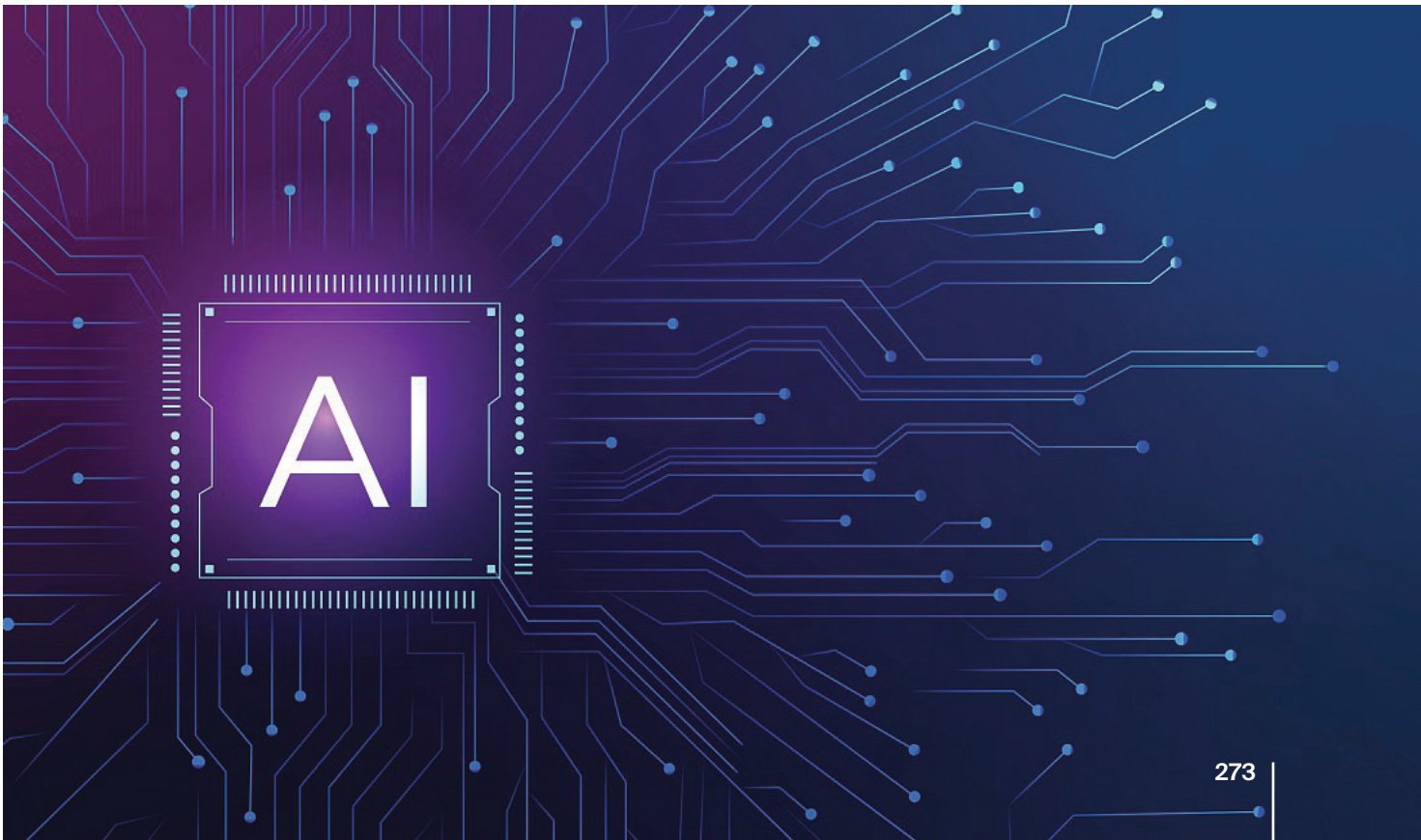
eXplainable Artificial Intelligence – eine Diskussion und Techniken für Aktuarinnen und Aktuare

von Prof. Dr. Anja Schmiedt und Dr. Martin Hüttemann



Der Artikel diskutiert die wachsende Bedeutung von eXplainable Artificial Intelligence in der Versicherungsbranche. Anhand eines eingängigen Beispiels wird illustriert, wie sogenannte Counterfactual Explanations zur Erklärung und damit zur Nachvollziehbarkeit einer Modellentscheidung beitragen können. Wann können ein Modell bzw. eine Modellentscheidung als hinreichend erklärt gelten? Was bedeutet Erklärbarkeit? Welche Modelle bedürfen einer Erklärung? Und warum ist aktuarielle Sorgfalt bei dem Einsatz von Erklärbarkeitstechniken unerlässlich? Der Artikel beinhaltet eine einführende Diskussion und stellt heraus, inwiefern Aktuarinnen und Aktuare gefordert sind sicherzustellen, dass AI nicht nur leistungsstark, sondern auch nachvollziehbar bleibt.

