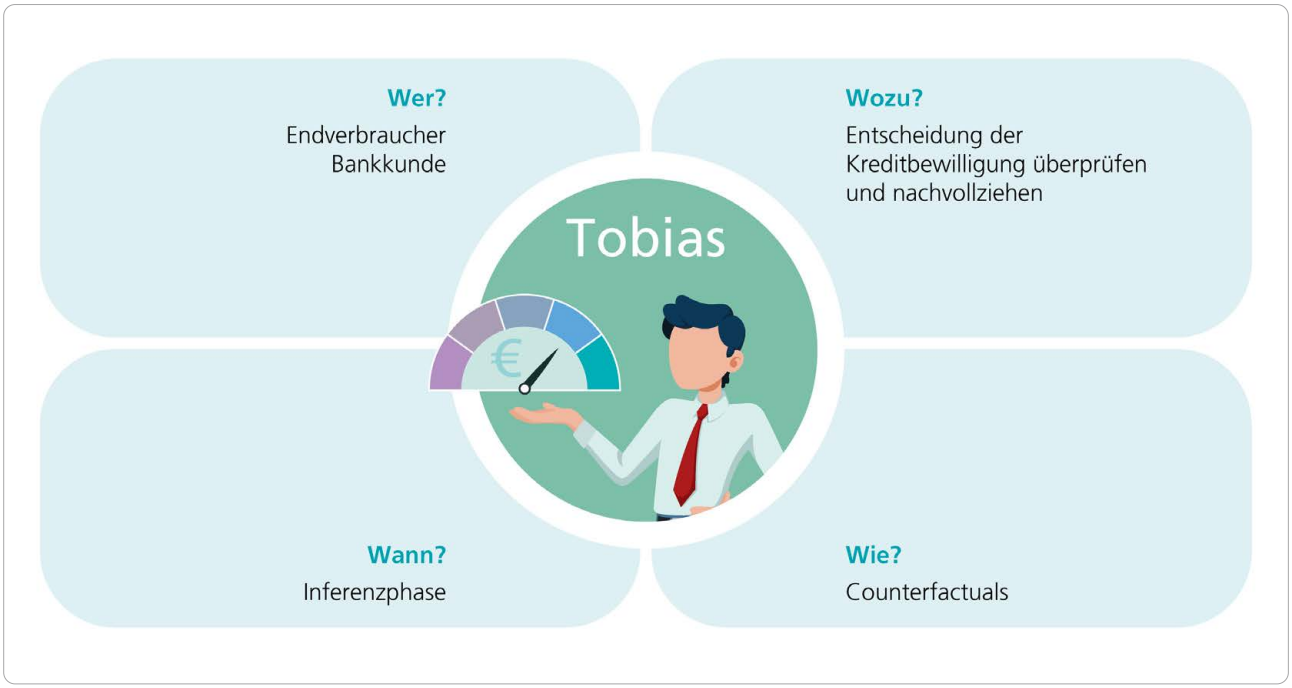


Tabelle 7: **Beispiele XAI für Endverbrauchende**

Quelle	Beispiel	Zielsetzung	Geeignete Methode
Busse (2023)	Klassifikation eines Bewerbenden	Szenario: Einem/r Bewerbenden wird mitgeteilt, welche Faktoren sich ändern müssten, damit er/sie die Stelle bekommen hätte.	Counterfactuals
Aljuneidi et al. (2024)	Verwaltungs-entscheidung	Szenario: Ein Bürger sieht sich mit der Entscheidung eines KI-Systems einer Behörde konfrontiert – einer Geldstrafe. Er erhält eine Erklärung anhand ausschlaggebender Faktoren (und ihrer Rangfolge).	LIME

Herausforderungen zielgruppenspezifische XAI

Eine aktuelle Bestandsaufnahme der Herausforderungen und Lösungsansätze in der Forschung findet sich unter anderem bei Biecek und Samek (2024), Longo et al. (2024) und Schmid und Wrede (2022) sowie Schwalbe und Finzel (2024). Im Folgenden wird auf diese Literatur zurückgegriffen und insbesondere auf die Herausforderungen der Zielgruppenorientierung eingegangen. Wie die verschiedenen Personae zeigen, können Anforderungen und Erwartungen an XAI sehr unterschiedlich ausfallen. Dies stellt die Forschung und Entwicklung von XAI vor eine Reihe von Herausforderungen, allen voran die Frage, wie Erklärungsgegenstand (was), Erklärungsadressat (wem), Erklärungsziel (wozu) und Erklärungsdurchführung (wie) im jeweiligen Fall ideal zu verknüpfen sind.

