erklärbare künstliche Intelligenz (XAI)

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Erklärbare künstliche Intelligenz (XAI) ist das Teilgebiet der Künstliche Intelligenz (KI), das sich damit befasst, die Vorhersagen von maschinelles Lernen (ML)-Methoden für Menschen verständlich zu machen. Vieles davon lässt sich als ein Problem der Approximation einer Funktion formulieren: Eine gelernte Hypothese wird durch eine einfachere Funktion erklärt, die ein Mensch nachvollziehen kann. Erklärungen unterscheiden sich darin, woraus sie bestehen. Eine Art gibt einen Wert für jedes Merkmal an, die Gewichte einer nahe dem Datenpunkt angepassten lineare Abbildung (lokale interpretierbare modellunabhängige Erklärungen (LIME)) oder additive Beiträge zur Vorhersage (SHAP (SHAP)). Eine Kontrafaktum benennt stattdessen einen anderen Merkmalsvektor, nämlich den nächstgelegenen, dessen Vorhersage abweicht. Ein Konzeptaktivierungsvektor (CAV) drückt die Erklärung in einem vom Nutzer benannten Konzept aus, das in den Aktivierungen innerhalb der trainierten Methode verortet ist. Alle drei Verfahren erfolgen post hoc und behandeln die gelernte Hypothese als feststehend. Die Alternative ist eine Hypothese, die keiner separaten Erklärung bedarf, was durch die Einschränkung des Hypothesenraums auf für Menschen nachvollziehbare Hypothesen oder durch eine Regularisierung, die solche Hypothesen bevorzugt, erreicht wird. Erklärbare empirische Risikominimierung (EERM) behält stattdessen den Hypothesenraum bei und bestraft Abweichungen von den Vorhersagen eines bestimmten Nutzers.

Definition

Das KI-System einer Bank lehnt einen Kreditantrag ab, und der Antragsteller fragt nach dem Grund. XAI ist das Teilgebiet der Künstliche Intelligenz (KI), das sich damit befasst, die Vorhersagen von maschinelles Lernen (ML)-Methoden für Menschen verständlich zu machen. Ziel ist es, jede Vorhersage durch eine Erklärung darüber zu ergänzen, wie sie zustande gekommen ist. Verwendet die ML-Methode einen hinreichend einfachen Hypothesenraum, kann die gelernte Hypothese inhärent interpretierbar sein, sodass keine separate Erklärung erforderlich ist (Rudin, 2019); siehe interpretierbares maschinelles Lernen, das weiter unten behandelt wird.

Der Begriff XAI wurde durch ein Programm der US-amerikanischen Defense Advanced Research Projects Agency bekannt gemacht (Gunning and Aha,

2019); im Kontext des MLs wird auch der synonyme Begriff erklärbares maschinelles Lernen (erklärbares ML) verwendet. Die zugrunde liegende Eigenschaft ist die Erklärbarkeit, welche die internationale Terminologienorm für KI für KI-Systeme im Allgemeinen definiert (International Organization for Standardization and International Electrotechnical Commission, 2022, Abschn. 3.5.7).

Ein Großteil von XAI lässt sich als ein Problem der Approximation einer Funktion formulieren. Eine trainierte Hypothese $\widehat{h} : \mathcal{X} \to \mathcal{Y}$, die z. B. von einem intransparenten tiefes Netz geliefert wird, ist typischerweise eine Verkettung vieler nichtlinearer Schichten, also eine Funktion auf einem hochdimensionalen Merkmalsraum $\mathcal{X}$ (Goodfellow et al., 2016, Kap. 6). Die Erklärung von $\widehat{h}$ an einem Datenpunkt mit dem Merkmalsvektor $\mathbf{x} \in \mathcal{X}$ bedeutet, eine Frage über diese Funktion in einer für Menschen nachvollziehbaren Form zu beantworten. Die nachfolgenden Beispiele für Erklärungen lassen sich in drei Arten unterteilen, je nachdem, woraus die Erklärung besteht: Die erste approximiert $\widehat{h}$ durch eine einfachere Funktion, die zweite fragt, wo $\widehat{h}$ einen anderen Wert annimmt, und die dritte bezieht sich auf ein Konzept innerhalb der trainierten Methode.

