Fraunhofer

Erklärbare Künstliche Intelligenz

Potenziale, Herausforderungen und Ansätze zur Zertifizierung von KI-Systemen

Benjamin Fresz | Danilo Brajovic | Andreas Aichele | Irene Edosomwan | Marco Huber
Vincent Philipp Göbels | Safa Omri | Janika Kutz | Jens Neuhüttler

Inhalt

Executive Summary

Mit der zunehmenden Marktreife von KI-Produkten stehen Entwicklerinnen und Entwickler sowie Unternehmen vor der Herausforderung, die Sicherheit dieser Produkte zu verifizieren und für eine eventuelle Zertifizierung bzw. Marktzulassung nachzuweisen. Diese Aufgabe wird durch die »Black-Box«-Natur von KI erschwert, da herkömmliche Prüfprozesse, die auf statischen Ausgaben und der Logik regelbasierter Algorithmen beruhen, bei selbstlernenden KI-Systemen nicht anwendbar sind. Obwohl klassische Softwaretests möglicherweise nie vollständig auf KI übertragbar sein werden, hat die Forschung in den letzten Jahren erhebliche Fortschritte im Bereich der erklärbaren Künstlichen Intelligenz (engl. eXplainable AI, kurz: XAI) gemacht. XAI zielt darauf ab, die inhärenten Logiken hinter Entscheidungen und Datenverknüpfungen in KI-Modellen nachvollziehbar zu machen.

Das vorliegende Whitepaper untersucht die Potenziale von XAI für die Entwicklung, Absicherung und Zertifizierung von KI-Systemen. Hierfür wurde relevante Literatur analysiert und Erkenntnisse aus Interviews mit Expertinnen und Experten (nachfolgend: Experteninterviews) aus Anwendung, Forschung und Zertifizierung wurden zusammengetragen. Die zentralen Erkenntnisse des Whitepapers sind:

Volle Absicherung durch XAI unrealistisch	Die umfassende Absicherung von KI mittels XAI, insbesondere durch vollständige globale Erklärungen von Modellen, wäre ideal, wird aber von Expertinnen und Experten als unrealistisch angesehen.
Mehrwert von XAI	XAI kann einen erheblichen Mehrwert für die Entwicklung sicherer KI-Systeme bieten, da XAI-Methoden tiefere Einblicke in die zugrunde liegende Datenbasis und mögliche Verzerrungen (Bias) in KI-Modellen ermöglichen.
Einsatz in Zertifizierungsprozessen	XAI kann Verzerrungen aufdecken und somit z. B. für mehr Fairness in Modellen sorgen. Allerdings gibt es keine Garantie für die vollständige Abwesenheit solcher Verzerrungen oder fehlerhaften KI-Entscheidungen.

User-fokussierte Methoden

Es besteht ein wesentlicher Bedarf an stärker nutzerfokussierten XAI-Methoden, da die aktuellen Erklärungen für Laien und Domänenexpertinnen und -experten oft schwer verständlich sind.

Praxisorientierte Forschung

Eine stärkere Industrie- und Praxisorientierung könnte helfen, XAI-Methoden für bisher untererforschte Bereiche wie Zeitreihendaten oder Large Language Models (LLMs) zu entwickeln.

Erweiterung der Absicherung

Es gibt erheblichen Spielraum für Methoden außerhalb der klassischen Erklärbarkeit zur Absicherung und Zertifizierung von KI. Dies umfasst die Unsicherheitsquantifizierung von KI-Modellen, die formale Verifikation bestimmter Eigenschaften, die KI-Unterstützung bei der Prüfung von KI sowie die Prüfung gegenüber physikalischen Gesetzen.

Zusammenfassend lässt sich feststellen, dass die Integration von XAI in die Absicherung und Zertifizierung von KI-Systemen vielversprechende Möglichkeiten zur Erhöhung der Transparenz und Vertrauenswürdigkeit dieser Technologien bietet. Trotz bestehender Herausforderungen zeigt das vorliegende Whitepaper, dass XAI einen wertvollen Beitrag zur sicheren und nachvollziehbaren Nutzung von KI leisten kann. Die kontinuierliche Weiterentwicklung von XAI-Methoden und deren praktische Anwendung sind entscheidend, um zukünftige Herausforderungen in der KI-Entwicklung und KI-Zertifizierung erfolgreich zu meistern.

Durch eine enge Zusammenarbeit zwischen Forschungseinrichtungen, Industrie und Regulierungsbehörden können harmonisierte Normen und Standards geschaffen werden, die den Einsatz von XAI in der Zertifizierung von KI-Systemen ermöglichen. Letztlich verdeutlicht dieses Whitepaper, dass erklärbare KI ein vielversprechendes Feld ist, das die Sicherheit und Vertrauenswürdigkeit von KI-Systemen signifikant verbessern kann.

1. Einleitung

In der sich rasant entwickelnden, zunehmend digitalen Welt stellt sich die Herausforderung, innovative Technologien vertrauensvoll und sicher zu entwickeln und zu nutzen. Künstliche Intelligenz (KI) ist dabei der größte Innovationstreiber, da sie das Potenzial hat, nahezu alle Aspekte des persönlichen Lebens zu verändern – von der Arbeitsweise bis hin zu alltäglichen Interaktionen. Durch die Verbreitung und den Einsatz in sicherheitskritischen Anwendungsfällen stehen die Qualität und Sicherheit der eingesetzten KI-Lösungen zunehmend im Mittelpunkt. Hierbei sind Absicherung und Zertifizierung dieser Technologien entscheidend, um das Vertrauen in die Integrität und Leistungsfähigkeit der Produkte und Services zu gewährleisten.

Dabei bezieht sich Absicherung grundsätzlich auf alle Maßnahmen, die getroffen werden, um ein Produkt – KI oder anderweitig – sicherer zu machen, während Zertifizierung eine Bestätigung über das Einhalten bestimmter Standards beschreibt. Diese Standards können sich auf Qualität, Sicherheit, Effizienz oder andere spezifische Merkmale beziehen, die in verschiedenen Branchen und Bereichen von Bedeutung sind. Die Zertifizierung erfolgt in der Regel nach einer gründlichen Prüfung und Bewertung durch eine unabhängige und akkreditierte Zertifizierungsstelle. Für KI ergibt sich dabei das Problem, dass entsprechende Standards meist noch in der Entstehung sind. Grund hierfür ist, dass die bestehenden Zertifizierungsprozesse für konventionelle Produkte an die Spezifika von KI-Systemen angepasst werden müssen. Existierende Normen und Standards zur Absicherung von KI bieten oft nicht ausreichend konkrete Handlungsanweisungen für die Entwicklung von KI-Systemen. An dieser Stelle ist insbesondere der EU AI Act zu beachten, der die Zulassung von KI-Systemen unter Auflagen in der EU reguliert und 2024 beschlossen wurde. Technische Details zum AI Act werden in Zukunft in den zugehörigen harmonisierten Normen spezifiziert, allerdings ist davon auszugehen, dass auch diese Interpretationsspielraum bieten werden.

Eine mögliche Lösung für die Absicherung und Zertifizierung von KI könnten Methoden aus dem Bereich der erklärbaren KI (engl. eXplainable AI, kurz: XAI) darstellen, mit deren Hilfe sich Einblicke in KI-Systeme gewinnen lassen. Allerdings ist XAI ein recht junges Forschungsfeld, somit sind die tatsächlichen Potenziale und momentanen Probleme im Bereich XAI noch nicht vollständig erschlossen. Dieses Whitepaper hat daher zum Ziel, durch Literaturanalyse und Experteninterviews die Potenziale von XAI in Bezug auf Absicherung und Zertifizierung zu ergründen.

Im Folgenden wird ein Überblick gegeben, wie sich Entwicklung, Absicherung und Zertifizierung bei KI-Systemen und bei herkömmlichen Produkten voneinander unterscheiden und welche Potenziale hier XAI bietet. Daraufhin werden anhand relevanter Literatur Eigenschaften und Probleme von XAI diskutiert, die schließlich mithilfe von Experteninterviews auf den Bereich der Absicherung und Zertifizierung heruntergebrochen werden. Zum Schluss wird auf Trends im Bereich XAI hingewiesen, die laut den Interviewpartnerinnen und Interviewpartnern bisherige Probleme des Forschungsbereichs und der KI-Zertifizierung in Zukunft lösen könnten.

2. Absicherung und Zertifizierung von KI-Systemen

Die Entwicklung von Produkten unterliegt zahlreichen rechtlichen und normenbedingten Vorgaben. Klassische Technologieprodukte werden dabei nach Normen geprüft. Diese beschreiben, wie Produkte im jeweiligen Geltungsbereich abgesichert sein sollten. In der Zertifizierung wird anschließend geprüft, ob die Dokumentation der Absicherung (»Evidenz«) aus dem Entwicklungsprozess darauf schließen lässt, dass alle Vorgaben der relevanten Norm erfüllt sind. Derartige Normen entstehen wiederum aus gesetzlichen Regulierungen und Anforderungen an die Produktsicherheit.