Nehmen Sie hypothetisch an, Sie möchten eine Teilkaskoversicherung für Ihr neues Auto abschließen und haben sich als Obergrenze einen Jahresbeitrag von 750 Euro gesetzt. Sie geben in einen Online-Tarifrechner prämienrelevante Informationen ein: Sie wohnen städtisch in Köln und haben keinen geschützten Stellplatz, Sie sind zwanzig Jahre alt und mit siebzehn Jahren nicht betreut gefahren. Ihnen wird ein jährlicher Beitrag von 1.497 Euro angegeben, sodass Sie sich die Frage stellen: „Warum ist mein Beitrag so hoch?“ Sie rechnen ein paar Varianten und nehmen dazu an, Sie wohnen außerstädtisch in Frechen, haben eine Einzel- oder Doppelgarage, sind mit 17 Jahren betreut gefahren und sind außerdem 40 Jahre alt[1]. Mit diesen hypothetischen Maßnahmen kommen Sie auf einen Jahresbeitrag von 724 Euro (vgl. Abbildung 1).

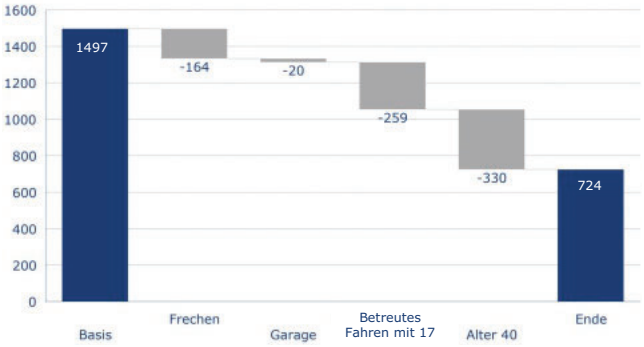
Sie haben hier intuitiv mit kontrafaktischen Erklärungen gearbeitet, die unter der Bezeichnung Counterfactual Explanations als eine Methode von eXplainable Artificial Intelligence (XAI) gelten.

eXplainable AI im europäischen Kontext

Die Integration von künstlicher Intelligenz (KI) und maschinellem Lernen (ML) in die Versicherungsindustrie im Allgemeinen und in aktuarielle Themen im Besonderen birgt Chancen und Risiken zugleich. Eine wesentliche Voraussetzung für die Akzeptanz von komplexen Verfahren ist die Gewährleistung von Transparenz und die damit verbundene Erklärbarkeit der zugrunde liegenden Modelle bzw. der darauf basierenden Entscheidungen.

Grundsätze der Transparenz und Erklärbarkeit von KI bzw. KI-Systemen werden auf europäischer Ebene seit mehreren Jahren im Kontext einer vertrauenswürdigen und ethischen KI diskutiert (vgl. Abbildung 2). Die hochrangige Expertengruppe für künstliche Intelligenz der Europäischen Kommission hat im Jahr 2019 Erklärbarkeit als eines von vier grundlegenden ethischen Prinzipien identifiziert, die für das langfristige Vertrauen in KI-Systeme unerlässlich sind (vgl. HEG-KI, 2019). Darauf aufbauend hat die europäische Versicherungsaufsicht EIOPA Leitlinien für eine ethische und vertrauenswürdige KI für die europäische Versicherungsbranche entwickelt; auch diese Leitlinien betonen das Prinzip der Erklärbarkeit im Zusammenhang mit dem weiter gefassten Begriff der Transparenz (vgl. EIOPA, 2021).

Abb. 1 **Analyse des Jahresbeitrags einer Teilkaskoversicherung mit kontrafaktischen Erklärungen (hypothetisches Beispiel, eigene Darstellung)**



Fußnote

[1] Gegebenenfalls fällt Ihnen auf, dass, wenn Sie heute 40 Jahre alt sind, das begleitete Fahren als Sie 17 Jahre alt waren, noch nicht eingeführt war. Damit haben Sie einen methodischen Nachteil erkannt: Es kann schwierig sein, eine gültige kontrafaktische Situation zu finden, um zu einem gewünschten Modellergebnis zu gelangen.

