Lernende Systeme
DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Nachvollziehbare KI

Whitepaper

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

Kurzfassung

Künstliche Intelligenz (KI) findet in vielen gesellschaftlichen Bereichen Anwendung, beispielsweise in der medizinischen Diagnostik, bei Finanzentscheidungen oder in der automatisierten Qualitätskontrolle in der Produktion. Mit ihrer zunehmenden Verbreitung steigt die Anforderung, die Funktionsweise und Entscheidungsprozesse von KI-Systemen transparent und nachvollziehbar zu gestalten. Denn wie und warum ChatGPT und andere KI-basierte Systeme zu ihren Ergebnissen kommen, also warum eine Anweisung an einen Chatbot genau zu der Wortfolge führt, die er ausgibt, bleibt häufig undurchsichtig. Für viele Anwendungsgebiete, insbesondere in sicherheitskritischen oder ethisch sensiblen Bereichen, wie der Medizin oder der Finanzwirtschaft, sind Transparenz und Nachvollziehbarkeit jedoch entscheidend, um die erzielten Ergebnisse einordnen und kritisch hinterfragen zu können.

Ganz allgemein sind KI-Systeme mit hoher Komplexität häufig sogenannte Black-Boxes. Aufgrund der Vielschichtigkeit der Modelle ist es oft nicht mehr möglich, die interne Funktionsweise des Modells oder die genauen Abläufe zwischen Eingabe und Ausgabe nachzuvollziehen, insbesondere im Bereich des maschinellen Lernens. Diese fehlende Transparenz kann nicht nur das Vertrauen der Nutzenden untergraben, sondern auch die Weiterentwicklung und Verbesserung von KI-Systemen erschweren. Genau hier setzt das Konzept der „Nachvollziehbaren KI“ an.

Ausgangssituation

In den letzten Jahren hat die Forschung im Bereich der erklärbaren KI (engl. explainable AI, kurz: XAI) erhebliche Fortschritte gemacht und sich zu einem vielversprechenden Forschungsfeld entwickelt. XAI zielt darauf ab, die Sicherheit, Qualität und Vertrauenswürdigkeit von KI-Systemen zu verbessern, indem sie beispielsweise Verzerrungen aufdeckt oder konkrete Modellverbesserungen ermöglicht. Damit trägt XAI einerseits dazu bei, Vertrauen sowie Werte wie Transparenz und Verlässlichkeit in KI-Systeme zu fördern. Werte wie sie etwa auch in regulatorischen Vorgaben wie der Datenschutzgrundverordnung (DSGVO) oder der KI-Verordnung (AI Act) angestrebt werden sowie auch in Normen der International Organization for Standardization (ISO) oder in den Kriterien im Prüfkatalog des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme (IAIS). Andererseits bietet der Einsatz von XAI Unternehmen die Möglichkeit, ihre Produktivität und damit auch ihre Gewinne zu steigern, indem sie beispielsweise Fehler in KI-Anwendungen zeitnaher erkennen, die Anwendungen optimieren oder das Vertrauen der Verbraucherinnen und Verbraucher stärken.

Erklärbarkeit sollte dabei stets in Benutzendenschnittstellen und konkreten Kontexten eingebettet sein, sodass sichergestellt wird, dass KI-Systeme menschliche Bedürfnisse auf verantwortungsvolle Weise bedienen können. Denn nicht alle Nutzenden benötigen dieselbe Art von Erklärung: Während Entwickelnde beispielsweise technische Informationen benötigen, um KI-Systeme zu verbessern, interessieren sich Verbrauchende eher für die Gründe einer konkreten Entscheidung, etwa bei der Kreditvergabe. Nachvollziehbare KI sollte daher einen menschengerechten Ansatz verfolgen, bei dem die Verständlichkeit auf die jeweiligen Bedürfnisse, Kenntnisse und Verantwortlichkeiten der jeweiligen Zielgruppe zugeschnitten ist – von der KI-Spezialistin bis hin zum einfachen Anwender.