Zum Beispiel werden die Anforderungen an die Funktionale Sicherheit von elektrischen/elektronischen Komponenten im Automotive-Umfeld in der ISO 26262 beschrieben. Diese beinhaltet ein Vorgehensmodell sowie anzuwendende Methoden und fordert diverse Aktivitäten und Arbeitsprodukte. Der prinzipielle Ablauf sieht hierbei vor, mögliche Situationen mit Gefahr für Leib und Leben zu identifizieren. Über eine Risikobewertung wird die Relevanz bzw. Gefährlichkeit von Situationen bestimmt. Bei besonders gefährlichen Situationen sind zur Vermeidung systematischer Fehler weitergehende Methoden anzuwenden. Bei zufälligen Fehlern ist eine quantitative Bewertung der elektronischen Komponenten mit einer maximal zulässigen Ausfallwahrscheinlichkeit gefordert. Im Folgenden werden die Besonderheiten bei der Zertifizierung von KI umrissen. Einen umfassenderen Überblick bietet die Publikation von (Kutz et al. 2023).

2.1 Unterschiede von klassischer Zertifizierung zu KI-Zertifizierung

Da bisher für den AI Act noch keine harmonisierten Normen vorliegen, ist eine generelle KI-Zertifizierung noch nicht durchführbar. In bestimmten Bereichen wie der Medizintechnik, einem traditionell stark regulierten Bereich, ist es hingegen möglich, einzelne KI-Funktionen als Teil eines Medizintechnikproduktes zertifizieren zu lassen, zum Beispiel im Abgleich zur EU-Verordnung für Medizinprodukte (Medical Device Regulation, MDR)[1].

KI und ML-Begriffseinordnung

Künstliche Intelligenz (KI) umfasst Technologien und Methoden, die es Computersystemen ermöglichen, eigenständig Aufgaben zu bewältigen und Probleme zu lösen. Ein zentraler Aspekt der KI ist das maschinelle Lernen (ML). Hierbei werden Modelle mit vorhandenen Daten trainiert, um Muster und Informationen zu identifizieren und darauf basierend Vorhersagen oder Entscheidungen zu treffen. Diese trainierten Modelle können anschließend auf neue, unbekannte Datensätze angewendet werden. Der Trainingsprozess nutzt spezifische Trainingsdaten, um das Modell zu entwickeln, und separate Testdaten, um die Genauigkeit und Leistungsfähigkeit des Modells zu überprüfen.

1 https://eur-lex.europa.eu/legal-content/DE/ALL/?uri=CELEX%3A32017R0745

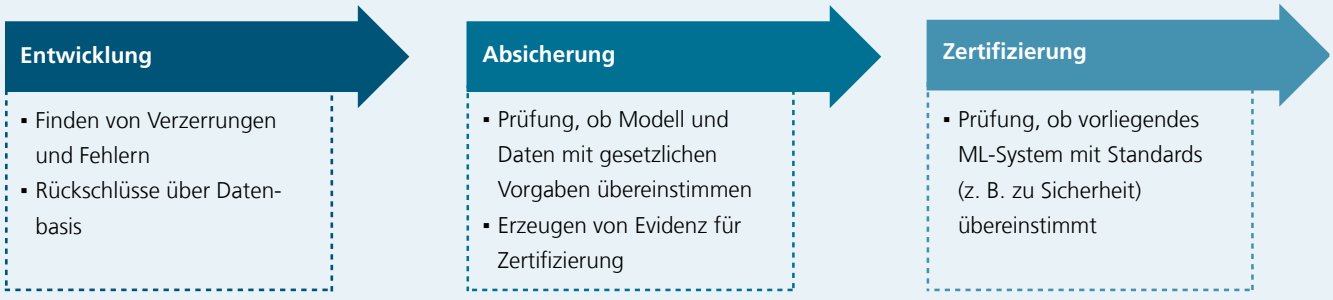
Während klassische Softwarekomponenten auf manuell definierten Funktionen basieren, deren Logik explizit von Menschen festgelegt wird und somit prüfbar ist, besteht der Kern von KI darin, aus Daten zu lernen, während Entwicklerinnen und Entwickler nur die Modell-Struktur vorgeben. Dies führt oft zu hochkomplexen Black-Box-Modellen, die Ergebnisse liefern, ohne den zugrunde liegenden Entscheidungsprozess – von den Eingabedaten zur Modellausgabe – zu erklären. Ohne Erklärung kann die Entscheidung nicht nachvollzogen werden, was wiederum zur Folge hat, dass eine Begründung für eine Entscheidung nicht gegeben ist. Im Sinne der ISO 26262 (siehe oben) müssten Fehler solcher Systeme durch die mangelnde Nachvollziehbarkeit als zufällig gelten, wobei wiederum – ebenfalls durch die mangelnde Nachvollziehbarkeit – der Nachweis über das zuverlässige Einhalten einer maximal zulässigen Ausfallwahrscheinlichkeit nicht möglich ist.

Zusätzlich sind manche KI-Systeme darauf ausgelegt, im laufenden Betrieb kontinuierlich weiter zu lernen. In solchen Fällen verändert sich diese hochkomplexe Funktion kontinuierlich und der Entscheidungsprozess der KI mit ihr, was bedeutet, dass eine Prüfung der Black-Box nach einiger Zeit veraltet sein kann. Diese Probleme ergeben sich teils aus dem grundsätzlichen Einsatzzweck von KI-Modellen: KI wird meist dann eingesetzt, wenn die manuelle Spezifikation von Programmen/Modellen nicht möglich ist, weil komplexe Probleme und Ursache-Wirkungs-Zusammenhänge vorliegen, die nicht oder nur mit sehr hohem Aufwand analytisch verstanden, formuliert oder gelöst werden konnten. Des Weiteren birgt KI immer Risiken im Zusammenhang mit den verwendeten Daten, beispielsweise wenn die ausgewählten Trainingsdaten nicht zur Aufgabenstellung passen, nicht ausreichend Daten gesammelt wurden oder es zu Umgebungsänderungen kommt. Selbst wenn eine KI zum Zeitpunkt der Entwicklung gut funktioniert, kann sich dies im Verlauf der Zeit durch kleinste Änderungen in der Umwelt verschlechtern. Schon die Änderung der Beleuchtungsverhältnisse in einer Fabrik könnte so dafür sorgen, dass ein System zur kamerabasierten Qualitätsprüfung nicht mehr ordnungsgemäß funktioniert. Zur Absicherung ist also nicht nur die Funktionalität des Modells während der Entwicklung zu überprüfen, sondern auch die kontinuierliche Überwachung der Repräsentativität von Trainings- und Testdaten, was die Absicherung von KI zu einer anspruchsvollen Aufgabe macht.

2.2 Potenzial von XAI für die Absicherung von KI

Zur Bewältigung der oben genannten Herausforderungen könnten Methoden aus dem Bereich XAI eingesetzt werden. XAI zielt darauf ab, die Entscheidungsprozesse von KI-Systemen für Menschen verständlich und überprüfbar zu machen. Bezüglich der Absicherung und Zertifizierung von KI sind hier verschiedene Zielsetzungen von XAI vorstellbar, wie in Abbildung 1 dargestellt. So könnte XAI schon frühzeitig im Entwicklungsprozess eingesetzt werden, um Fehler und Verzerrungen (Bias) in der Modellentwicklung zu finden oder das Verständnis der zugrunde liegenden Datenbasis zu verbessern. Zu einem späteren Zeitpunkt in der Entwicklung könnte mittels XAI geprüft werden, ob das trainierte KI-Modell und die zugehörige Datenbasis mit rechtlichen Vorgaben bzw. relevanten Normen übereinstimmen. In diesem Schritt könnte XAI ebenfalls genutzt werden, um Evidenz über das KI-Modell zu erzeugen, die dann wiederum als Grundlage der Zertifizierung des Modells dienen kann. Nachfolgend könnte XAI helfen, nach einem eingetretenen Schadensfall die rechtliche Verantwortung für ebendiesen Schaden zu klären. So wurden zum Beispiel in einem der ersten Gerichtsverfahren zur Haftbarkeit von KI-Entscheidungen Methoden aus dem Bereich XAI eingesetzt, um zu klären, ob die Funktionsweise eines Algorithmus von Trivago bewusst falsch wiedergegeben wurde, so beschrieben von Fraser et al. (2022). Durch die Abhängigkeit vom jeweiligen zu klärenden Sachverhalt ist allerdings davon auszugehen, dass für den XAI-Einsatz bezüglich Haftbarkeit keine allgemeinen Bedingungen beschrieben werden können. Deshalb beschränkt sich dieses Whitepaper auf die ersten drei Schritte – Entwicklung, Absicherung und Zertifizierung von KI.

Abbildung 1: Mögliche Anwendungen für XAI im Rahmen von Entwicklung, Absicherung und Zertifizierung von KI.