Abb. 2 Schematische Darstellung zu den (1) Ethik-Leitlinien für eine vertrauenswürdige KI der hochrangigen Expertengruppe für künstliche Intelligenz der Europäischen Kommission (vgl. HEG-KI, 2019) und zu den (2) Grundsätzen für eine ethische und vertrauenswürdige KI für die europäische Versicherungsbranche der EIOPA Consultative Expert Group on Digital Ethics in Insurance (vgl. EIOPA, 2021) (eigene Darstellung)

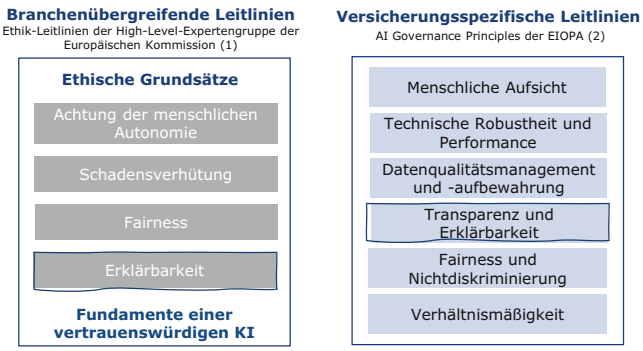
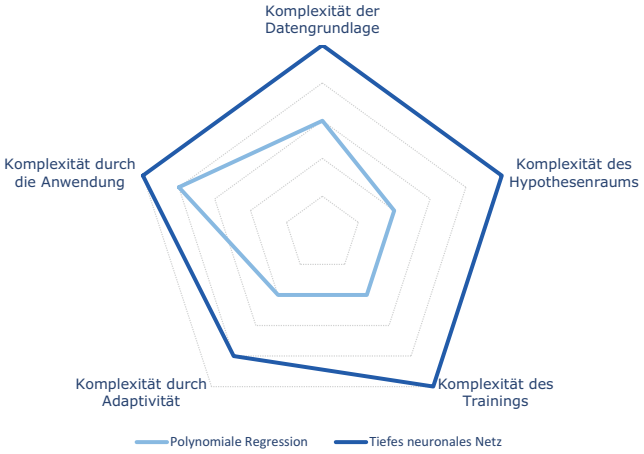


Abb. 3 Schematische Darstellung zu möglichen Charakteristika von Modellkomplexität, hier illustrativ am Beispiel einer polynomialen Regression und eines tiefen neuronalen Netzes (eigene Darstellung)



Darüber hinaus verweist die im Juli 2024 veröffentlichte KI-Verordnung des Europäischen Parlaments und des Rates der Europäischen Union auf die Notwendigkeit, spezifische Transparenzpflichten für bestimmte (Hochrisiko-)KI-Systeme zu etablieren (vgl. AI Act, 2024).

eXplainable AI als Arbeitsgruppe der Deutschen Aktuarvereinigung

Die Arbeitsgruppe eXplainable Artificial Intelligence des Ausschusses Actuarial Data Science hat einen Ergebnisbericht veröffentlicht (vgl. DAV, 2024), der sich intensiv mit den Begrifflichkeiten und Methoden der erklärbaren künstlichen Intelligenz im Versicherungswesen auseinandersetzt. Der Bericht beginnt mit der Definition der Komplexität von Modellen, die als Grundlage für die Diskussion über Erklärbarkeit dient. Dabei werden verschiedene Charakteristika vorgestellt, die zur Erfassung dieser Komplexität beitragen. Im weiteren Verlauf sind unterschiedliche Definitionsansätze und Merkmale von Erklärbarkeit erörtert, die als Basis für die Kategorisierung von Erklärbarkeitsmethoden dienen. Diese Methoden wurden von der Arbeitsgruppe in einer Reihe von Notebooks aufbereitet und u. a. im Kontext aktuarieller Sorgfalt diskutiert. Abschließend thematisiert der Bericht prinzipienbasierte Kriterien, die dabei unterstützen festzulegen, wann bzw. ob ein Modell als ausreichend erklärt gelten kann. Im Folgenden werden zentrale Aspekte des Berichts zusammengestellt.

Artificial Intelligence, Machine Learning und komplexe Modelle

Eine Literaturrecherche zeigt schnell, dass es keine allgemeingültigen Definitionen von künstlicher Intelligenz und deren Teilgebiet des maschinellen Lernens gibt. Genauso wenig existiert eine allgemeingültige begriffliche Abgrenzung von ML-Modellen zu gemeinhin klassischen statistischen Modellen. Daher stellen wir auf die Komplexität von Modellen ab und schlagen einen Charakteristika-Ansatz vor, der eine definitorische Unterscheidung zwischen ML-Modellen und klassischen statistischen Modellen umgeht (vgl. DAV, 2024, S. 7 ff.).