Anpassung von XAI an Ziele, Zielgruppen und Domänen

- **Ziele:** Auch die Ziele derselben Nutzerinnen und Nutzer können sich je nach Kontext unterscheiden, wie die Persona der Produktentwicklerin zeigt, die zu einem Zeitpunkt eher die Kundenzufriedenheit und zu einem anderen Zeitpunkt eher die (Kosten-)Effizienz des KI-Modells im Blick hat. Das Zeitbudget kann sich ebenfalls je nach Zeitdruck von Situation zu Situation unterscheiden und damit auch, wie stark sich Nutzende mit welchem Ziel auf eine Erklärung einlassen können.

- **Zielgruppen:** Es ist eine anspruchsvolle Aufgabe, Erklärungen zu entwerfen und anzupassen, die sowohl inhaltlich als auch in Bezug auf Format und Präsentation für jede Art von Erklärungsadressaten geeignet sind. Insbesondere aufgrund der unterschiedlichen Bildungshintergründe, Sprachgewohnheiten, Bedürfnisse und Erwartungen.

- **Domänen:** Jede Domäne bringt ihre eigenen Annahmen, Wissensbestände, Umgebungen, Erwartungen und Herausforderungen mit sich.

Menschenzentrierte XAI

Bei der Anpassung von XAI an unterschiedliche Zielgruppen sollte stets der Mensch im Mittelpunkt stehen (siehe hierzu auch das Forschungsgebiet Human-Centered AI).

- **Personalisierung:** Im Idealfall sollten die Nutzenden in der Lage sein, die Erklärungsperspektive anzupassen und im Dialog mit dem KI-System auf natürliche Weise zu kommunizieren, um ihre individuellen Informationsbedürfnisse zu decken. Dies kann bedeuten, dass Nutzende bestimmte Erklärungsformate (z. B. Bild, Text, quantitative Darstellungen etc.) auswählen und die Erklärungen durch Interaktion selbst verfeinern können.

- **Handlungsfähigkeit:** Sollen Erklärungen auch handlungsleitend und befähigend sein und nicht nur das Verständnis des KI-Systems verbessern, erhöht sich die Komplexität für mögliche technische Lösungen. Erklärungen sollten dann auch hinterfragt werden können und es sollte eine Begründung des Prozesses von der Eingabe bis zur Ausgabe eines KI-Systems enthalten sein, die nicht nur erklärt, warum der Prozess technisch korrekt ist, sondern auch warum

er rechtlich und ethisch vertretbar ist. Die Stärkung der Handlungsfähigkeit berührt weiterhin den Aspekt der Kausalität. XAI-Methoden wie auch die KI-Modelle selbst basieren häufig auf Korrelationen und nicht auf Kausalitäten, was einer handlungsanleitenden Funktion von Erklärungen abträglich ist, insbesondere wenn die Entscheidung eines KI-Systems wiederum Menschen betrifft.

Evaluation für zielgruppenorientierte XAI

Wie kann beurteilt werden, ob eine Erklärung für eine bestimmte Zielgruppe auch eine hinreichend gute und verständliche Erklärung ist? Hierfür werden Systematiken und Bewertungskriterien benötigt, die über Studien, Kontexte und Einstellungen hinweg allgemein anwendbar sind. Darüber hinaus legt die Betonung von Anpassbarkeit und Personenzentrierung nahe, dass für die Evaluierung von XAI vermehrt Studien mit Nutzenden durchgeführt werden sollten. Auch bei der Evaluation gibt es Unterschiede in den relevanten Bewertungskriterien. So können für die Modellerstellenden Metriken für die Modelltreue von Aussagen besonders wichtig sein, während für die Endverbrauchenden eher die Verständlichkeit und Nützlichkeit einer Aussage relevant sind, also Kriterien, die wiederum stark kontextabhängig sind. Erschwerend kommt hinzu, dass es keine begriffliche Klarheit über Konzepte wie Verständlichkeit gibt.