3. Erklärbare Künstliche Intelligenz (XAI)

Um einen Einblick zu geben, was mit aktuellen XAI-Methoden bereits möglich ist, werden im Folgenden kurz die relevanten Begrifflichkeiten und zugehörige Anschauungsbeispiele dargestellt.

3.1 Globale und lokale Erklärungen

Bei XAI unterscheidet man zwischen globalen Erklärungen, die den generellen Entscheidungsprozess eines Modells darlegen, und lokalen Erklärungen, die auf einzelne Entscheidungen abzielen. Globale Erklärungen beschreiben Regeln, die das Modell generell anwendet. Im Beispiel eines KI-Modells zur Beurteilung der Kreditwürdigkeit von Personen könnte eine globale Erklärung folgendermaßen lauten: »Alle Personen mit einem Einkommen über X Euro erhalten einen Kredit«. Lokale Erklärungen hingegen beleuchten individuelle Fälle wie etwa: »Person P wurde der Kredit aufgrund eines niedrigen Einkommens und einer kurzen Beschäftigungsdauer verweigert«. Aufgrund der Komplexität von KI-Systemen sind globale Erklärungen oft schwierig zu erfassen, wohingegen lokale Erklärungen meist greifbarer und anwendungsorientierter sind.

3.1.1. Globale Erklärungen

Obwohl die Umsetzung globaler Erklärungen eine Herausforderung darstellt, wird an dieser Stelle zum Verständnis der Interviewergebnisse auf konzeptbasierte Erklärungen eingegangen. Diese fassen wichtige menschenverständliche Konzepte zusammen, welche für die Entscheidungen eines KI-Systems von Bedeutung sind. Beispielsweise können in einem Modell zur Identifikation von Fahrzeugen auf Bildern »Reifen« als essenzielle Komponente identifiziert werden. Auch für Text sind solche Erklärungen erzeugbar. Mittels dieser Konzepte sind je nach Methode sowohl Modelle als auch Einzelentscheidungen erklärbar. Allerdings sind entsprechende Konzepte oft schwer zu finden und die entsprechenden XAI-Methoden meist modellspezifisch und nur mit großem Aufwand anwendbar.

Begriffserklärung

Modellspezifisch bedeutet im Kontext einer XAI-Methode, dass diese nur für eine bestimmte Art KI-Modell (zum Beispiel neuronale Netze) funktioniert und damit auch meist an ein spezifisches Modell und den damit verbundenen Anwendungsfall angepasst wer-den muss. Im Gegensatz dazu stehen modell-agnostische Methoden, die für alle Modelltypen funktionieren und oft mit weniger Implementierungsaufwand ver-bunden sind.

3.1.2. Lokale Erklärungen

Die einfachste Form lokaler Erklärungen sind Feature Attributionen, die jedem Eingabeparameter eine Bedeutung beimessen, zum Beispiel »Das Einkommen hatte den stärksten negativen Einfluss auf die Kreditentscheidung«. Solche Erklärungen sind verbreitet und können beispielsweise mit Tools wie SHAP (Lundberg und Lee 2017) erzeugt werden. Allerdings sind sie für Laien oft verwirrend, da die reine Wichtigkeit eines Parameters nicht unmittelbar aufzeigt, wie sich Veränderungen an diesem auf die Modellausgabe auswirken. Sogenannte kontrafaktische Erklärungen (engl. Counterfactuals) sind dagegen intuitiver und nachvollziehbarer, da sie gewissermaßen als Handlungsanweisung verstanden werden können, welche Parameter zu modifizieren sind, damit sich die KI-Entscheidung ändert. Der Begriff »kontrafaktisch« bezieht sich hierbei darauf, dass meist mit einem Gegenbeispiel zu dem real vorliegenden (»faktischen«) Fall argumentiert wird, so zum Beispiel »Person P bekäme einen Kredit, wenn ihr Einkommen um X höher wäre«. In der Praxis sind solche Erklärungen jedoch schwierig zu generieren und liefern je nach Anwendungsfall oft unklare oder unrealistische Empfehlungen.

Die Herausforderung besteht darin, XAI-Methoden effektiv einzusetzen, um vertrauenswürdige KI-Systeme zu entwickeln, ihre Funktionsweise abzusichern und sie letztendlich zertifizieren zu können. Hierbei gilt es, ein Gleichgewicht zwischen technischer Machbarkeit, Verständlichkeit für den Endnutzenden und regulatorischen Anforderungen zu finden.

Erschwerend kommt hinzu, dass die Anwendbarkeit und Interpretierbarkeit von Erklärverfahren je nach zugrunde liegendem Datentyp unterschiedlich sind. Während Feature-Attributionen für die meisten Modalitäten anwendbar sind, helfen diese nur weiter, wenn ein Mensch sie auch interpretieren kann und die entsprechenden Zusammenhänge versteht. Bei Bildern und Text funktioniert das in der Regel relativ gut, da hier beispielsweise einzelne Worte oder Bildbereiche hervorgehoben werden. Hierbei liefern verschiedene XAI-Verfahren allerdings unterschiedliche Ergebnisse für das gleiche KI-Modell, bekannt als sogenanntes »Disagreement Problem«. Abbildung 2 verdeutlicht dies; für die Klassifikation des Schmetterlings im linken Bild liefern die sechs dargestellten Verfahren teils unterschiedliche Erklärungen.

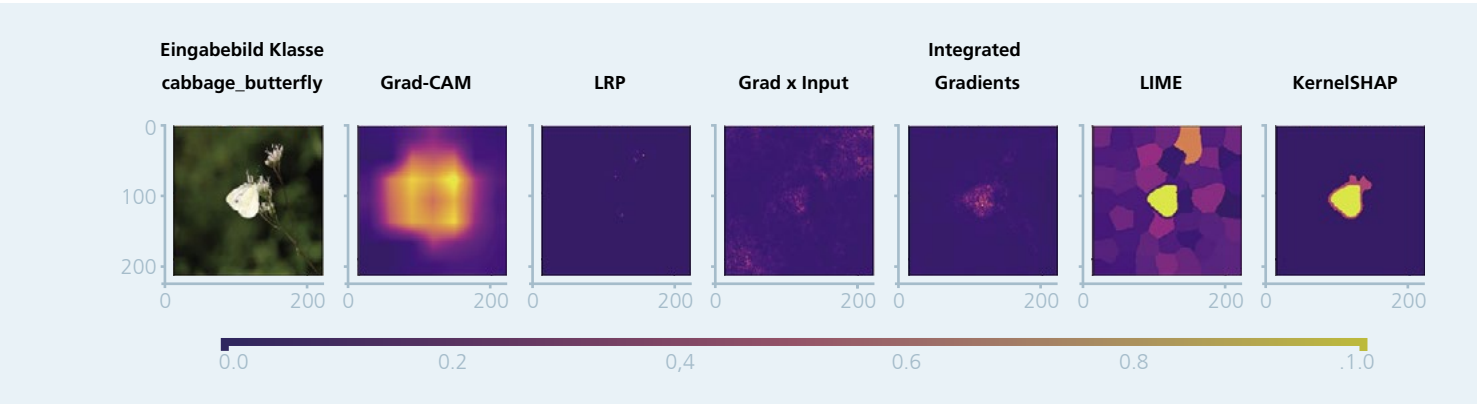


Abbildung 2: Lokale Feature Attribution Erklärungen (sog. Heatmaps oder Saliency Maps) für ein Bild eines Kohlweißlings. Bild aus Schaaf et al. (2021); Schmetterlingsbild: Yongseok Lee.

3.2 Eigenschaften von XAI-Methoden

Um XAI-Methoden zuverlässig für verschiedene Anwendungszwecke nutzen zu können, sollten klare Anforderungen an Erklärungen bzw. an für Erklärungen genutzte Methoden gestellt werden. Eine mögliche Grundlage für solche Anforderung bieten Eigenschaften von XAI-Methoden wie jene von Nauta et al. (2022), die im Folgenden kurz beschrieben sind.

- **Korrektheit (Correctness)** gibt an, ob und inwieweit eine Erklärung mit dem dazugehörigen KI-Modell übereinstimmt. Obwohl dies sehr grundlegend für Erklärungen klingt, sind Erklärungen von XAI-Methoden nicht notwendigerweise korrekt. Insbesondere bei den sehr populären stichprobenbasierten Verfahren LIME und SHAP kann keine Korrektheit garantiert werden.

- **Konsistenz (Consistency)** bedeutet, dass gleiche Eingaben stets zu gleichen Erklärungen führen. Das ist insbesondere für stichprobenbasierte Methoden relevant, die in ihrem Ausgabeverhalten nicht deterministisch sind. Eine weitere Quelle inkonsistenter Erklärungen können verschiedene Parameter-Einstellungen einer XAI-Methode sein.