In Anlehnung an einen Ansatz der Bundesanstalt für Finanzdienstleistungsaufsicht (vgl. BaFin, 2021) definieren wir spezifische Charakteristika, die das Kontinuum zwischen maschinellen und klassischen statistischen Modellen anhand ihrer Komplexität kennzeichnen. Der Grad der Komplexität eines Modells kann gemäß diesem Ansatz durch fünf Charakteristika – Datengrundlage, Hypothesenraum, Training, Adaptivität und Anwendung – beschrieben werden. Eine skizzierende Beschreibung der genannten

Abb. 4 Wortwolke (eigene Darstellung)

Charakteristika ist im Kasten rechts zu finden; in Abbildung 3 sind mögliche Ausprägungen exemplarisch für eine polynomiale Regression – als Vertreter weniger komplexer Modelle – und für ein tiefes neuronales Netz – als Vertreter komplexer Modelle – dargestellt.

Beschreibung der Charakteristika, die das Kontinuum zwischen maschinellen und klassischen statistischen Modellen anhand ihrer Komplexität kennzeichnen

Komplexität der Datengrundlage: **Die Verwendung von Big Data oder hochdimensionalen Eingangsdaten ist charakteristisch für ein komplexeres Modell. Ein weniger komplexes Modell verarbeitet tendenziell strukturierte und weniger dimensionale Eingangsdaten mit geringerem Datenvolumen. Die Erklärbarkeit eines Modells kann z. B. bei unstrukturierten oder hochdimensionalen Daten herausfordernder sein.**

Komplexität des Hypothesenraum: **Weniger komplexe Modelle schränken den Hypothesenraum (den Zusammenhang von Eingangsdaten und Modellergebnis) aufgrund von vordefinierten statistischen Assoziationen ein. Komplexere Modelle erlernen eine Problemstruktur hingegen aus den vorhandenen Daten, wobei in der Hypothesenbildung komplexere Muster generiert werden können. Letzteres kann die Erklärbarkeit des Wirkungszusammenhangs zwischen Eingangsdaten und Ergebnis erschweren.**

Komplexität des Trainings: **Der Trainingsprozess von komplexeren Modellen involviert in der Regel eine umfassende Abfolge verschachtelter Rechenvorschriften und setzt iterative Verfahren ein. Des Weiteren werden bei komplexeren Modellen verstärkt Hyperparameter verwendet, die nicht automatisch erlernt werden und gleichzeitig einen wesentlichen Einfluss auf die Modellergebnisse haben können, sodass deren Auswahl bzw. Optimierung einer Erklärung bedarf.**

Komplexität durch Adaptivität: **Weniger komplexe Modelle sind in der Regel statisch und benötigen manuelle Modellanpassungen bzw. Rekalibrierungen, wenn sich Modellanforderungen oder Daten ändern. Komplexere Modelle können hingegen auch in hoher Frequenz an neue Daten angepasst und ihre Leistungsfähigkeit durch inkrementelles Lernen verbessert werden. Durch eine hohe Adaptivität können u. a. Validierbarkeit und Reproduzierbarkeit von Modellergebnissen beeinträchtigt werden.**

Komplexität durch die Anwendung: **Weniger komplexe Modelle werden typischerweise als eigenständige Modelle für wohldefinierte, abgegrenzte Aufgabenstellungen verwendet. Komplexere Modelle bedienen eher ein breiteres Aufgabenspektrum und können in komplexere Modellabhängigkeit eingebettet sein. Die insgesamt komplexeren Anwendungen können die Erklärbarkeit erschweren, wobei gleichzeitig die Anforderungen an Erklärbarkeit von der Anwendung abhängig sind. (Zur Frage, wann ein Modell als hinreichend erklärt gelten kann, siehe S. 278 f.)**

Erklärbarkeit und Erklärbarkeitsmethoden

Mit einer zunehmender Komplexität eines Modells steigen in der Regel auch die Bedeutung von Erklärbarkeit und deren Herausforderungen. Erklärbarkeit ist dabei ein vielschichtiges Konzept, für das es weder in der Literatur noch in der praktischen Anwendung einheitliche Bezeichnungen oder allgemeingültige Definitionen gibt. In Abbildung 4 sind aus der englischsprachigen Literatur verwandte Begriffe aufgeführt, die u. a. den weitergefassten Begriff der Transparenz und den zumeist synonym verwendeten Begriff der Interpretierbarkeit umfassen.