Interdisziplinarität als Lösung für zielgruppenorientiert XAI?

Die interdisziplinäre Zusammenarbeit in Forschung, Entwicklung und Anwendung von XAI ist eng mit dem Thema Anpassungsfähigkeit und menschenzentrierter XAI verbunden. Interdisziplinäre Forschung kann dazu beitragen, die Herausforderungen der Adaptivität zu bewältigen, indem unterschiedliche Perspektiven in den Entwicklungsprozess eingebracht werden. Sie birgt aber auch ihre eigenen Herausforderungen. So sind am Thema XAI die Informatik, die Psychologie, die Sozial- und Geisteswissenschaften und je nach Domäne entsprechende Expertinnen und Experten aus den Anwendungsdisziplinen beteiligt. Dies ist insbesondere bei Studien mit Nutzenden der Fall. Dabei besteht oft a priori kein Konsens über die Definition von Begriffen und die Fachkulturen unterscheiden sich.

7 Gestaltungsoptionen

Vor dem Hintergrund der Potenziale, insbesondere für unterschiedliche Adressaten nachvollziehbarer KI, sowie der dargelegten Herausforderungen werden im Folgenden einige allgemeine Gestaltungsoptionen sowie spezifische Optionen hinsichtlich zielgruppenorientierter KI ausgeführt.[8]

Allgemein

Insgesamt sollten die beiden allgemeinen Ziele von XAI, nämlich der Einsatz von XAI zur Realisierung vertrauenswürdiger KI und der Einsatz von XAI als Ingenieurswerkzeug zur Verbesserung von Modellen, gleichermaßen verfolgt werden. Beide sollten in der öffentlichen Diskussion als Chancen dieser KI-Technologie thematisiert werden. Deutlich sollte dabei jeweils werden, dass XAI nicht nach dem Motto „one-size-fits-all" eingesetzt werden kann, sondern dass die Potenziale dieser Technologie nur dann ausgeschöpft werden können, wenn die Zielgruppen der Erklärung und ihre jeweiligen Ziele berücksichtigt werden. Um dies zu fördern, werden folgende Gestaltungsoptionen vorgeschlagen.

Forschende könnten…

XAI-Methoden für neue Arten von KI entwickeln und etablierte Methoden verbessern: Dies gilt gerade für die rasante Entwicklung bei großen KI-Modellen (z. B. Sprachmodelle, multimodale Modelle etc.), aber auch für verteiltes und interaktives bzw. kollaboratives maschinelles Lernen. Etablierte XAI-Methoden wie Attributions-basierte und Konzept-basierte XAI sollten weiterentwickelt werden. Gerade für große, generative Modelle sollten skalier- und automatisierbare Ansätze mit Blick auf die Inspizier- und Kontrollierbarkeit dieser Modelle (weiter)entwickelt werden und daraus zum Beispiel, in Kooperation mit Unternehmen und anderen Anwendenden, einfach nutzbare Standardwerkzeugkisten nach dem Prinzip „Erklären, um zu revidieren" aufgebaut werden. Bedeutend ist dabei, dass solche Modelle einfach angepasst werden können, ohne dass ein erneutes Modelltraining erforderlich wird. Ferner sind ganzheitliche XAI-Ansätze (Dreyer et al. 2025), die sowohl die Eingabe-, Modell- als auch Ausgabeebene in den Blick nehmen, für große, generative Modelle von Interesse, da sie zahlreiche Perspektiven auf das Modell bieten und der Komplexität dieser Modelle gerecht werden.

Lehrende könnten…

XAI im Sinne eines Engineering-Tools als Teil von AI Engineering stärker in den Curricula von KI- und Data Science-Studiengängen verankern: Dabei sollte nicht nur vermittelt werden, wie verschiedene XAI-Methoden und deren Kombination zur Verbesserung von KI-Modellen jenseits eines Fokus auf Performanz eingesetzt werden können, sondern auch der Entdeckergeist gefördert werden. Modellexploration sollte als kreative Tätigkeit verstanden werden, bei der XAI-Methoden, KI-Wissen, Phantasie und Neugier zusammenkommen, um das Verhalten von Modellen systematisch zu analysieren.

8 Die Optionen sind inspiriert durch Schmid und Wrede (2022), Decker et al. (2023), Saeed und Omlin (2023), Biecek und Samek (2024), Teso et al. (2024), Longo et al. (2024) sowie Samek (2025).