- **Kontinuität (Continuity)** fordert, dass minimale Veränderungen in den Eingaben oder Ausgaben des Modells nur minimale Änderungen in der Erklärung nach sich ziehen sollten. Ein Beispiel hierfür sind Saliency-Maps bei Bildern: Verändert sich die Modellvorhersage trotz Überlagerung des Bildes mit Rauschen nicht, sollte auch die Erklärung (fast) gleichbleiben.

- **Kontrastivität (Contrastivity)** zielt auf die Unterscheidungskraft von Erklärungen ab. Besonders kontrastreiche Erklärungen zeigen klar auf, wie sich eine gewählte Entscheidung von anderen, nicht gewählten Alternativen unterscheidet.

- Die **Kovarianz-Komplexität (Covariate Complexity)** gibt die Komplexität der Merkmale und ihrer Wechselwirkungen innerhalb einer Erklärung wieder. Die einfachste Herangehensweise wäre die direkte Nutzung der Eingabemerkmale ohne Beachtung ihrer Interaktionen, was zu Lasten der Korrektheit gehen kann. Eine etwas komplexere Erklärung könnte lineare oder monoton verlaufende Feature-Interaktionen beinhalten.

- **Vollständigkeit (Completeness)** misst, in welchem Umfang das Erklärungsobjekt – das KI-Modell, die Daten oder eine einzelne Vorhersage – in der Erklärung abgebildet wird. Hier erstreckt sich die Bandbreite von der Darstellung des gesamten KI-Modells mit allen mathematischen Operationen bis hin zur Reduktion auf ein einzelnes Eingabemerkmal als Hauptfaktor für ein bestimmtes Ergebnis. Per Definition sind White-Box-Modelle vollständige Erklärungen ihrer selbst, können allerdings häufig auf unterschiedliche Arten und Weisen dargestellt werden.

- **Kompaktheit (Compactness)** beschreibt die Größendimension einer Erklärung. Zu umfangreiche Erklärungen können Nutzerinnen und Nutzer überfordern, während zu kurze Erklärungen zu wenig Einblick in eine Entscheidung oder das Modellverhalten geben.

- **Kompositionalität (Compositionality)** betrachtet die Darbietungsweise einer Erklärung, d. h. ihr Format, ihre Organisation und Struktur. Unterschiedliche Präsentationsformen können die Rezeption einer Erklärung verbessern, etwa durch das Bereitstellen von high-level Informationen (z. B. zur grundlegenden Funktionsweise von KI-Algorithmen) oder durch das Verwenden von einfacher Terminologie.

- **Unsicherheitsschätzungen (Confidence)** unterstützen die Anwendung von XAI auf zweierlei Weise: Sie können sowohl das Vertrauen des Modells in seine eigene Vorhersage als Teil der Erklärung anzeigen als auch die Zuverlässigkeit der Erklärung selbst, falls Erklärungen nicht per se korrekt sind.

- Der **Kontext (Context)** einer Erklärung, insbesondere die Kenntnis des Adressaten über ein Subjekt, z. B. über ein bestimmtes Forschungsfeld, ist entscheidend, um diese an individuelle Bedürfnisse der Nutzerinnen und Nutzer anzupassen und ihre praktische Anwendung zu optimieren.

- **Kohärenz (Coherence)** beschreibt, ob eine Erklärung mit vorhandenem Vorwissen – sei es das von Nutzerinnen und Nutzern, allgemeiner Konsens oder etablierte Überzeugungen – übereinstimmt und dadurch verständlicher wird. Obwohl Menschen ihr Vorwissen nutzen, um die Vertrauenswürdigkeit einer Erklärung einzuschätzen, unterscheidet sich Kohärenz von Richtigkeit: Eine Erklärung kann korrekt sein, also die Logik eines Modells widerspiegeln, ohne mit den vorherigen Überzeugungen von Nutzerinnen und Nutzern übereinzustimmen.

- **Kontrollierbarkeit (Controllability)** schließlich bezieht sich darauf, wie und in welchem Umfang mit einer Erklärung interagiert werden kann. Dadurch können individuelle Erklä-rungsbedarfe befriedigt werden (siehe Kontext). Zudem verbringen Nutzerinnen und Nutzer mehr Zeit mit interaktiven Erklärungen, was dazu beiträgt, dass diese besser verstanden werden (Cheng et al. 2019).

3.3 Bekannte Probleme im Forschungsfeld XAI

In der Literatur besteht kein Konsens darüber, wie wichtig die zuvor beschriebenen XAI-Eigenschaften jeweils sind und wie diese geprüft werden können, insbesondere, da sich verschiedene Bewertungsmethoden für die gleiche Eigenschaft widersprechen können (Tomsett et al. 2019). Ebenfalls ist es möglich, dass sich wünschenswerte Eigenschaften von Erklärungen widersprechen. So sollte eine Erklärung zum Beispiel korrekt und vollständig, aber auch möglichst kompakt und damit gut verständlich sein. Weitere Schwierigkeiten für den Einsatz von XAI sind ebenfalls bekannt und werden von Longo et al. (2023) folgendermaßen zusammengefasst:

- Da sich die KI-Landschaft entwickelt, müssen auch immer neue Erklärungsmöglichkeiten für neue KI-Verfahren gefunden werden.
- Der jetzige Stand der Technik bietet noch Raum für Verbesserungen, z. B. hin zu robusteren Erklärungen.
- Die Bewertung von Methoden von Erklärungen ist bisher unklar (siehe oben).
- Innerhalb verschiedener Disziplinen unterscheidet sich die Terminologie. Dies führt zu Verständigungsproblemen und unklaren Beziehungen zwischen verschiedenen Konzepten.
- Der Zusammenhang von XAI mit anderen Konzepten der vertrauenswürdigen KI wie Sicherheit und Fairness ist bisher noch nicht ausreichend erforscht.
- Erklärungen sollten besser an menschliche Bedürfnisse angepasst werden, z. B. durch high-level- statt low-level-Erklärungen.
- Erklärungen sollten bestenfalls an Stakeholder, Domänen oder bestimmte Einsatzziele angepasst werden.
- Bisher wird noch nicht genug getan, um negative Auswirkungen von Erklärungen (z. B. durch Automation Bias oder bewusst falsche Erklärungen) abzufangen.
- Der gesellschaftliche Einfluss von XAI könnte sich durch bestimmte Themensetzungen verbessern, z. B. die Umsetzung des Rechts auf Vergessenwerden oder den Einbezug von zukünftig durch ein KI-System Betroffenen in dessen Erstellung.

Trotz bestehender Schwierigkeiten bieten XAI-Methoden das Potenzial, die Entwicklung und Überprüfung von KI-Systemen zu verbessern, doch aktuell herrscht ein Mangel an systematischen Ansätzen für ihren effektiven Einsatz. Dieses Defizit ist teils auf technische Herausforderungen zurückzuführen, etwa bei der Qualität generierter Erklärungen, teils auf die unzureichende regulatorische Orientierung. Obwohl der AI Act in Artikel 13 Transparenz und Erklärbarkeit einfordert, bleibt die Definition konkreter Standards aus. Die DIN Spec 92001-3 widmet sich der Erklärbarkeit, jedoch beschränkt sie sich auf die Definition von Begrifflichkeiten und liefert keine praxisbezogenen Handlungsanleitungen.

Der derzeitige Stand der Regulierung spiegelt die Unsicherheit in der Anwendung von XAI innerhalb der Zertifizierung wider: Zwar erscheint ein menschlich nachvollziehbarer Entscheidungsprozess von KI-Systemen als intuitiver Kern der Zertifizierung. Doch ob diese Nachvollziehbarkeit des Entscheidungsprozesses zwangsweise für eine erfolgreiche Zertifizierung notwendig ist und wie hierfür die praktische Umsetzung aussehen könnte, bleibt unbestimmt.

Um zu untersuchen, wie XAI bisher eingesetzt wird und ob sich dabei die Probleme aus der Praxis mit den oben beschriebenen decken, wurden Interviews durchgeführt, deren Ergebnisse im folgenden Kapitel dargestellt werden.

4. Interviewergebnisse

Im Folgenden werden die qualitativen Interviewergebnisse aus 15 Experteninterviews bezüglich der Anwendung von XAI und der besonderen Herausforderungen in der Zertifizierung von KI mittels XAI zusammengefasst. Eine Kurzfassung der Ergebnisse findet sich ebenfalls in Tabelle 1 am Ende von Abschnitt 4.1. Die Ergebnisse dieser Befragung wurden in ähnlicher Form ebenfalls in einem wissenschaftlichen Paper publiziert (Fresz et al. 2024a).

4.1 XAI in der Praxis

Die Interviewpartnerinnen und -partner stammen aus Industrieunternehmen, Prüf- und Zertifizierungsstellen sowie anwendungsbezogenen Forschungseinrichtungen. Ihre fachlichen Hintergründe sind überwiegend in der Informatik, Mathematik, Ingenieurstechnik, Psychologie und Medizintechnik anzusiedeln.