Konsens besteht in der Literatur (zumindest) darüber, dass es herausfordernd ist, Erklärbarkeit gar mathematisch zu definieren oder anderweitig zu formalisieren (vgl. Molnar, 2019). Eine beispielhafte Definition von Erklärbarkeit von der europäischen Versicherungsaufsicht lautet wie folgt (s. EIOPA, 2021, S. 41):

„Explainability is part of the concept of transparency and concerns the ability to explain the output of the AI system to a particular audience, in particular the weight / influence and causal relationship of a specific variable (or group of variables) in the final output."

Eine Erklärung richtet sich folglich an ein bestimmtes Publikum. Den Adressaten einer Erklärung greifen wir im weiteren Verlauf dieses Beitrags als eine Determinante der Antwort auf die Frage auf, wann ein Modell als hinreichend erklärt gelten kann. Zuvor werden verschiedene Charakteristika von Erklärbarkeit bzw. von Erklärbarkeitsmethoden aufgeführt (vgl. Abbildung 5), die sich – nicht zuletzt aufgrund der Ermangelung einer einheitlichen Definition von Erklärbarkeit – herausgebildet haben:

Erklärbarkeit hat einen lokalen oder globalen Geltungsbereich:

- Lokale Erklärbarkeit fokussiert sich darauf, spezifische Modellergebnisse (z. B. einzelne Prognosen) zu begründen und wird daher auch als Entscheidungserklärung bezeichnet.
- Globale Erklärbarkeit bezieht sich hingegen auf die gesamte Logik des Modells bzw. dessen gesamtes Verhalten und wird dah er auch als Modellerklärung bezeichnet.

Erklärbarkeit wird intrinsisch oder post-hoc erreicht:

- Intrinsische Erklärbarkeit wird bereits während der Modellentwicklung angestrebt, indem z. B. die Komplexität des Modells bewusst reduziert wird, sodass das Modell aufgrund einer entsprechend einfachen Struktur als inhärent erklärbar gilt. Man spricht auch von inhärenter oder modellbasierter Erklärbarkeit.

Abb. 5 Schematische Darstellung zu Charakteristika von Erklärbarkeit (eigene Darstellung)

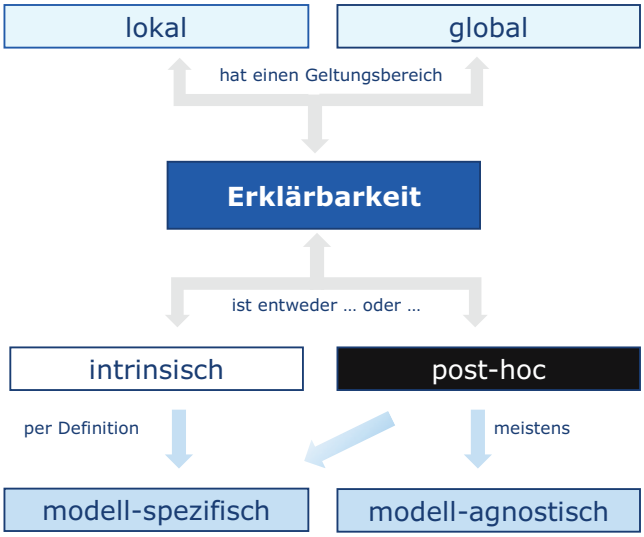
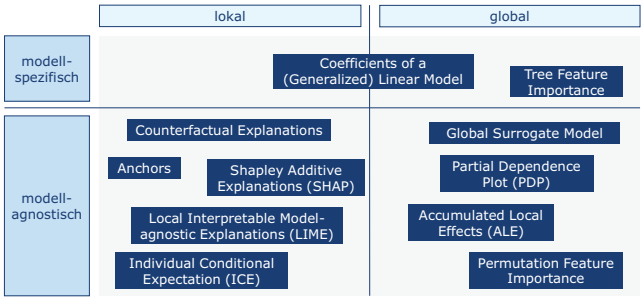


Abb. 6 Einordnung ausgewählter Erklärbarkeitsansätze (eigene Darstellung)



- Methoden der Post-hoc-Erklärbarkeit kommen nach der Entwicklung und Anpassung eines Modells zum Einsatz. Hierbei wird ein zusätzliches Modell bzw. eine zusätzliche Methode verwendet, um die Funktionsweise des originären (komplexen) Modells oder dessen Ergebnisse zu erläutern.