Zielgruppenspezifische XAI

Politikerinnen und Politiker/staatliche Einrichtungen könnten…

Im Aktionsplan Künstliche Intelligenz des Bundesministeriums für Bildung und Forschung (BMBF) ist bereits das Ziel formuliert, deutliche Fortschritte bei der Erklärbarkeit und Vertrauenswürdigkeit von KI zu demonstrieren. Ein Beitrag dazu wird durch die Förderung der Interpretier- und Erklärbarkeit, insbesondere auch in der interdisziplinären Forschung, geleistet (BMBF 2023, S. 7 und 25)[9]; dies ist ein Schritt in die richtige Richtung, der durch weitere Maßnahmen verstärkt werden sollte.

- **Partizipative Projekte fördern:** Diese sollen die betroffenen Stakeholder mit in den Entwicklungs- und Designprozess von KI und XAI einbeziehen.

- **Anreize für interdisziplinäre Forschung schaffen:** Das heißt der Interdisziplinarität Vorrang einräumen, indem Zuschüsse, Preise und andere Anerkennungen für gemeinsame Projekte vergeben werden.

Forschende könnten…

- **Potenziale für mehr Adaptivität ausschöpfen:** Dabei sollte jeweils die Adaptivität und der Fokus auf den Menschen berücksichtigt werden. Insbesondere Sprachmodelle, multimodale Modelle oder auch interaktives Lernen und computergestütztes Argumentieren bieten neue Möglichkeiten für interaktive, personalisierbare Erklärungen in sprachlicher oder visueller Form, die für Nutzende intuitiv sind. Während für KI-Forschende und Modellerstellende die Modellexploration durch effizientes Navigieren durch komplementäre Erklärungen und deren Evaluationsergebnisse im Vordergrund steht, wird für Nutzende eher die Frage im Vordergrund stehen, wie sie (automatisch) eine kompakte, situativ verständliche und hilfreiche Erklärung erhalten oder wie sie selbst die Art und Form der Erklärung an ihre Bedürfnisse anpassen können. Dabei ist es nötig, dass die Nutzenden über Gehalt der Erklärung ausreichend informiert sind, sodass die Erklärungen weder ungerechtfertigtes Misstrauen noch überzogenes Vertrauen in das KI-System generieren.

- **Evaluation von (interaktiver) XAI noch mehr Aufmerksamkeit widmen:** Es sollten Evaluationssystematiken (inkl. Benchmarks, Standards, Tools etc.) entwickelt werden, die über Studien, Kontexte und Einstellungen hinweg allgemein anwendbar sind. Insbesondere die Evaluation aus der Perspektive der Nutzenden wird hier auch einen verstärkten Fokus auf Studien mit Nutzenden erfordern. In beiden Fällen wird die Entwicklung geeigneter Benutzeroberflächen von Bedeutung sein.

9 Siehe hierzu die Bekanntmachungen „Flexible, resiliente und effiziente Machine-Learning-Modelle“ (BMBF 2025a) sowie KI-Erklärbarkeit und Transparenz (BMBF 2025b)

- **Interdisziplinarität in der Forschung stärken:** Die Interdisziplinarität in der Forschung sollte gestärkt werden, da an der XAI-Forschung viele verschiedene Disziplinen beteiligt sind, von der Informatik über die Psychologie bis hin zur Philosophie. Interdisziplinäre Foren sollten gestärkt werden, um den Dialog und das gegenseitige Verständnis zwischen Forschenden mit unterschiedlichem Hintergrund zu fördern. Besonders hilfreich wäre es, wenn aus diesem Dialog eine zentrale Plattform entsteht, um Wissen über XAI interdisziplinär zur Verfügung zu stellen und die Terminologie disziplinübergreifend zu vereinheitlichen. Leitlinien und Standards für kontextsensitive Erklärungen sollten entwickelt werden.

Unternehmen könnten…

- **XAI vermehrt nutzen, um sich von Wettbewerbern abzugrenzen (Decker et al. 2023):** Dabei sollten sie ihre Zielgruppe im Blick behalten und XAI nutzen, um sowohl Vertrauen zu gewinnen als auch um die Performanz der Nutzer-System-Interaktion zu verbessern.