Generell zeigte sich in den Interviews, dass Transparenz und Erklärbarkeit als wichtige Aspekte in der Entwicklung von KI-Anwendungen angesehen werden, in den meisten Fällen allerdings noch keine entsprechenden Methoden in die gängigen Entwicklungsprozesse integriert sind. Dies liegt teilweise unter anderem daran, dass bei der externen Beauftragung von KI-Projekten nicht ausreichend Mehrwert durch Erklärbarkeit gesehen wird, um hierfür die Finanzierung aufzubringen. Wenn Methoden der Erklärbarkeit zum Einsatz kommen, sind die Zielsetzungen vielfältig und meist recht abstrakt, so zum Beispiel die positive Außenwirkung solcher Projekte oder die generelle Compliance mit Behörden oder Kundenanforderungen. Weitere Zielsetzungen umfassen das Abfangen von Fehlern in KI-Systemen (insbesondere, bevor diese in deutlich kostenintensiveren Testphasen wie Feldtests Probleme bereiten) und die Möglichkeit zur Kommunikation über die Fähigkeiten und Funktionsweise von KI-Methoden mit anderen Abteilungen innerhalb des Unternehmens (zum Beispiel bezüglich Compliance- und Ethik-Prüfungen).

Über die Performance-Bewertung von XAI-Methoden ließ sich in den Interviews – auch bedingt durch die Vielzahl an verschiedenen Zielsetzungen des Einsatzes von Transparenz- und XAI-Methoden – keine wirkliche Bewertungsstruktur der Expertinnen und Experten ausmachen. Ein weiterer Grund hierfür lässt sich in der Bandbreite der eingesetzten XAI-Methoden vermuten: So finden neuro-symbolische KI-Systeme, graphen- und konzeptbasierte Erklärungen, White-Box-Modelle und Feature-Importance-Methoden wie SHAP Einsatz. Hierbei wurde in mehreren Interviews darauf hingewiesen, dass es durch die Anzahl an möglichen Methoden und die Unklarheiten bei der Bewertung der jeweiligen Zielsetzung in der Praxis schwierig sei, herauszufinden, welche Methode für den entsprechenden Anwendungszweck besonders geeignet ist. Oft werden dann Methoden genutzt, deren Anwendung besonders einfach ist und über deren Implementierung bereits umfangreiche Informationen zur Verfügung stehen. Hier ist insbesondere auf die frei verfügbaren Implementierungen von SHAP zu verweisen, wobei SHAP jedoch von den Interviewten und in der wissenschaftlichen Literatur aus anderen Gründen kritisch gesehen wird (Sundararajan und Najmi 2018; Kumar et al. 2020).

Trotz der bekannten Probleme berichten die Interviewten über die erfolgreiche Anwendung von XAI-Verfahren, insbesondere beim Finden von Fehlern in bestehenden KI-Systemen, der Plausibilitätsprüfung von Modellen während der Entwicklung oder bei der Verbesserung des Datenverständnisses. Allerdings zeigt sich auch, dass jetzige XAI-Verfahren für andere Einsatzzwecke nicht gut geeignet sind. Oft scheitern entsprechende Projekte und Erklärungen an ähnlichen Gründen: Entweder sind sie für die Zielpersonen ungeeignet und somit unverständlich, Fachexpertinnen und -experten haben nicht ausreichend Zeit, mit den Erklärungen zu interagieren oder die eigentlich zu findenden Zusammenhänge sind – teils durch fehlende KI- oder Domänenexpertise – unbekannt, womit es schwerfällt,

gefundene Zusammenhänge zu verifizieren. Zusätzlich kompliziert wird der Einsatz von XAI durch die schwierige Anwendung der meisten XAI-Methoden, da diese nur mit großem Aufwand implementiert werden können, teilweise instabil sind und somit nicht replizierbare Ergebnisse erzeugen, oder Methoden für spezifische Datentypen wie beispielsweise Zeitreihen bisher nicht ausreichend erforscht sind. So sind zum Beispiel gängige Feature-Importance-Verfahren wie SHAP für Zeitreihen grundsätzlich nutzbar, können aber relevante Informationen wegen ihrer grundsätzlichen Funktionsweise nicht identifizieren. Bei Zeitreihen finden sich relevante Informationen oft im Frequenzbereich, dieser steht gängigen XAI-Methoden allerdings nicht zur Verfügung.

Häufig wird in der wissenschaftlichen Literatur auf die Nutzung von XAI-Methoden zur Schaffung von Vertrauen bei Endverbrauchern hingewiesen. Allerdings wurde insbesondere diese Anwendung in den Interviews oft kritisch gesehen: Durch die interne Komplexität von XAI-Methoden wird das Problem einer nicht vertrauenswürdigen Black-Box (des KI-Systems) auf eine weitere Black-Box (das XAI-System) verschoben, deren Vertrauenswürdigkeit ebenfalls infrage steht, da bisherige XAI-Methoden selbst teils umstritten sind. Hier wurde speziell auf das Disagreement-Problem hingewiesen – verschiedene XAI-Methoden liefern für eine Entscheidung eines KI-Modells unterschiedliche Erklärungen. Somit sei unklar, was die »wahre« Erklärung für eine Entscheidung ist. Als möglicher Ausweg aus dem durch die doppelte Black-Box entstehenden Vertrauensproblem wurden Schulungen bezüglich KI und XAI für KI-Anwenderinnen und Anwender genannt.

4.2 XAI für Absicherung und Zertifizierung

Weitere Anwendungsfälle für XAI finden sich rund um den AI Act und die Zertifizierung von KI-Systemen (siehe Kapitel 2). Hier spielen für den rechtssicheren Einsatz von XAI-Methoden insbesondere die Messbarkeit von »angemessener« Transparenz und Erklärbarkeit eine wichtige Rolle. In den Interviews zeigte sich, dass hier sowohl bezüglich der Messbarkeit als auch bezüglich der rechtlichen Vorgaben angemessener XAI-Methoden noch größere Lücken gesehen werden. Da sich die XAI-Methoden je nach Einsatzzweck deutlich unterscheiden, müsste sich auch das jeweilige Ziel der Messung und somit das entsprechende Maß unterscheiden. Hierfür existieren zwar technische Metriken, diese werden in der Praxis aber häufig als nicht zielführend angesehen. Der Grund dafür ist, dass sie zwar Informationen über relevante Eigenschaften wie Korrektheit liefern können, aber keinen wirklichen Aufschluss über die Performance in der relevanten Aufgabe geben. So kann eine Erklärung zum Beispiel korrekt, aber für Nutzerinnen und Nutzer grundsätzlich unverständlich sein. Der wesentliche Nutzen von technischen Messungen an XAI (und gleichzeitig größter Kritikpunkt an XAI-Methoden) stellt die Sicherstellung der Zuverlässigkeit von Erklärungen dar, zum Beispiel via Quantifizierung der Unsicherheit bzw. Konfidenz von Erklärungen. Im Kontext der Zertifizierung wird insbesondere die Vollständigkeit der Erklärung, also das Erfassen aller relevanter Informationen, die zu einer KI-Entscheidung führen, als relevante – aber vermutlich nicht erreichbare – Eigenschaft angesehen.

Um den tatsächlichen Nutzen der Erklärungen zu messen, wird auf Nutzerstudien verwiesen, welche vorzugsweise mit Domänenexpertinnen und -experten durchzuführen sind, zur Verringerung der Kosten allerdings auch mit Laien durchgeführt werden können. Auch diese Nutzerstudien sind nicht frei von Kritik. So können sie das Problem bergen, dass eher der gefühlte Nutzen einer Erklärung anstelle des tatsächlichen Nutzens gemessen wird. So könnten auch irreführende Erklärungen bei Nutzerinnen und Nutzern als positiv empfunden werden (Chromik et al. 2021).

Die sich aus der schwierigen Messbarkeit und der Unklarheiten beim Einsatz von XAI ergebenden Schlüsse unterscheiden sich bei den Expertinnen und Experten wesentlich in zwei Gruppen: Von manchen Expertinnen und Experten wird nur ein geringer Einfluss von XAI auf die Zertifizierung von KI-Systemen gesehen. Entweder weil bereits anderweitige Vorgaben, zum Beispiel aus der Medizinprodukteverordnung, existieren, oder weil XAI-Methoden nicht als ausreichend belastbar und

vollständig angesehen werden, insbesondere für komplexe Anwendungsfälle, für die die Erklärungen selbst wieder zu komplex werden. Andererseits wird argumentiert, dass der erfolgreiche Einsatz von XAI zur Verbesserung von KI-Modellen bzw. zum Finden von Fehlern und Bias in der Entwicklung führt und dass sich mittels XAI später auftretende Probleme bereits frühzeitig identifizieren lassen. Da dies das wesentliche Ziel der Absicherung darstellt, wird ein positiver Mehrwert von XAI für die Absicherung und Zertifizierung gesehen – allerdings nur in Kombination mit menschlichen Prüferinnen und Prüfern, nicht als Methodik zur Automatisierung von Zertifizierungsprozessen.