Erklärbarkeitsmethoden können modell-spezifisch oder modell-agnostisch sein:

- Modell-spezifische Ansätze beziehen sich auf bestimmte Modellklassen, wie z. B. neuronale Netze oder baumbasierte Modelle.
- Methoden der modell-agnostischen Erklärbarkeit sind hingegen so konzipiert, dass sie auf eine Vielzahl von Modellen anwendbar sind, unabhängig von deren spezifischer Architektur oder Funktionsweise.

Die Methode **Counterfactual Explanations** aus dem einführenden Beispiel ist in diese Taxonomie wie folgt einzuordnen:

- Counterfactuals sind eine lokale Methode, da ein spezifisches Modellergebnis (der Jahresbeitrag) erklärt wird;
- Counterfactuals finden Post-hoc-Anwendung und sind in diesem Beispiel modell-agnostisch, da sie nur mit den Eingaben (den Risikomerkmalen) und den Ausgaben des Modells (dem Jahresbeitrag) arbeiten und nicht mit der internen Struktur des (Tarifierungs-)Modells.

In Abbildung 6 sind eine Reihe weiterer Erklärbarkeitsansätze den Charakteristika global vs. lokal und modell-spezifisch vs. modell-agnostisch zugeordnet. Zu diesen (und weiteren) Methoden wurden in der Arbeitsgruppe verschiedene Notebooks entwickelt, die über das GitHub-Repository der Arbeitsgruppe zugänglich sind.

In dem Repository finden sich übersichtliche Katalogisierungen der Notebooks sowie der darin diskutierten Erklärbarkeitsmethoden. Die Notebooks wurden mit verschiedenen Zielsetzungen erarbeitet und sind entsprechend in vier Kategorien unterteilt:

- Toy Examples: In dieser Kategorie sind modell-agnostische Erklärbarkeitsmethoden für aktuarielle Regressions- und Klassifikationsprobleme mit überschaubaren Datensätzen einführend beschrieben, praktisch implementiert und kritisch diskutiert.
- Reimplementations: Zielsetzung dieser Kategorie ist es, durch eine Reimplementierung ausgewählter Erklärbarkeitsmethoden (aus oft verwendeten Python-Bibliotheken) die jeweiligen Methoden grundlegend zu verstehen und diese anschließend auf spezifische aktuarielle Beispiele anzuwenden.
- Simulation Study: In einem Notebook dieser Kategorie wird das Verhalten der Methoden Partial Dependence Plots und Accumulated Local Effects, die die Bedeutung von Einflussvariablen erklären, unter Verwendung simulierter Daten analysiert, um Unterschiede – insbesondere in den Voraussetzungen – beider Methoden herauszuarbeiten.
- Use Cases: Betrachtet werden zum einen der vom Ausschuss Actuarial Data Science veröffentlichte Anwendungsfall *Use (this Solvency II) case! Neural Networks Meet Least Squares Monte Carlo* und zum anderen ein von der Society of Actuaries bereitgestellter Datensatz zur Reaktivierung bei Berufsunfähigkeit in den USA. Herausforderungen aufgrund der Komplexität der zugrunde liegenden Datensätze werden herausgearbeitet und – über übliche Erklärbarkeitsmethoden hinaus – für den spezifischen Use Case weitere interessante, nicht standardisierte Methoden dargestellt.

Wann gilt ein Modell als hinreichend erklärt?

Erklärbarkeit komplexer Modelle kann vielschichtig und mitunter selbst komplex sein. Daher liegt es nahe, die Frage zu diskutieren, wann ein Modell als ausreichend erklärbar bzw. erklärt gilt. Die Literatur identifiziert keine konkreten allgemeingültigen Regeln, sondern vielmehr allgemeine Prinzipien, die je nach Einzelfall auszugestalten sind. Beispielhafte Prinzipien zur Erklärbarkeit hat die hochrangige Expertengruppe für künstliche Intelligenz der Europäischen Kommission vorgeschlagen (vgl. HEG-KI, 2019):

- Vom KI-System getroffene Entscheidungen können vom Menschen verstanden und rückverfolgt werden.
- Beeinflusst das KI-System Menschenleben, muss eine geeignete Erklärung des Entscheidungsprozesses rechtzeitig und auf die Sachkenntnisse des jeweiligen Interessententrägers angepasst, erhältlich sein.
- Es müssen Erläuterungen darüber vorliegen, inwieweit ein KI-System die Entscheidungsprozesse einer Organisation beeinflusst und gestaltet, sowie Entwurfsentscheidungen und Gründe für die Einführung.