Begriffsverständnis im XAI-Kontext

Was ist aber unter Transparenz zu verstehen und wie grenzt sich dieser Begriff von Erklärbarkeit ab? Eine definitorische Abgrenzung zur Einordnung zentraler Begrifflichkeiten um XAI erweist sich hier notwendig, da bisher noch keine allgemein anerkannte einheitliche Taxonomie existiert: **Transparenz**, verstanden in einem eher engen technischen Sinn, bezeichnet eine Systemeigenschaft, die das innere Funktionieren des Systems einschließlich seiner Funktionsweise, internen Schlussfolgerungen und Verarbeitungsschritte nachvollziehbar macht. Darüber hinaus erklärt sie die Systemeingaben und -ausgaben, wie etwa Systementscheidungen, in einer angemessenen Weise. Bei **Erklärbarkeit** geht es darum, den Nutzenden Begründungen für das Modellverhalten eines KI-Modells bereitzustellen, damit sie das Ergebnis (z. B. Vorhersagen oder Klassifikationen) nachvollziehen können. Da diese XAI-Methode oft außerhalb des Modells liegt, erklärt sie keine modellinternen Prozesse. **Interpretierbarkeit** bezieht sich dagegen auf das Modell selbst (White-Box-Modell). Denn es folgt standardmäßig einem für Menschen nachvollziehbaren Entscheidungsprozess und liefert somit selbst eine Erklärung für die Klassifikation oder Vorhersage. **Nachvollziehbarkeit**, als übergeordnetes Ziel, wird durch Interpretierbarkeit und Erklärbarkeit erreicht.

XAI entlang der Frage Wem, soll was wozu und wie erklärt werden

Wie lässt sich die „Black-Box“ mithilfe von XAI öffnen? Der Schlüssel liegt in einem reflektierten Einsatz von XAI entlang der Fragestellung „Wem soll was wozu und wie erklärt werden?“. XAI ist somit kein Selbstzweck und bietet keine one-size-fits-all-Lösungen. Vielmehr gilt es, XAI mit einer realistischen und angemessenen Erwartungshaltung zu entwickeln und anzuwenden, da sie letztlich nur Aspekte des Modellverhaltens erklären kann. Sie vermag keine vollständige und umfassende Erklärung zu geben, was in vielen Fällen auch nicht unbedingt notwendig ist.

Zielgruppe – Wem erklären? Die Frage nach dem Wem zielt auf die jeweilige Zielgruppe einer Erklärung ab. Erklärungen müssen an diese angepasst werden, da Erwartungen und Anforderungen der unterschiedlichen Zielgruppen stark variieren. Faktoren wie Vorwissen der Betroffenen im Bereich KI, Zeitdruck, Motivation und Interessen spielen dabei eine Rolle. So können für die Modellerstellenden Metriken für die Modelltreue von Erklärungen bedeutend sein, während für die Endverbrauchenden eher die Verständlichkeit und Nützlichkeit einer Erklärung relevant sind, also Kriterien, die wiederum stark kontextabhängig sind.

Grundsätzlich lassen sich drei Zielgruppen unterscheiden. Die erste Zielgruppe, Modellentwickelnde und KI-Forschende, benötigt technisches Wissen und Details in den Erklärungen, um KI-Modelle mit XAI-Methoden zu verbessern und zu überprüfen. Die zweite Zielgruppe, KI-Laien und Endverbrauchende, legt Wert auf Vertrauenswürdigkeit und erwartet einfache und zugängliche Erklärungen. Die dritte Zielgruppe, etwa Produktentwickelnde, Zertifizierende und Domänenforschende, erwartet, dass XAI-Methoden sie in ihrer spezifischen Tätigkeit unterstützen. Für diese ist Vertrauenswürdigkeit und Qualitätsverbesserung von KI wichtig.

Erklärungsgegenstand – Was erklären? Mit der Anpassung der Erklärung an die entsprechende Zielgruppe stellt sich die Frage, was der konkrete Gegenstand der Erklärung ist. Dieser kann im Prozess des maschinellen Lernens die Eingabe-, Modell- und Ausgabe-Ebene sein. XAI-Methoden liefern auf diesen drei Ebenen wichtige Erkenntnisse, sei es zur Datenqualität, zum Modellverständnis oder zur Erklärung konkreter Klassifikationen und Vorhersagen: Auf der Eingabeebene können so fehlerhafte und verzerrte Datensätze aufgedeckt werden. Auf der Modellebene beispielsweise beim Deep Learning kann analysiert werden, welche Konzepte durch einzelne Neuronen eines künstlichen neuronalen Netzwerks repräsentiert werden. Auf der Ausgabeebene kann ermittelt werden, wie eine konkrete Modellausgabe zustande kam, beispielsweise die Antwort eines Chatbots auf eine Anfrage oder ein Vorschlag für eine medizinische Diagnose. Die drei Ebenen können entweder als separate Elemente einer Erklärung dienen oder mithilfe neuerer Methoden wie SemanticLens als ganzheitlicher Ansatz verwendet werden, um die Rolle und Funktionsweise aller Komponenten von KI systematisch auszuwerten und zu validieren.