- **XAI vermehrt als Werkzeug einsetzen und weiterentwickeln:** Damit können Hürden für den Austausch zwischen Domänenexpertinnen und -experten, Modellerstellenden, Produktentwickelnden und Verbrauchenden abgebaut werden (Decker et al. 2023).

- **Angaben zur Nachvollziehbarkeit in Model Cards für die von ihnen verwendeten KI-Modelle integrieren (Mitchell et al. 2018):** In Zusammenarbeit mit Forschenden könnte die automatische Erstellung von Model Cards speziell für die Bedürfnisse von Unternehmen und ihrer Zielgruppen weiterentwickelt werden, um so den Aufwand für die Dokumentation zu minimieren.

Literatur

Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W. & Lapuschkin, S. (2023): From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence, 5(9), 1006-1019. https://doi.org/10.1038/s42256-023-00711-8

AI Office (2025): Third Draft of the General-Purpose AI Code. Online unter: EU AI Act: General-Purpose AI Code of Practice · Draft 3

Aljuneidi, S., Heuten, W., Abdenebaoui, L., Wolters, M. K. & Boll, S. (2024): Why the Fine, AI? The Effect of Explanation Level on Citizens' Fairness Perception of AI-based Discretion in Public Administrations. *In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (S. 1-18).

Anders, C. J., Weber, L., Neumann, D., Samek, W., Müller, K.-R. & Lapuschkin, S. (2022): Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. Information Fusion, 77:261-295. https://doi.org/10.1016/j.inffus.2021.07.015

Atzmueller, M., Fürnkranz, J., Kliegr, T. & Schmid, U. (2024): Explainable and interpretable machine learning and data mining. Data Mining and Knowledge Discovery, 38(5), 2571-2595.

Bahlmann, C., Felix, R., Hahn, A. et al. (2024): Vertrauen in KI-basierte Mobilität. Technologische und ethische Aspekte. Whitepaper aus der Plattform Lernende Systeme, München. https://doi.org/10.48669/pls_2024-7

Beckh, K., Müller, S., Jakobs, M., Toborek, V., Tan, H., Fischer, R. (2023): Harnessing prior knowledge for explainable machine learning: An overview. In: *1st IEEE Conference on Secure and Trustworthy Machine Learning SaTML.* https://doi.org/10.1109/SaTML54575.2023.00038

Bekkemoen, Y. (2024): Explainable reinforcement learning (XRL): a systematic literature review and taxonomy. Machine Learning, 113(1), 355-441.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M. & Eckersley, P. (2020): Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 648-657. https://doi.org/10.1145/3351095.3375624

Biecek, P. & Samek, W. (2024): Position: Explain to question not to justify. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR, 235:3996-4006. https://proceedings.mlr.press/v235/biecek24a.html

Boehm, J., Grennan, L., Singla, A. & Smaje, K. (2022): Why digital trust truly matters. Online unter: https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-digital-trust-truly-matters#/

Brath, R., Keim, D.A., Knittel, J., Pan, S., Sommerauer, P. & Strobelt, H. (2023): The Role of Interactive Visualization in Explaining (Large) NLP Models: from Data to Inference. https://doi.org/10.48550/arXiv.2301.04528

Budde, K. et al. (2023): KI für Gesundheitsfachkräfte – Chancen und Herausforderungen von medizinischen und pflegerischen KI-Anwendungen. Whitepaper aus der Plattform Lernende Systeme, München. https://doi.org/10.48669/pls_2023-2

Bundesministerium für Bildung und Forschung (BMBF) (2023): BMBF-Aktionsplan Künstliche Intelligenz. Neue Herausforderungen chancenorientiert angehen. Online unter: https://www.bmbf.de/SharedDocs/Downloads/DE/2023/230823-executive-summary-ki-aktionsplan.pdf?__blob=publicationFile&v=1

Bundesministerium für Bildung und Forschung (BMBF) (2025a): KI-Erklärbarkeit und Transparenz. https://www.bmbf.de/SharedDocs/Bekanntmachungen/DE/2019/04/2392_bekanntmachung.html

Bundesministerium für Bildung und Forschung (BMBF) (2025b): Flexible, resiliente und effiziente Machine-Learning-Modelle. https://www.bmbf.de/SharedDocs/Bekanntmachungen/DE/2023/09/2023-09-07-Bekanntmachung-Machine-Learning-Modelle.html?templateQueryString=flexible+resilienz+und+effizienz++-id_%3A913100