Die erste Art ist ein Wert für jedes Merkmal des Datenpunkts. Die Werte sind die Gewichte einer einfacheren Hypothese, wie etwa einer lineare Abbildung, die in der Nähe von $\mathbf{x}$ mit $\widehat{h}$ übereinstimmt und aus einem Hypothesenraum stammt, dessen Hypothesen ein Nutzer ablesen kann, so wie lokale interpretierbare modellunabhängige Erklärungen (LIME) sie berechnet. Die zweite Art ist ein anderer Merkmalsvektor, dessen Vorhersage abweicht. Die dritte Art ist ein vom Nutzer benanntes Konzept, das innerhalb der trainierten Methode statt unter den Merkmale des Datenpunkts verortet ist. Alle drei erfolgen post hoc: Sie behandeln die gelernte Hypothese als feststehend und konstruieren die Erklärung nach dem Training.

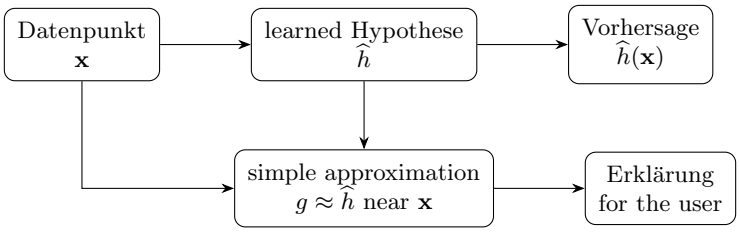


Abbildung 1: Die erste Art von Erklärung. Die gelernte Hypothese $\widehat{h}$ liefert die Vorhersage $\widehat{h}(\mathbf{x})$ für einen Datenpunkt mit dem Merkmalsvektor $\mathbf{x}$. Die Erklärung wird aus einer einfachen Hypothese g, z. B. einer lineare Abbildung, konstruiert, die $\widehat{h}$ nahe $\mathbf{x}$ approximiert; der für ein Merkmal angegebene Wert ist das Gewicht, das g ihm zuweist (vgl. LIME)

Die erste Art von Erklärung gibt einen Wert pro Merkmal an, der besagt, wie stark dieses Merkmal zur Vorhersage beigetragen hat. LIME ermittelt die Werte als die Gewichte einer nahe $\mathbf{x}$ an $\widehat{h}$ angepassten lineare Abbildung (Ribeiro et al., 2016); Abb. 1 veranschaulicht diese Konstruktion. SHAP (SHAP) ermittelt

sie durch Zerlegung der Vorhersage in einen Basiswert und einen Beitrag pro Merkmal, wobei die Beiträge Shapley-Werte sind (Lundberg and Lee, 2017; Molnar, 2025). Beide liefern dieselbe Art von Erklärung und unterscheiden sich darin, wie deren Werte berechnet werden. Sind die Merkmale Eigenschaften von Bildpixeln, bilden die Werte der Merkmale ein zweites Bild derselben Größe: ein Relevanzwert pro Pixel (siehe Abb. 2). Die Klassenaktivierungskarte (CAM) ist eine Konstruktionsweise eines solchen pixelweisen Wertes.