Ähnlich wie beim Nutzen von XAI für die Zertifizierung unterscheiden sich auch die mit zukünftigen Entwicklungen von XAI assoziierten Hoffnungen. So werden insbesondere vollständige, globale Erklärungen als nicht realisierbar angesehen, auch wenn diese für eine vollumfängliche Absicherung eines KI-Systems besonders hilfreich wären. Andererseits besteht die Hoffnung, dass neue XAI-Ansätze, darunter vor allem konzeptbasierte, aber auch neuro-symbolische und weitere, eine neue Darstellung von Erklärungen ermöglichen, um so doch die grundsätzliche Funktionsweise eines KI-Modells verständlich zu machen. Auch neue Darstellungen von Erklärungen, insbesondere datenbasierte und multi-modale, werden als aussichtsreich für das Forschungsfeld angesehen.

Während laut einigen Befragten XAI-Methoden durchaus die Möglichkeit bieten, Fehler in KI-Systemen zu erkennen, und somit in Zukunft bestenfalls direkt in Entwicklungsprozesse integriert werden sollten, werden im Zusammenhang mit der zukünftigen Prüfung von KI-Modellen mehrheitlich Methoden außerhalb von XAI genannt. So zum Beispiel die Prüfung von KI durch weitere KI-Systeme, gegenüber physikalischen Gesetzmäßigkeiten (wo angemessen), die formale Verifikation bestimmter Eigenschaften und die Unsicherheitsquantifizierung von KI-Entscheidungen. Mehrfach wurde darauf hingewiesen, dass – wo möglich – möglichst simple, intrinsisch interpretierbare KI-Lösungen verwendet werden sollen. Diese Einstellung findet sich ebenfalls in entsprechender Fachliteratur, so zum Beispiel in Rudin (2019). Im Extremfall wurde darauf hingewiesen, dass KI grundsätzlich nicht in Hochrisiko-Anwendungen Einsatz finden sollte. Insgesamt besteht hier der Wunsch, dass für die Zertifizierung von KI klare Vorgaben wie beispielweise bestimmte Metriken existieren sollten, während Methoden für die Erklärbarkeit solcher Systeme durch eine stärkere Nutzerinnen- und Nutzer- sowie Industriefokussierung ihre Potenziale besser entfalten könnten.

Trotz der genannten Alternativen und Einschränkungen wird XAI ein bedeutendes Potenzial in Bezug auf die Absicherung und Zertifizierung von KI-Systemen zugesprochen. Es wird jedoch deutlich, dass XAI allein in diesem Bezug nicht als umfassende Lösung dienen kann. Tabelle 1 liefert einen Überblick über die Einschätzungen der Interviewpartnerinnen und -partner zum Einfluss von XAI in den verschiedenen Anwendungsgebieten.

Tabelle 1: Zusammenfassung der Experteninterviews zu XAI und der Zertifizierung von KI-Systemen.

	Hohes Potenzial	**Geringes Potenzial**
XAI im Allgemeinen	▪ Kommunikation zwischen Domänen-/KI-Expertinnen und Experten ▪ Klare Informationen, wann welche XAI-Methode verwendet werden sollen	▪ Erklärungen für Laien ▪ Erklärungen, wenn zugrundeliegende Prozesse zu komplex oder bisher nicht gut verstanden sind
XAI in Zertifizierung / für sichere KI-Entwicklung	▪ Plausibilitätsprüfung des KI-Modells in der Entwicklung ▪ Entdeckung von Verzerrungen/Fehlern ▪ Verbesserung des Datenverständnisses	▪ Verlässliche Aussagen über die Sicherheit von KI-Systemen
Zukunft von XAI	▪ Erhöhter Fokus auf User-Bedürfnisse ▪ Neue Erklärungstypen (konzeptbasiert, multimodal, …) ▪ Unsicherheitsquantifizierung von (X)AI	▪ Aussagekräftige technische Messung von Transparenz/Erklärbarkeit
Zukunft der KI-Zertifizierung	▪ XAI als zusätzliche Evidenz für die Zertifizierung von KI ▪ Formale Verifikation sicherheitsrelevanter KI-Eigenschaften ▪ Neue KI-Ansätze (z. B. neuro-symbolisch)	▪ XAI als umfassende Antwort auf die Zertifizierung von KI

4.3 Diskussionsbeiträge

Neben den allgemeineren Beiträgen, die ursprünglich das Ziel dieser Befragung waren, kommentierten die Interviewten auch immer wieder spezifische Themen bezüglich XAI oder der Zertifizierung von KI. Diese Beiträge werden – in gekürzter, paraphrasierter Form – im Folgenden vorgestellt, da sie einen interessanten Einblick in die Unterschiede zwischen klassischer und KI-Zertifizierung und in bestehende Probleme von XAI-Verfahren geben.

4.3.1. Zertifizierung auf der Grundlage falscher Evidenz?

Zertifizierung prüft bisher, ob die Evidenz den Anforderungen der Standards und Normen entspricht. Es ist kein Verfahren etabliert, um zu prüfen, ob diese Evidenz korrekt ist.

In den Gesprächen mit den Expertinnen und Experten, insbesondere denen mit Erfahrung in der Zertifizierung von technischen Systemen, zeigte sich mehrfach, dass die Zertifizierung von KI neue Herausforderungen bezüglich der bisherigen Prozesse für Produktzertifizierungen bietet. So wurde insbesondere darauf hingewiesen, dass Zertifizierung bisher prüft, ob von Herstellern bereitgestellte Evidenz mit den Anforderungen aus Standards und Normen übereinstimmt. Dabei wird davon ausgegangen, dass diese Evidenz korrekt ist, also mit den verwendeten Prozessen bzw. dem tatsächlich entwickelten Produkt übereinstimmt. Sollte Evidenz nun mittels XAI erzeugt werden, kann diese Annahme eventuell nicht getroffen werden. So kann inkorrekte Evidenz entweder absichtlich (wie bisher ebenfalls) oder aber durch die Nutzung von XAI unabsichtlich erzeugt werden. Hier müssen insbesondere die Zuständigkeiten geklärt werden: Garantieren Hersteller für die Korrektheit der Evidenz, was nach bisherigem Stand von XAI kaum möglich sein wird, oder muss auf der Seite der Zertifizierer in Zukunft der Erzeugungsprozess der Evidenz mitbedacht werden, was wiederum tiefgreifende Kenntnisse bezüglich KI im Allgemeinen und XAI im Spezifischen voraussetzt?

Bis jetzt waren die »Uniformity Hypothesis« und die »Competent Programmer Hypothesis« wichtige Bausteine in der Entwicklung und Zertifizierung von sicherheitskritischer Software.

Ähnliche Herausforderungen wurden ebenfalls in Bezug auf die Entwicklung von sicherheitskritischen Anwendungen genannt. Bei diesen galt bisher zum einen die »Uniformity Hypothesis«, dass für Tests bestimmte Datenpunkte ausgewählt werden können, deren Erkenntnisse sich auf eine Äquivalenzklasse ähnlicher Daten generalisieren lassen. Zum anderen wurde unter der »Competent Programmer Hypothesis« angenommen, dass sicherheitsrelevante Software nicht vollkommen unvorhersehbare Fehler erzeugt, weil sie durch kompetente Programmiererinnen und Programmierer erzeugt wurde, die zufällig wirkende Fehler vermeiden können. Beide Annahmen sind durch die Black-Box-Natur und die Datenbasiertheit von KI jedoch verletzt: Für die Generalisierbarkeit von Tests bestimmter Daten sind Äquivalenzklassen nur schwierig zu finden und die Entscheidungsfindung eines KI-Systems wurde nicht explizit programmiert. Durch die Komplexität von KI-Modellen können so entstehende Fehler zufällig wirken. Allerdings sei ein guter Schritt in Richtung sichere KI, die in KI und XAI getroffenen Annahmen zu dokumentieren, was auch für Annahmen wie die »Uniformity Hypothesis« und die »Competent Programmer Hypothesis« oft nicht geschehe. Ein Experte wies explizit darauf hin, dass wissenschaftlicher Fortschritt typischerweise durch das Herausfordern von bestehenden Ideen entsteht. Im Bereich XAI führt das zu einer für die Praxis schwer zu durchdringenden Komplexität. So wurde XAI ursprünglich zur Bewertung bzw. Prüfung von KI vorgeschlagen, inzwischen existieren allerdings auch Metriken zur Prüfung von XAI und wiederum Metriken zur Prüfung dieser (Fresz et al. 2024b). Daraus folgt, dass ein Ziel der Anwendungsforschung nun sein könnte, explizite Empfehlungen für die Auswahl von XAI-Methoden für spezifische Anwendungsfälle zu treffen.

4.3.2. Zertifizierung von KI nach utilitaristischen Gesichtspunkten

Wenn KI zertifiziert werden soll, muss es eine Diskussion über Zertifizierung nach utilitaristischen Gesichtspunkten statt nach den bisherigen Werte-basierten Prozessen geben.