Busse, F. (2023): Potenziale von erklärbarer KI zur Aufklärung von Verbraucher*innen. Online unter: https://www.zvki.de/storage/publications/ZVKI-FachAG1_Thesenpapier_2023.pdf

CAIS (2024): MeMo:KI – Meinungsmonitor Künstliche Intelligenz. Online unter: https://www.cais-research.de/forschung/memoki/

Common Crawl Foundation (2025): Common Crawl. https://commoncrawl.org/

Deck, L., Schomäcker, A., Speith, T., Schöffer, J., Kästner, L., Kühl, N. (2024): Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle. https://arxiv.org/abs/2404.18736

Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R. & Weber, S. H. (2023, July): The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In: *International Conference on Human-Computer Interaction* (S.184-208). Cham: Springer Nature Switzerland.

Deiseroth, B., Deb, M., Weinbach, S., Brack, M., Schramowski, P. & Kersting, K. (2023): AtMan: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation. https://arxiv.org/abs/2301.08110v6

Dreyer, M., Berend, J., Labarta, T., Vielhaben, J., Wiegand, T., Lapuschkin, S. & Samek, W. (2025): Mechanistic understanding and validation of large AI models with SemanticLens. http://arxiv.org/abs/2501.05398

European Data Protection Supervisor (EDPS) (2023): EDPS TechDispatch on Explainable Artificial Intelligence. Online unter: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

Finzel, B., Tafler, D. E., Scheele, S. & Schmid, U. (2021): Explanation as a process: User-centric construction of multi-level and multi-modal explanations. In: *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44* (pp. 80-94). Springer International Publishing.

Fischer, R., Jakobs, M., Mücke, S., Morik, K. (2022): A unified framework for assessing energy efficiency of machine learning. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases,* pp 39–54. ECML PKDD 2022. https://doi.org/10.1007/978-3-031-23618-1_3

Fischer, R., Liebig, T. & Morik, K. (2024): Towards more sustainable and trustworthy reporting in machine learning. Data Mining and Knowledge Discovery (2024)38, 1909–1928. https://doi.org/10.1007/s10618-024-01020-3

Fischer, R., Wischnewski, M., van der Staay, A. & Poitz K. (2024): Practical insights for trustworthy AI development via model labeling (working title), 2024. https://lamarr.cs.tu-dortmund.de/ml-label-interviews/

Göbel, K., Niessen, C., Seufert, S. & Schmid, U. (2022): Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. Frontiers in Artificial Intelligence, 5, 919534.

Gramelt, D., Höfer, T. & Schmid, U. (2024): Interactive Explainable Anomaly Detection for Industrial Settings. DOI: https://doi.org/10.48550/arXiv.2410.12817

Gyevnar, B., Ferguson, N. & Schafer, B. (2023): Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act? European Conference on Artificial Intelligence. https://doi.org/10.48550/arXiv.2302.10766

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., ... & Höhne, M. M. C. (2023): Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research, 24*(34), 1-11.

Heidrich, L., Slany, E., Scheele, S. & Schmid, U. (2023): FairCaipi: A Combination of Explanatory Interactive and Fair Machine Learning for Human and Machine Bias Reduction. Machine Learning and Knowledge Extraction, 5(4), 1519-1538.

Houdeau, D. (2022): Wie wir KI-Systeme vor Cyberangriffen schützen. https://www.plattform-lernende-systeme.de/reden-und-beitraege-newsreader/wie-wir-ki-systeme-vor-cyberangriffen-schuetzen.html

Huchler, N. et al. (Hrsg.) (2020): Kriterien für die menschengerechte Gestaltung der Mensch-Maschine-Interaktion bei Lernenden Systemen. Whitepaper aus der Plattform Lernende Systeme, München.

Jacovi, A. (2023): Trends in explainable AI (XAI) literature. https://arxiv.org/abs/2301.05433

Kluge Corrêa, N. et al. (2023): Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns, 4*(10). https://doi.org/10.1016/j.patter.2023.100857

Koebler, A., Decker, T., Lebacher, M., Thon, I., Tresp, V. & Buettner, F. (2023): Towards explanatory model monitoring. In *XAI in Action: Past, Present, and Future Applications.*

Kraus, T., Ganschow, L., Eisenträger, M., Wischmann, S. (2021): Erklärbare KI: Anforderungen, Anwendungsfälle und Lösungen. Herausgeber: Institut für Innovation und Technik. Online unter: https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.html

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & 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. *Artificial Intelligence, 296,* 103473.