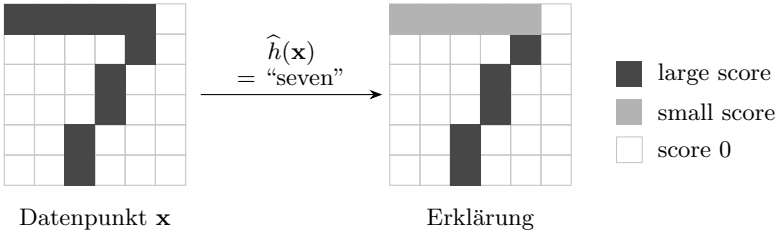


Abbildung 2: Eine Erklärung erster Art für ein Bild, ein Wert pro Merkmal, wie ihn eine CAM liefert. Die Objekte sind wie in Abb. 1 benannt. Links: ein Datenpunkt $\mathbf{x}$, ein Bild auf einem Gitter aus 6×6 Pixeln, für das die gelernte Hypothese die Vorhersage $\widehat{h}(\mathbf{x}) = \textit{sieben}$ liefert. Rechts: die Erklärung dieser Vorhersage, ein Relevanzwert pro Pixel. Die Pixel des diagonalen Strichs tragen die größten Werte, der obere Balken kleinere, und die übrigen Pixel haben den Wert null

Die zweite Art von Erklärung ist eine Kontrafaktum. Sie gibt keine Werte an und approximiert nichts. Stattdessen fragt sie, wo $\widehat{h}$ einen anderen Wert annimmt, und zeigt auf, wie die Merkmale verändert werden müssen, damit die Vorhersage $\widehat{h}(\mathbf{x})$ anders ausfällt. Eine Kontrafaktum identifiziert die kleinste derartige Veränderung, wobei die Nähe anhand einer gewählten Metrik gemessen wird (Wachter et al., 2018) (siehe Abb. 3).

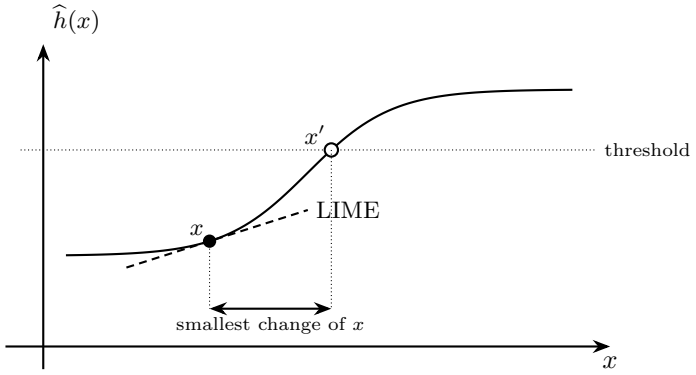


Abbildung 3: Die ersten beiden Arten von Erklärungen für eine Vorhersage. Die trainierte Hypothese $\widehat{h}$ (durchgezogene Kurve) liefert eine Vorhersage für jeden Wert des Merkmals. An einem gegebenen Datenpunkt mit dem Merkmal-Wert x (ausgefüllter Kreis) konstruiert LIME (gestrichelte Linie) eine lokale lineare Approximation. Eine Kontrafaktum identifiziert bezüglich einer gewählten Metrik den nächstgelegenen Punkt x' (offener Kreis), an dem die Vorhersage eine Entscheidungsschwelle überschreitet. Der Doppelpfeil markiert die Änderung des Merkmals, die sie angibt

Für den abgelehnten Kreditantragsteller gibt LIME beispielsweise an, welchem der Merkmale des Antragstellers, wie Einkommen und Kreditwürdigkeit, seine lokale lineare Approximation bei der Ablehnung das größte Gewicht zuweist. Eine Kontrafaktum nennt stattdessen die kleinste Veränderung dieser Merkmale, welche die Vorhersage in eine Bewilligung umwandeln würde.

Die dritte Art von Erklärung bezieht sich nicht auf die Merkmale des Datenpunkts, sondern auf ein vom Nutzer spezifiziertes Konzept. Ein solches Konzept wird durch einen vom Nutzer bereitgestellten Datensatz von Beispielen definiert. Ein Konzeptaktivierungsvektor (CAV) verortet ein solches Konzept als eine Richtung im Raum der Aktivierungen einer verdeckten Schicht eines künstliches neuronales Netz (KNN): den Normalenvektor einer Entscheidungsgrenze, welche die bereitgestellten Beispiele, die das Konzept aufweisen, von jenen trennt, bei denen dies nicht der Fall ist (Kim et al., 2018).