Zur Konkretisierung dieser Prinzipien hat die hochrangige Expertengruppe eine Bewertungsliste entwickelt (vgl. HEG-KI, 2019). Da diese nicht ausschließlich auf die Versicherungswirtschaft beschränkt ist, empfiehlt es sich, die in der Bewertungsliste formulierten Fragen im aktuariellen Kontext zu verfeinern und in der praktischen Anwendung zu erproben. Auf Basis von Veröffentlichungen der Bank of England (2019) und der EIOPA (2021) wurde daher im Ergebnisbericht der Arbeitsgruppe (vgl. DAV, 2024, S. 22 ff.) eine auflistende Darstellung dazu erarbeitet, welche Informationen und Fragen bereitgestellt bzw. beantwortet werden sollten, um Erklärbarkeit und Transparenz zu erzielen. Die Auflistung unterscheidet verschiedene interne und externe Adressaten einer Erklärung. Die Unterscheidung ist sinnvoll, da Adressaten – neben dem konkreten Anwendungsfall – den erforderlichen Grad an Erklärbarkeit determinieren.

Der erforderliche Grad an Erklärbarkeit wird zum einen durch den Adressaten bestimmt:
Es gibt eine Vielzahl von Stakeholdern, die eine Erklärbarkeit komplexer Modelle erwarten können. Innerhalb des Unternehmens zählen dazu z. B. Modellierungsteams, Validierungsteams, interne Prüferinnen und Prüfer sowie das Management und der Vorstand. Extern sind es z. B. die Endkundinnen und Endkunden, externe Prüferinnen und Prüfer, Aufsichtsbehörden und die Wissenschaft.

Dabei hat jeder Adressat eine individuelle Interessenslage hinsichtlich der Erklärbarkeit komplexer Modelle. Beispielsweise haben Endkunden in der Regel einen **Mikroblick**: Sie sind u. a. daran interessiert, welche spezifischen Merkmale die einzelnen Modellergebnisse beeinflussen, wie z. B. bei der Tarifierung in dem einführenden Beispiel zu kontrafaktischen Erklärungen. Der Vorstand oder die Aufsicht haben hingegen einen **Makroblick**: Sie benötigen u. a. globale Informationen über das gesamte Modellverhalten, um z. B. potenzielle Risiken, die sich aus den Modellen für die Solvenzlage des Versicherungsunternehmens ergeben, zu verstehen.

Literatur

Dieser Artikel referenziert eine Auswahl an Quellen. Eine umfassende Literaturschau sowie die entsprechenden Quellen sind im Ergebnisbericht der Arbeitsgruppe (vgl. DAV, 2024) verfügbar.

Bank of England (2019). Staff Working Paper No. 816 Machine learning explainability in finance: an application to default risk analysis. Online verfügbar unter https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis

Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin, 2021). Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte – Konsultationspapier. Online verfügbar unter https://www.bafin.de/SharedDocs/Veroeffentlichungen/DE/Meldung/2021/meldung_2021_07_15_Konsultation_Maschinelles_Lernen.html

Deutsche Aktuarvereinigung e. V. (DAV, 2024). Explainable Artificial Intelligence: ein aktueller Überblick für Aktuarinnen und Aktuare. Online verfügbar unter https://aktuar.de/unsere-themen/fachgrundsaetze-oeffentlich/2024-05-27_Ergebnisbericht_DAV_AG_XAI.pdf

EIOPA Consultative Expert Group on Digital Ethics in Insurance (EIOPA, 2021). Artificial intelligence governance principles: towards ethical and trustworthy artificial intel-ligence in the European insurance sector. Online verfügbar unter https://www.eiopa.europa.eu/publications/artificial-intelligence-governance-principles-towards-ethical-and-trustworthy-artificial_en

Europäisches Parlament und Rat der Europäischen Union (AI Act, 2024). Verordnung (EU) 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz. Online verfügbar unter https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_DE.pdf

Hochrangige Expertengruppe für künstliche Intelligenz eingesetzt von der Europäischen Kommission (HEG-KI, 2019). ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜR-DIGE KI. Online verfügbar unter https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Molnar, C. (2020). Interpretable machine learning – a guide for making black box models explainable. Online verfügbar unter https://christophm.github.io/interpretable-ml-book/index.html

Der erforderliche Grad an Erklärbarkeit wird zum anderen durch den Anwendungsfall bestimmt:
Je höher das Risiko ist, das mit der Anwendung eines Modells verbunden ist, desto größer ist in der Regel auch der erforderliche Grad an Erklärbarkeit. Beispielhafte Leitfragen zur Bewertung des Risikos einer Anwendung sind:

- Welche Auswirkungen hat die Anwendung auf die Kundinnen und Kunden?
- Trifft die Anwendung direkte Entscheidungen oder fungiert sie lediglich als Entscheidungshilfe?
- Hätte der Ausfall der Anwendungen Auswirkungen auf das Tagesgeschäft und bestünde ein Reputationsschaden?