Ziele der Erklärung – Wozu erklären? Die Fragestellung klärt, was die konkreten Ziele sind. Diese sollten realistisch sein, um grundsätzlich abzuklären, wann erklärbare KI sinnvoll ist und wann nicht. Mit dem Einsatz von XAI-Methoden und -Verfahren lassen sich vornehmlich zwei allgemeinere Ziele von XAI identifizieren, die sich teilweise überschneiden: Einerseits die Nutzung von XAI-Methoden, um den Dimensionen vertrauenswürdiger KI gerecht zu werden (z. B. Kontrolle, Fairness, Verlässlichkeit etc.). Damit wird das allgemeine Ziel verfolgt, Vertrauen in KI zu stärken. Andererseits die Nutzung von XAI-Methoden als Ingenieurswerkzeug, um die Qualität von KI-Modellen zu verbessern (z. B. Datenbasis und Modell verbessern, Domänenanpassung etc.).

Auch wenn die verschiedenen Zielgruppen von XAI jeweils eigene Gründe haben, XAI einzusetzen – beispielsweise, um etwas über die Daten und eine konkrete Problemstellung zu lernen, neue Antibiotika zu entwickeln oder eine monetär oder existentiell wichtige Entscheidung nachvollziehbar zu machen –, werden die beiden Ziele „XAI für vertrauenswürdige KI“ und „XAI als Ingenieurswerkzeug“ in weiten Teilen ein Mittel zum Zweck der Realisierung zielgruppenspezifischer Ziele darstellen.

Umsetzung der Erklärung – Wie erklären? Das „Wie“ setzt sich damit auseinander, 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). Damit lässt sich die Wirksamkeit eines KI-Modells im praktischen Einsatz besser verstehen und optimieren.

Die Form der Erklärung muss zielgruppengerecht, verständlich und kontextbezogen sein – ob visuell, verbal oder multimodal: beispielsweise per visueller Hervorhebungen bei zeitkritischen Entscheidungen oder als detaillierte verbale Erklärungen bei komplexen Sachverhalten. Die **Methode** einer Erklärung legt dar, wie eine Erklärung technisch erzeugt wird: zum Beispiel Relevanz-, Konzept- und Beispiel-basierte Methoden (Local Interpretable Model-Agnostic Explanations (LIME), Layer-wise Relevance Propagation (LRP), Shapley Additive Explanation (SHAP) etc.). Neuere Ansätze wie interaktive XAI, binden Nutzende aktiv in den maschinellen Lern- und Erklärungsprozess ein und generative KI stellt XAI unter anderem durch die schiere Modell- und Datensatzgröße vor neue Herausforderungen. Zu den neueren Ansätzen zählt auch mechanistische Interpretierbarkeit, die versucht, die inneren Strukturen neuronaler Netze gezielt zu analysieren und zu deuten.

Da Erklärungen nicht zwangsläufig korrekt sind, wird auch eine **Evaluation der Erklärung** mittels spezieller Metriken nötig, die wiederum eine systematische Bewertung und Verbesserung von XAI-Erklärungen ermöglichen. Hierzu werden unter anderem Modelltreue und Robustheit gemessen: Einerseits die Übereinstimmung der Erklärung mit dem tatsächlichen Verhalten und andererseits die Stabilität der Erklärungen bei leichten Störungen der Modelleingaben.

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.