Lapuschkin, S., Wäldchen, S., Binder, A. et al. (2019): Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* 10, 1096. https://doi.org/10.1038/s41467-019-08987-4

Levy, S. (2024, Mai 21): AI Is a Black Box. Anthropic Figured Out a Way to Look Inside. Wired. Online unter: https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features/

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., ... & Tegmark, M. (2024): Kan: Kolmogorov-arnold networks. https://arxiv.org/abs/2404.19756

Longo, L. et al. (2024): Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion, 106,* 102301. https://doi.org/10.1016/j.inffus.2024.102301

McKinsey (2022): Why businesses need explainable AI – and how to deliver it. Online unter: https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. & Gebru, T. (2018): Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency.*

Morik, K., Kotthaus, H., Fischer, R., Mücke S., Jakobs, M. et al. (2022): Yes we care! – Certification for machine learning methods through the care label framework. Front. Artif. Intell. 5. https://doi.org/10.3389/frai.2022.975029

Mullenbach, J. et al. (2018): Explainable Prediction of Medical Codes from Clinical Text. (A. f. Linguistics, Hrsg.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), S. 1101-1111. http://dx.doi.org/10.18653/v1/N18-1100

Müller-Quade, J. et al. (2020): Sichere KI-Systeme für die Medizin. Whitepaper aus der Plattform Lernende Systeme, München. Online unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_6_Whitepaper_07042020.pdf

Nandis, S. (2024): Novel Architecture Makes Neural Networks More Understandable. Online unter: https://www.quantamagazine.org/novel-architecture-makes-neural-networks-more-understandable-20240911/

Pahde, F., Dreyer, M., Samek, W. & Lapuschkin, S. (2023, October): Reveal to revise: An explainable AI life cycle for iterative bias correction of deep models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 596-606). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43895-0_56

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J. & Gomez, E. (2023): The role of explainable AI in the context of the AI Act. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1139–1150. https://doi.org/10.1145/3593013.3594069

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., ... & Wrobel, S. (2021): Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog). https://www.iais.fraunhofer.de/content/dam/iais/publikationen/studien-und-whitepaper/2021/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf

Rostalski, F., Janal, R. et al. (2024): Künstliche Intelligenz und Recht. Auf dem Weg zum Robo-Richter? Whitepaper aus der Plattform Lernende Systeme, München. https://doi.org/10.48669/pls_2024-6

Saeed, W. & Omlin, C. (2023): Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems 263, C (Mar 2023). https://doi.org/10.1016/j.knosys.2023.110273

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. und Müller, K.-R. (2019): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, Springer, 11700:1-439. https://doi.org/10.1007/978-3-030-28954-6

Samek, W. (2025): Explaining and Interpreting Generative AI. In: *The Oxford Handbook of the Foundations and Regulation of Generative AI.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198940272.001.0001

Schmid, U. & Finzel, B. (2020): Mutual explanations for cooperative decision making in medicine. KI-Künstliche Intelligenz, 34(2), 227-233. https://doi.org/10.1007/s13218-020-00633-2

Schmid, U., Wrede, B. (2022): What is Missing in XAI So Far? Künstliche Intelligenz 36, 303–315. https://doi.org/10.1007/s13218-022-00786-2

Schmid, U. (2024): Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable. In: Werthner, H. et al.: *Introduction to Digital Humanism.* Springer, Cham. https://doi.org/10.1007/978-3-031-45304-5_10

Schneider, J. (2024): Explainable generative AI (GenXAI): A survey, conceptualization, and research agenda. Artificial Intelligence Review, 57(11), 289.

Schramm, S., Wehner, C. & Schmid, U. (2023): Comprehensible Artificial Intelligence on Knowledge Graphs: A survey. Journal of Web Semantics,79, 100806.

Schwalbe, G. & Finzel, B. (2024): A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38(5), 3043-3101.