Im Gegensatz zu LIME und einer Kontrafaktum als Erklärungen benötigt ein CAV Zugriff auf die internen Berechnungen eines ML-System (ML-System)s. Eine solche interne Berechnung ist die Aktivierung $\mathbf{z} = f(\mathbf{x})$ einer verdeckten Schicht in einem tiefes Netz. Ein tiefes Netz berechnet $\widehat{h}$, indem es die Aktivierungen $\mathbf{z}$ an die letzten Schichten weitergibt, die für jedes Label einen Wert $s(\mathbf{z})$ liefern; die Vorhersage $\widehat{h}(\mathbf{x})$ folgt aus diesen Werten (siehe CAV). In einem Bild-Klassifikator stellt beispielsweise ein Nutzer, der vermutet, dass Fotografien über das Konzept *Streifen* mit dem Label Zebra versehen werden, Fotografien mit Streifen und Fotografien ohne Streifen bereit. Der CAV für *Streifen* ist dann der Normalenvektor des linearer Klassifikators, der die Aktivierungen $\mathbf{z}$ der beiden Mengen von Fotografien trennt. Wie stark das Konzept zur Vorhersage einer

Zebrafotografie beigetragen hat, wird dann durch die Richtungs-Ableitung des Wertes s entlang dieses CAVs gemessen.

Die oben genannten XAI-Methoden konstruieren eine Erklärung nach dem Training für eine gelernte Hypothese, die bereits feststeht. Die Alternative besteht darin, eine zu lernen, die keiner separaten Erklärung bedarf. Interpretierbares maschinelles Lernen erreicht dies, indem es nur Hypothesen zulässt, die ein Mensch direkt nachvollziehen kann. Die Einschränkung erfolgt durch die manuelle Wahl eines hinreichend einfachen Hypothesenraums oder durch eine Regularisierung, die Hypothesen mit bestimmten Eigenschaften wie dünner Besetztheit oder Vorhersagbarkeit bevorzugt (Tibshirani, 1996; Zhang et al., 2024). Erklärbare empirische Risikominimierung (EERM) nutzt zwar Regularisierung, jedoch nicht zu diesem Zweck: Es belässt den Hypothesenraum unverändert und liefert Erklärbarkeit für einen bestimmten Nutzer. Dieser Nutzer stellt seine eigenen Vorhersagen für die Datenpunkte eines Trainingsdatensatzs bereit, und der Strafterm misst die Abweichung der Vorhersagen der gelernten Hypothese von den Vorhersagen des Nutzers (Zhang et al., 2024).

Drei Aspekte einer Erklärung werden untersucht: wem sie dient, wie viel von der gelernten Hypothese sie abdeckt und ob sie originalgetreu ist. Wie viel eine Erklärung leistet, hängt davon ab, wer sie erhält: Erklärbarkeit wird relativ zu einem bestimmten Nutzer gemessen (Colin et al., 2022; Jung and Nardelli, 2020). Erklärungen können lokal sein und eine einzelne Vorhersage betreffen, oder global, indem sie die gelernte Hypothese als Ganzes charakterisieren (Molnar, 2025).

Eine Erklärung muss originalgetreu (engl. *faithful*) sein, d. h. die Berechnung widerspiegeln, welche die gelernte Hypothese ausführt. Wenn die Erklärung die Merkmale bewertet, lässt sich dies testen, statt es nur vorauszusetzen. Eine CAM ist für eine Vorhersage originalgetreu, wenn das Invertieren der von ihr am höchsten bewerteten Pixel diese Vorhersage häufiger ändert als das Invertieren ebenso vieler der von ihr am niedrigsten bewerteten Pixel. Das Stören der am höchsten bewerteten Regionen zuerst und das Erfassen, wie schnell der vorhergesagte Klassenwert fällt, ist die Standardform dieses Tests (Samek et al., 2017, Abschn. III-A, S. 2666).