Durch die Unwägbarkeiten bei der Prüfung von KI-Systemen vermutete ein Experte, dass sich die Kultur der Zertifizierung grundsätzlich ändern müsse, um KI zuzulassen: Bisherige Zertifizierung ist prinzipiell durch gesellschaftliche Werte und Normen gesteuert. So kann zum Beispiel für die Norm »Sicherheit« eines technischen Systems ein Grenzwert festgelegt werden, der dann basierend auf einer Detailanalyse (zum Beispiel via Methoden wie FMEA) von einem Gesamtsystem eingehalten werden kann. Sollte dieser Wert nicht eingehalten werden, muss entsprechend seitens der Entwicklung gegengesteuert werden. Ein besonders bekanntes Beispiel für die Nachvollziehbarkeit von ethischen Werten in technischen Systemen findet sich im Bereich des autonomen Fahrens: das sogenannte »Trolley-Problem«. Bei diesem muss unmittelbar vor einem Unfall gewählt werden, welche der beteiligten Personen einem höheren Risiko für schwere Verletzungen ausgesetzt wird. Für KI sind solche Grenzwerte und ethischen Entscheidungen allerdings im Moment nicht ausreichend ermittelbar (siehe

oben). Der Experte vermutete deshalb, dass der Einsatz von KI mehr nach utilitaristischen Standpunkten bewertet werden müsste, also zum Beispiel: »Wenn durch den Einsatz von KI von insgesamt weniger Verletzten oder Toten im Straßenverkehr auszugehen ist, dann ist der Einsatz sinnvoll.«

4.3.3. Zuständigkeiten für die Vorgaben zu XAI

Erklärbarkeit ist auch ein soziales, nicht nur ein technisches Thema. Damit sind die üblichen Standardisierungsgremien nicht zuständig und teils überfordert.

Ein weiterer Kritikpunkt an dem jetzigen Vorgehen zur Absicherung und Zertifizierung bezieht sich auf den Prozess der Standard-Erstellung. So wird kritisiert, dass zuständige Organisationen wie DIN, CEN/CENELEC und ISO für die technische Standardisierung zuständig seien, aber über Themen wie Erklärbarkeit und fundamentale Überlegungen wie die Einhaltung und Aushandlung von ethischen Werten (siehe oben) explizit nicht technische, sondern gesellschaftspolitische Diskussionen zu führen seien. Auch hier ist insbesondere darauf hinzuweisen, dass für XAI technische Bewertungsmethoden existieren, diese von der Mehrzahl der Befragten aber nicht als zielführend angesehen wurden. Das bedeutet, dass rein technische Standardisierung die offenen Fragen um die Absicherung und Zertifizierung von KI vermutlich nicht lösen kann.

4.3.4. Potenzielle Limitierungen von XAI

An KI-Modelle werden sehr strikte Anforderungen bezüglich der Erklärbarkeit von Entscheidungen gestellt, während menschliche Entscheidungen selbst meist intuitiv getroffen werden und nicht umfänglich begründet werden können. Dabei wird XAI nie eine wirkliche Erklärung, sondern nur eine Rechtfertigung für KI-Verhalten liefern.

Ebenfalls kritisiert wurde, dass an KI deutlich detailliertere Anforderungen gestellt werden, als sie für Menschen tatsächlich erfüllbar sind. Hierfür wurde der Vergleich zu menschlichen Erklärungen bzw. explizit »Rechtfertigungen« gezogen. Während Menschen ihr Verhalten bei Aufforderung im Nachhinein zwar rechtfertigen, seien solche Rechtfertigungen nicht mit den tatsächlichen Beweggründen identisch, insbesondere für oft eher intuitiv getroffene Entscheidungen. An dieser Stelle wurde vermutet, dass auch für XAI das gleiche gelte: XAI könne zwar eine für Menschen verständliche Rechtfertigung von KI-Verhalten erzeugen, die tatsächliche Erklärung sei aber in der Berechnungskette des KI-Systems zu finden und damit zwar grundsätzlich nachvollziehbar, aber nicht in Gänze für Menschen verständlich. Hier könnte erwidert werden, dass die Idealvorstellung von XAI ermöglicht, eben diese Berechnungskette auf eine Art und Weise zusammenzufassen, dass doch eine korrekte Erklärung entsteht, die für Nutzerinnen und Nutzer Aufschlüsse über das KI-Verhalten bietet.

4.3.5. Fehlende User-Fokussierung in der XAI-Forschung

In der Forschung werden Cognitive Load und Interaktionszeit mit Erklärungen zu wenig beachtet.

Mehrere Teilnehmerinnen und Teilnehmer kritisierten, dass in der XAI-Forschung die expliziten Erfahrungen und Ziele von Fachexpertinnen und-experten nicht ausreichend berücksichtigt werden. So operiere die XAI-Forschung anscheinend unter der Annahme, dass zu allen Zeitpunkten vollständige Erklärungen generiert werden sollten. In der Praxis benötigen Fachexpertinnen und -experten wie Ärztinnen und Ärzte typischerweise nur in spezifischen Fällen Erklärungen.

Darüber hinaus verbessern sich die User-Erfahrung und die Interaktionszeit mit Erklärungen, wenn diese dazu dienen, den mit der Aufgabe verbundenen »Cognitive Load« zu reduzieren. Die Mehrheit der Benutzerinnen und Benutzer hat in ihren täglichen Routinen weder die Zeit noch die kognitiven Ressourcen, sich mit übermäßig komplexen Erklärungen auseinanderzusetzen. Somit wird effektiv das Hauptziel von XAI, die KI zugänglicher zu machen, untergraben. Zur leichteren Verständlichkeit der Erklärungen müssten diese kontextualisiert werden, erwähnte ein Experte. Dies könne die kognitive Belastung je nach spezifischer Implementierung potenziell erhöhen oder verringern. Obwohl Interaktionszeit und kognitive Belastung in der XAI-Literatur häufig nicht beachtet werden, gibt es mehrere Veröffentlichungen, die Ideen zur Reduzierung der kognitiven Belastung von Erklärungen untersuchen (Herm 2023).

4.3.6. Neue Paradigmen für XAI

Das direkte Feedback von XAI-Informationen in den Modelltrainingsprozess ist ein besonders aussichtsreicher Ansatz.

Als besonders vielversprechend für den Bereich XAI wurde die Kombination des Erklärungsprozesses mit der zugehörigen Modellverbesserung bezeichnet. So wird XAI bisher meist unidirektional gesehen – auch wenn Fehler und Verzerrungen in bestehenden Modellen gefunden werden können, bietet sich noch kein simpler Eingriff in Modell- oder Trainingsdaten an, um bestehende Probleme zu korrigieren. Ein neues Paradigma könnte hier die Möglichkeit bieten, mit Erklärungen zu interagieren, diese zu korrigieren und die Korrekturen wieder in den Modelltrainingsprozess zu integrieren. Somit könnten die aus XAI gewonnenen Erkenntnisse effizient für die Fehlerkorrektur von KI-Modellen genutzt werden.

5. Fazit und Zukunftsaussichten

Im Rahmen der durchgeführten Experteninterviews wurden potenzielle Entwicklungspfade für die Weiterentwicklung von XAI beleuchtet, die in diesem Abschnitt näher betrachtet werden. Diese Erkenntnisse liefern wichtige Impulse für die Gestaltung effektiver Zertifizierungsprozesse in Innovationsökosystemen.

Erkenntnisse aus den Experteninterviews
Durch die Experteninterviews konnte ein besseres Verständnis für die Rolle von XAI in Zertifizierungsprozessen gewonnen werden. So zeigte sich, dass die derzeitigen Fähigkeiten von XAI die strengen Anforderungen der Zertifizierungsstandards noch nicht erfüllen. Dies verdeutlicht eine kritische Lücke zwischen den theoretischen Vorteilen von XAI-Methoden und deren praktischer Anwendbarkeit in Zertifizierungsprozessen.

Es wurde ebenfalls deutlich, dass der Übergang von XAI als Debugging-Werkzeug zu einem Zertifizierungsinstrument ein besseres Verständnis (wenn nicht sogar eine Garantie) für die Korrektheit von Erklärungen und Verhaltensweisen von KI-Systemen erfordert. Dieser Übergang ist entscheidend, um XAI effektiv in Zertifizierungsprozessen zu integrieren, und steht noch am Anfang seiner Entwicklung.

Gesellschaftliche und ethische Implikationen
Die Diskussion um XAI berührt auch breitere gesellschaftliche und ethische Implikationen. Die Notwendigkeit, ethische Standards und gesellschaftliche Auswirkungen in die Zertifizierungsansätze zu integrieren, erfordert einen Paradigmenwechsel von rein technischen zu ganzheitlichen Bewertungen, die die gesellschaftlichen Implikationen der KI-Technologien berücksichtigen.