Sperrle, F., Jeitler, A. V., Bernard, J., Keim, D. A., El-Assady, M. (2021): Co-Adaptive Visual Data Analysis and Guidance Processes, Computers & Graphics, 100, 93-105.

Spinner, T., Schlegel, U., Schäfer, H., El-Assady, M. (2019): explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning IEEE Transactions on Visualization and Computer Graphics. https://ieeexplore.ieee.org/document/8807299

Stowasser, S. & Suchy, O. et al. (Hrsg.) (2020): Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management. Whitepaper aus der Plattform Lernende Systeme, München.

Teso, S. & Kersting, K. (2019, January): Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 239-245).

Teso, S., Alkan, Ö., Stammer, W. & Daly, E. (2023): Leveraging explanations in interactive machine learning: An overview. Frontiers in Artificial Intelligence, 6, 1066049.

Vakilzadeh Hatefi, S. M., Dreyer, M., Achtibat, R., Wifegand, T., Samek, W. & Lapuschkin, S. (2024): Pruning By Explaining Revisited: Optimizing Attribution Methods to Prune CNNs and Transformers. Computer Vision – ECCV 2024 Workshops. Lecture Notes in Computer Science, vol 15643. Springer, Cham (im Erscheinen).

van Aken, B. et al. (2022): This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis. Prediction from Clinical Text. AACL/IJCNL 2022.

Weber, L., Lapuschkin, S., Binder, A. & Samek, W. (2023): Beyond explaining: Opportunities and challenges of XAI-based model improvement. Information Fusion, 92, 154-176. https://doi.org/10.1016/j.inffus.2022.11.013

Yeom, S. K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K. R. & Samek, W. (2021): Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition, 115,* 107899. https://doi.org/10.1016/j.patcog.2021.107899

Über dieses Whitepaper

Die Autorinnen und Autoren sind Mitglieder der Arbeitsgruppe *Technologische Wegbereiter und Data Science der Plattform Lernende Systeme.* Als eine von insgesamt sieben Arbeitsgruppen thematisiert sie Fragen zu KI-Forschungsfeldern und Potenzialen von KI-Technologien sowie zu Ausbildung von KI-Talenten und Transfer in die Anwendung.

Autorinnen und Autoren
Prof. Dr. Wojciech Samek, Fraunhofer Heinrich-Hertz-Institut (HHI)
Prof. Dr. Ute Schmid, Universität Bamberg

Dr. Johannes Hoffart, SAP
Prof. Dr. Daniel Keim, Universität Konstanz
Prof. Dr. Gitta Kutyniok, LMU München
Philipp Schlunder, daibe.io

Redaktion
Dr. Maximilian Hösl, Plattform Lernende Systeme
Christine Wirth, Plattform Lernende Systeme

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
Juni 2025

Bildnachweis
janiecbros/iStock (Titel), freepik/eigene Darstellung

Empfohlene Zitierweise
Samek, W., Schmid, U. et al. (2025):
Nachvollziehbare KI: Erklären, für wen, was und wofür.
DOI: https://doi.org/10.48669/pls_2025-2

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Dr. Thomas Schmidt (Leiter der Geschäftsstelle):
info@plattform-lernende-systeme.de

Über die Plattform Lernende Systeme

Die Plattform Lernende Systeme ist ein Netzwerk von Expertinnen und Experten zum Thema Künstliche Intelligenz (KI). Sie bündelt vorhandenes Fachwissen und fördert als unabhängiger Makler den interdisziplinären Austausch und gesellschaftlichen Dialog. Die knapp 200 Mitglieder aus Wissenschaft, Wirtschaft und Gesellschaft entwickeln in Arbeitsgruppen Positionen zu Chancen und Herausforderungen von KI und benennen Handlungsoptionen für ihre verantwortliche Gestaltung. Damit unterstützen sie den Weg Deutschlands zu einem führenden Anbieter von vertrauenswürdiger KI sowie den Einsatz der Schlüsseltechnologie in Wirtschaft und Gesellschaft. Die Plattform Lernende Systeme wurde 2017 vom Bundesforschungsministerium auf Anregung von acatech – Deutsche Akademie der Technikwissenschaften gegründet und wird von einem Lenkungskreis gesteuert. Die Leitung der Plattform liegt bei Dorothee Bär (Bundesministerin für Forschung, Technologie und Raumfahrt) und Jan Wörner (Präsident acatech).