Der erforderliche Grad an Erklärbarkeit am Beispiel der Tarifierung:
Dass in der Tarifierung sensible Daten zum Einsatz kommen und dass die auf Basis der Modelle abgeleiteten Entscheidungen eine große Tragweite haben, spiegelt sich in dem Anforderungsgrad an Erklärbarkeit wider. So erwarten Vorstand und Aufsicht weitreichende Erklärbarkeit, v. a. zur Modellarchitektur sowie den dazugehörigen Methoden und Annahmen; gegenüber Kundinnen und Kunden ist es notwendig, die Versicherbarkeit, einen Ausschluss von Risiken oder die individuelle Preisgestaltung zu erklären.

Die Rolle der Aktuarinnen und Aktuare

Es gibt eine Reihe von Erklärbarkeitsmethoden, die zum Verständnis der Funktionsweise bzw. der Ergebnisse von komplexen Modellen beitragen können. Einige Post-hoc-Methoden haben sich als Standard etabliert, da sie zunächst leicht verständlich und in Softwarepaketen bzw. -bibliotheken implementiert sind. Im Sinne einer aktuariellen Sorgfalt kann es jedoch nicht ausreichen, Standardmethoden ohne ein konkretes Verständnis von Methode und Anwendung und ohne Hinterfragen der Ergebnisse anzuwenden (vgl. auch Fußnote [1]). Im Ergebnisbericht der Arbeitsgruppe (vgl. DAV, 2024, S. 19 ff.) sind daher mögliche Prinzipien für die sorgfältige und kritische Anwendung speziell von Erklärbarkeitsmethoden formuliert, die im idealen aktuariellen Anwendungsprozess berücksichtigt werden sollten. Dazu zählen die Beantwortungen von Fragen zum Ziel der Erklärung, zur Dokumentation und Implementierung der Erklärbarkeitsmethode, zu den statistischen Voraussetzungen der Methode, zu deren Vor- und Nachteilen usw.

Aktuarinnen und Aktuare spielen eine entscheidende Rolle im Bereich der Erklärbarkeit komplexer Modelle bei aktuariellen bzw. versicherungsspezifischen Anwendungsfällen. Durch ihre Ausbildung und Erfahrung – u. a. im Umgang mit Daten und statistischen Modellen – können sie komplexe Modelle sowie Erklärbarkeitsmethoden verstehen, anwenden und einordnen. Dabei verstehen Aktuarinnen und Aktuare zudem das Fachgebiet profund.

Über die Autorin und den Autor

→ Prof. Dr. Anja Schmiedt
ist Professorin für Mathematik mit den Schwerpunkten Aktuarwissenschaften und Statistik an der OTH Regensburg. Vor ihrer Erstberufung im Oktober 2021 war sie als Aktuarin DAV in leitenden Funktionen in der Rückversicherung und aktuariellen Beratung tätig. Sie promovierte an der RWTH Aachen auf einem Gebiet der mathematischen Statistik. Für die Deutsche Aktuarvereinigung engagiert sie sich nebenberuflich in der Ausbildung angehender Aktuare und ehrenamtlich in verschiedenen Arbeitskreisen und Funktionen. So ist sie beispielsweise Mitglied im Ausschuss Actuarial Data Science, für den sie die Arbeitsgruppe Erklärbare künstliche Intelligenz leitete, und Co-Leiterin der gleichnamigen Fachgruppe.

→ Dr. Martin Hüttemann
ist seit 2018 in der Abteilung „Quantitative Risikomodellierung" der Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin) tätig. Dort prüft er interne Modelle von Versicherungsunternehmen mit einem Schwerpunkt auf versicherungstechnischen Risiken im Bereich Non-Life sowie operationellen Risiken. Zudem beschäftigt er sich unter anderem mit Themen rund um maschinelles Lernen und künstliche Intelligenz.
Dr. Hüttemann hat einen Master in Wirtschaftsmathematik an der Universität zu Köln und einen Master in Financial Computing am University College London erworben. Seine Promotion erfolgte ebenfalls an der Universität zu Köln. Er ist Mitglied der DAV und erwarb im Jahr 2023 die Zusatzqualifikation als Certified Actuarial Data Scientist.