Bedeutung für Innovationsökosysteme
Diese Erkenntnisse fließen direkt in die Arbeit von Innovationsökosystemen wie dem Ökosystem Heilbronn-Franken mit dem Innovationspark AI (IPAI) oder dem Cyber Valley ein. Derartige Ökosysteme fördern die Zusammenarbeit verschiedener Akteure aus Forschung, Industrie und öffentlichem Sektor, die gemeinsam an der Entwicklung und Implementierung neuer Zertifizierungsansätze arbeiten. Die enge Zusammenarbeit zwischen den Akteuren ermöglicht es, schnell auf technologische und marktbezogene Veränderungen zu reagieren und die Entwicklung von Zertifizierungsstandards voranzutreiben, die den Anforderungen moderner KI-Systeme gerecht werden.

Integration von XAI in Forschungsprojekte
Durch die Integration von XAI in Forschungsprojekte wird die Erklärbarkeit von KI von Beginn an berücksichtigt, was nicht nur Transparenz schafft, sondern auch dazu beiträgt, das Vertrauen in KI-Technologien zu stärken. Zusätzlich ermöglichen Innovationsökosysteme die Bildung spezialisierter Einheiten und Arbeitsgruppen, die eng mit akademischen und Forschungsinstitutionen verknüpft sind. Diese Einheiten können auf neueste wissenschaftliche Erkenntnisse zurückgreifen, um relevante und aktuelle Zertifizierungsstandards zu entwickeln, die ergänzend zu etablierten Zertifizierungsstellen wie dem TÜV agieren und die speziellen Anforderungen von KI-Systemen adressieren.

Zukunftsaussichten
In diesen Ökosystemen können auch Selbstverpflichtungsregeln und gemeinsame Standards entwickelt werden, die ein Siegel für geprüfte und verifizierte XAI-Produkte vergeben. Diese Form der Zertifizierung könnte entlang der gesamten Wertschöpfungskette von der Datengenerierung bis zur Anwendung implementiert werden und eine umfassende und lebensnahe Überprüfung der Systeme gewährleisten.

Abschließend lässt sich sagen, dass die Weiterentwicklung von XAI und deren Integration in Zertifizierungsprozesse eine interdisziplinäre Anstrengung erfordert. Nur durch die Zusammenarbeit von Forschung, Industrie und öffentlichen Akteuren können effektive und vertrauenswürdige Zertifizierungsstandards entwickelt werden, die den Herausforderungen moderner KI-Systeme gerecht werden und deren gesellschaftliche Akzeptanz fördern.

6. Danksagung

Wir bedanken uns herzlich für die zur Verfügung gestellte Zeit und Expertise der Interviewten. Während hier auf eine persönliche Nennung verzichtet wird, findet sich im Folgenden – in Absprache mit den Interviewten – eine Liste an Organisationen, aus denen die Beteiligten stammen.

- IAV GmbH
- TÜV SÜD Digital Service GmbH
- Frauhofer IIS Projektgruppe Comprehensible AI und Bamberger Zentrum für KI (BaCAI) an der Otto-Friedrich-Universität Bamberg
- ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland.
- Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland.
- Centre for Assuring Autonomy, Department of Computer Science, University of York, UK
- Fraunhofer-Institut für Kognitive Systeme IKS
- VDI/VDE Innovation + Technik GmbH
- Experian GmbH
- Bundesdruckerei GmbH
- C-AI Group, Freie Universität Berlin

Für die Verfeinerung des Entwurfes dieses Whitepapers wurde FhGenie (Version 4.0 Turbo) verwendet.

Literaturverzeichnis

Cheng, H.; Wang, Ruotong; Zhang, Zheng; O'Connell, Fiona; Gray, Terrance; Harper, F. M.; Zhu, Haiyi (2019): Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Chromik, Michael; Eiband, Malin; Buchner, Felicitas; Krüger, Adrian; Butz, Andreas (2021): I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In: Tracy Hammond, Katrien Verbert, Dennis Parra, Bart Knijnenburg, John O'Donovan und Paul Teale (Hg.): 26th International Conference on Intelligent User Interfaces. IUI ,21: 26th International Conference on Intelligent User Interfaces. College Station TX USA, 14 04 2021 17 04 2021. New York, NY, USA: ACM, S. 307–317.

Fraser, Henry; Simcock, Rhyle; Snoswell, Aaron J. (2022): AI Opacity and Explainability in Tort Litigation. In: Timo Speith (Hg.): A review of taxonomies of explainable artificial intelligence (XAI) methods. FAccT ,22: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea, 21 06 2022 24 06 2022. New York, Saarbrücken: ACM; Saarländische Universitäts- und Landesbibliothek, S. 185–196.

Fresz, Benjamin; Göbels, Vincent Philipp; Omri, Safa; Brajovic, Danilo; Aichele, Andreas; Kutz, Janika et al. (2024a): The Contribution of XAI for the Safe Development and Certification of AI: An Expert-Based Analysis. Online verfügbar unter https://arxiv.org/pdf/2408.02379.

Fresz, Benjamin; Lörcher, Lena; Huber, Marco (2024b): Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery (FAccT '24), S. 1–19.

Herm, Lukas-Valentin (2023): Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study. Online verfügbar unter http://arxiv.org/pdf/2304.08861v1.

Kumar, I. Elizabeth; Venkatasubramanian, Suresh; Scheidegger, Carlos; Friedler, Sorelle (2020): Problems with Shapley-value-based explanations as feature importance measures, 30.06.2020. Online verfügbar unter https://arxiv.org/pdf/2002.11097.pdf.

Kutz, Janika; Göbels, Vincent Philipp; Brajovic, Danilo; Fresz, Benjamin; Renner, Niclas; Omri, Safa et al. (2023): KI-Zertifizierung und Absicherung im Kontext des EU AI Act. Hg. v. Fraunhofer-Gesellschaft. Online verfügbar unter https://publica.fraunhofer.de/entities/publication/6ab76f95-756c-4d52-98a3-fda7a9f959de/details.

Longo, Luca; Brcic, Mario; Cabitza, Federico; Choi, Jaesik; Confalonieri, Roberto; Del Ser, Javier et al. (2023): Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions, 30.10.2023. Online verfügbar unter http://arxiv.org/pdf/2310.19775.pdf.

Lundberg, Scott; Lee, Su-In (2017): A Unified Approach to Interpreting Model Predictions, 25.11.2017. Online verfügbar unter http://arxiv.org/pdf/1705.07874.

Nauta, Meike; Trienes, Jan; Pathak, Shreyasi; Nguyen, Elisa; Peters, Michelle; Schmitt, Yasmin et al. (2022): From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. Online verfügbar unter http://arxiv.org/pdf/2201.08164v1.

Rudin, Cynthia (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 22.09.2019. Online verfügbar unter https://arxiv.org/pdf/1811.10154.pdf.

Schaaf, Nina; Wiedenroth, Saskia Johanna; Wagner, Philipp (2021): Erklärbare KI in der Praxis. Anwendungsorientierte Evaluation von xAI-Verfahren. Hg. v. Thomas Bauernhansl, Marco Huber und Philipp Wagner. Fraunhofer IPA. Online verfügbar unter https://publica.fraunhofer.de/handle/publica/300845.

Sundararajan, Mukund; Najmi, Amir (2018): The Many Shapley Values for Model Explanation. In:. Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery: PMLR, S. 9269–9278. Online verfügbar unter http://proceedings.mlr.press/v119/sundararajan20b/sundararajan20b.pdf.

Tomsett, Richard; Harborne, Dan; Chakraborty, Supriyo; Gurram, Prudhvi; Preece, Alun (2019): Sanity Checks for Saliency Metrics, 29.11.2019. Online verfügbar unter https://arxiv.org/pdf/1912.01451.

Impressum

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA
Nobelstraße 12
70569 Stuttgart
www.ipa.fraunhofer.de

Kontakt
Benjamin Fresz
Tel. +49 711 970-1404
benjamin.fresz@ipa.fraunhofer.de

Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO
Bildungscampus 9
74076 Heilbronn
www.iao.fraunhofer.de

Kontakt
Vincent Philipp Göbels
Mobil +49 152 22543933
vincent-philipp.goebels@iao.fraunhofer.de

Satz und Layout
Franz Schneider, Fraunhofer IAO

Titelbild
© Irina Strelnikova – Adobe Stock

Fraunhofer-Publica
http://dx.doi.org/10.24406/publica-3694

Kontakt

Benjamin Fresz
Tel. +49 711 970-1404
benjamin.fresz@ipa.fraunhofer.de

Fraunhofer-Institut für Produktions-
technik und Automatisierung IPA
Nobelstraße 12
70569 Stuttgart

www.ipa.fraunhofer.de

Vincent Philipp Göbels
Fraunhofer IAO
Mobil +49 152 22543933
vincent-philipp.goebels@iao.fraunhofer.de

Fraunhofer-Institut für Arbeits-
wirtschaft und Organisation IAO
Bildungscampus 9
74076 Heilbronn

www.iao.fraunhofer.de