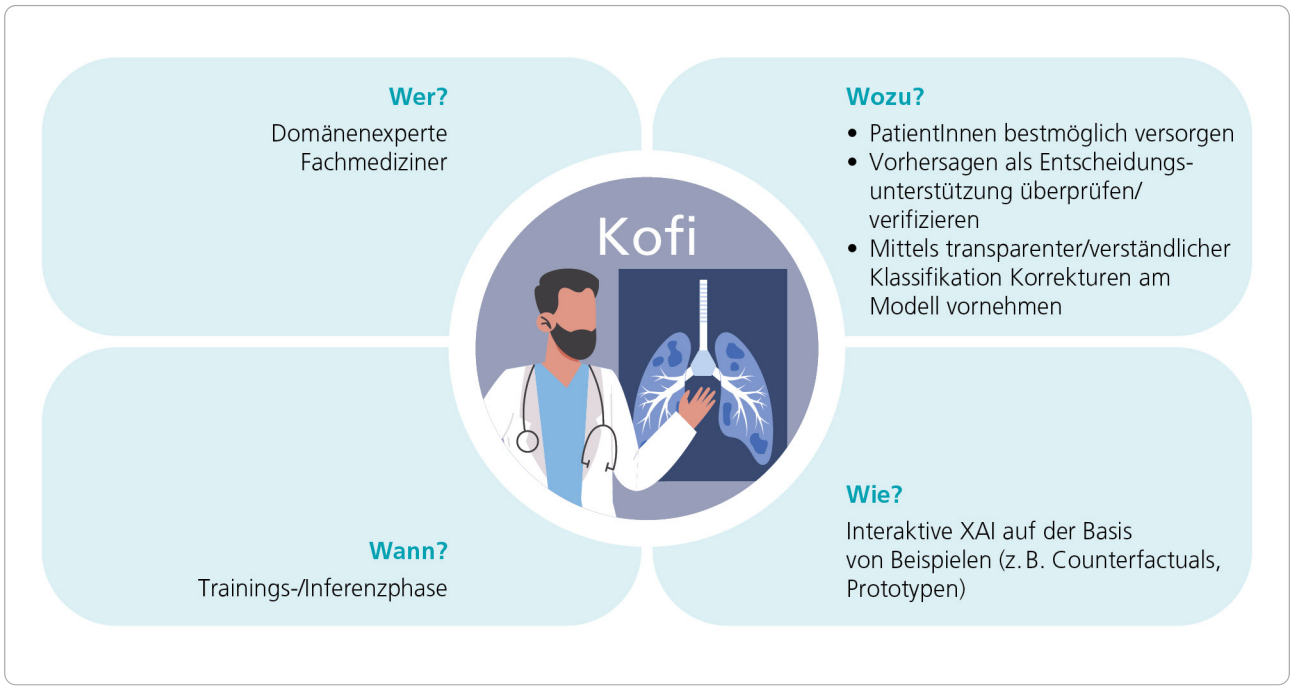
Konzept-basiert

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).

Wie eine XAI-gestützte Erklärungsstrategie aus unterschiedlichen Perspektiven entlang der W-Frage aussehen kann, wird anhand von sieben verschiedenen Personae vorgestellt: KI-Forscher, Domänenexpertin und Domänenexperte (Chemiker/Arzneimittelforschung, Fachmediziner), KI-Modellarchitektin, Software-Produktmanager und Endverbraucher (Prozessführung): Dabei werden übergeordnete Ziele und konkrete Erklärbarkeitsanforderungen aus Sicht der jeweiligen Persona identifiziert und entsprechende XAI-Erklärungswege beschrieben.

Beispiel der Persona „Kofi – Domänenexperte“



Ausblick

Erklärbare KI ist ein vielversprechendes Feld, das die Qualität und Vertrauenswürdigkeit von KI-Systemen erheblich verbessern kann. Aufgrund ihrer gesellschaftlichen und wirtschaftlichen Relevanz ist die Weiterentwicklung von XAI-Methoden und ihre praktische Anwendung von entscheidender Bedeutung, um zukünftige Herausforderungen in der KI-Entwicklung erfolgreich zu bewältigen. Daher sollten die beiden allgemeinen Ziele von XAI – der Einsatz von XAI zur Realisierung vertrauenswürdiger KI und der Einsatz von XAI als Ingenieurswerkzeug zur Verbesserung von Modellen – gleichermaßen verfolgt werden.

Um das Potenzial von XAI noch besser auszuschöpfen, sollte die Anpassungsfähigkeit und Verknüpfung an Ziele, Zielgruppen und Domänen verfolgt und verstärkt berücksichtigt werden. Dabei sollte stets der Mensch im Mittelpunkt stehen, um diesem Handlungsoptionen zu ermöglichen. Hierzu sind Evaluationssystematiken zu entwickeln, die über Studien, Kontexte und Einstellungen hinweg allgemein anwendbar sind. Seitens der Forschung sollten etablierte Methoden verbessert und XAI-Methoden für neue Arten von

KI (weiter-)entwickelt werden, insbesondere in Hinblick auf holistische XAI-Ansätze, die Daten-, Modell- und Ausgabeebene einbeziehen sowie Ansätze, die auf die Inspizier- und Kontrollierbarkeit großer Modelle ausgerichtet sind. 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 und Organisationen sollten die Implementierung von XAI als integralen Bestandteil ihrer KI-Strategie betrachten und vermehrt auf XAI setzen, um beispielsweise etwa Produkte zu verbessern, interne Kommunikationshürden abzubauen oder sich von Wettbewerbern abzugrenzen.

Impressum
Herausgeber: Lernende Systeme – Die Plattform für Künstliche Intelligenz | Geschäftsstelle | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Folgen Sie uns auf LinkedIn: de.linkedin.com/company/plattform-lernende-systeme | Mastodon: social.bund.de/@LernendeSysteme | Stand: Juni 2025 | Bildnachweis: janiecbros/iStock/Titel

Diese Kurzfassung entstand auf Grundlage des Whitepapers *Nachvollziehbare KI. Erklären, für wen, was und wofür*, München, 2025. Es wurde erstellt von Mitgliedern der Arbeitsgruppe Technologische Wegbereiter und Data Science. Die Originalfassung der Publikation ist online verfügbar unter: *https://doi.org/10.48669/pls_2025-2*

Gefördert vom

Bundesministerium
für Forschung, Technologie
und Raumfahrt

acatech
DEUTSCHE AKADEMIE DER
TECHNIKWISSENSCHAFTEN