Für das Bild aus Abb. 2 und einen an Bilder dieser Art angepassten linearer Klassifikator ändert das Invertieren der drei am höchsten bewerteten Pixel die Vorhersage, während das Invertieren der Pixel in der umgekehrten Reihenfolge sie über alle 36 Pixel hinweg unverändert lässt. Die durch die Karte vorgegebene Reihenfolge ändert die Vorhersage auch bei jeder von 200 verrauschten Varianten des Bildes früher. Die Originaltreue begrenzt zudem, was eine Post-hoc-Erklärung leisten kann: Eine, die überall mit der gelernten Hypothese übereinstimmen würde, wäre diese Hypothese selbst, sodass eine einfachere Erklärung irgendwo von ihr abweicht (Rudin, 2019).

Zwei Rechtsinstrumente formulieren ausdrücklich Anforderungen an XAI. Gemäß der Datenschutz-Grundverordnung (DSGVO) hat eine Person, die einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung unterworfen ist, Anspruch auf „aussagekräftige Informationen über die involvierte Logik“ dieser Entscheidung (European Parliament and Council of

the European Union, 2016, Art. 15 Abs. 1 Buchst. h). Die Mitteilung des Algorithmus selbst stellt keine hinreichend präzise und verständliche Erklärung dar (Court of Justice of the European Union, 2025). Eine Kontrafaktum als Information kann geeignet sein: das Ausmaß, in dem eine Abweichung bei den personenbezogene Daten zu einem anderen Ergebnis geführt hätte. Das zweite Instrument ist der KI-Verordnung, der für ein Hochrisiko-KI-System verlangt, dass betroffene Personen „klare und aussagekräftige Erläuterungen zur Rolle des KI-Systems im Entscheidungsverfahren“ erhalten (European Parliament and Council of the European Union, 2024, Art. 86) (siehe Recht auf Erklärung).

Synonyme: erklärbares maschinelles Lernen (erklärbares ML).

Siehe auch: Erklärbarkeit, Interpretierbarkeit, Erklärung, interpretierbares maschinelles Lernen, EERM, LIME, SHAP, Kontrafaktum, Merkmal, mechanistische Interpretierbarkeit, CAM, Recht auf Erklärung.

Querverweise

Literatur

1. Court of Justice of the European Union (2025). Judgment of 27 February 2025, CK v Dun & Bradstreet Austria GmbH and Magistrat der Stadt Wien, C-203/22, ECLI:EU:C:2025:117. Court of Justice of the European Union, First Chamber. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62022CJ0203.

2. Colin et al. (2022). What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. In: Adv. Neural Inf. Process. Syst., pp. 2832–2845.

3. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

4. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

5. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

6. Gunning and Aha (2019). DARPA's Explainable Artificial Intelligence Program. AI Magazine, 40(2), pp. 44–58.

7. International Organization for Standardization and International Electrotechnical Commission (2022). ISO/IEC 22989:2022 — Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. International Standard, 1st edition. URL: https://www.iso.org/standard/74296.html.

8. Jung and Nardelli (2020). An Information-Theoretic Approach to Personalized Explainable Machine Learning. IEEE Signal Process. Lett., 27, pp. 825–829.

9. Lundberg and Lee (2017). A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774.

10. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.

11. Ribeiro et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1135–1144.

12. Samek et al. (2017). Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Trans. Neural Netw. Learn. Syst., 28(11), pp. 2660–2673.

13. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. J. Roy. Statist. Soc.: Ser. B (Methodological), 58(1), pp. 267–288.

14. Wachter et al. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), pp. 841–887.

15. Zhang et al. (2024). Explainable empirical risk minimization. Neural Comput. Appl., 36(8), pp. 3983–3996.

16. Kim et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Mach. Learn., pp. 2668–2677.

17. Rudin (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Mach. Intell., 1(5), pp. 206–215.