 https://doi.org/10.33196/juridikum202402024401

# Spotlight Erklärbare KI

## Eine Besprechung ausgewählter Use Cases aus rechtlicher und technologischer Perspektive

Elisabeth Paar / Timothée Schmude / Cansu Cinar

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

### 1. Einleitung

Erklärungen spielen eine zentrale Rolle in der zwischenmenschlichen Kommunikation. So intuitiv dies auf den ersten Blick erscheint, so wenig ist damit gesagt, was dies im Einzelfall tatsächlich bedeutet. Art, Ausmaß und Funktion einer Erklärung variieren nämlich je nach den Umständen der konkreten Situation. Ein Bereich, in dem das Verständnis von und die Anforderungen an Erklärungen sowie die Umsetzung ebendieser aktuell besonders intensiv diskutiert werden, ist jener der künstlichen Intelligenz (KI). Dabei steht die Frage im Zentrum, ob – und wenn ja, auf welche Weise – die Funktionsweisen und Entscheidungen von KI-Systemen erklärbar sind und inwieweit sie die allfälligen rechtlichen Anforderungen, wie sie an unterschiedliche Formen von Erklärungen gestellt werden, abzubilden vermögen.
In diesem Beitrag möchten wir uns diesem Thema der erklärbaren KI (Explainable AI, XAI) widmen und dabei sowohl die rechtlichen als auch technologischen Dimensionen zweier ausgewählter Use Cases beleuchten. Während dieser Beitrag keine umfassende Besprechung von XAI leisten kann, da diese einen größeren Umfang erfordern würde, werden die zentralen Fragestellungen anhand der zwei Anwendungsfälle illustriert. Zu diesem Zweck werden sowohl für White Box Modelle als auch für Black Box Modelle konkrete Use Cases aus technologischer Perspektive vorgestellt. Der erste Use Case – das schwedische Trelleborg-Modell zur automatisierten Verarbeitung von Anträgen auf Sozialleistung – dient als Beispiel des Einsatzes sogenannter White Box Modelle in der öffentlichen Verwaltung. Der zweite Use Case – ein Modell zur Vorhersage des Risikos bei herzchirurgischen Eingriffen – stellt demgegenüber die Erklärbarkeit von Black Box Modellen in Hochrisiko-Anwendungen in den Vordergrund. Im Lichte ihrer jeweiligen Anwendungsbereiche wird daran anknüpfend aus rechtlicher Perspektive skizzenhaft aufgezeigt, welche Anforderungen an die Erklärung von KI-Systemen jeweils gestellt werden und inwieweit diese realisierbar sind.

### 2. Erklärbare KI: Ein Überblick

KI ist ein heterogener Sammelbegriff und umfasst eine Vielzahl algorithmischer Modelle, die zumeist durch das Training auf bestehenden Daten Vorhersagen, Empfehlungen

© Verlag Österreich 2024

BiblioScout

oder Entscheidungen über neue Datenpunkte treffen können.[1] XAI hat zum Ziel, die Prozesse sowie Entscheidungen von KI-Systemen transparent und nachvollziehbar zu gestalten.[2]

Aktuelle Taxonomien ordnen Erklärungsansätze nach verschiedenen Eigenschaften ein.[3] Die wichtigsten dieser Eigenschaften orientieren sich im Wesentlichen an drei Polen: Der erste betrifft die Unterscheidung nach dem Zeitpunkt, zu dem eine Erklärung gegeben wird, und damit jene zwischen Post-Hoc und Ante-Hoc: Während *Post-Hoc* Erklärungen das Modell anhand bereits getroffener Entscheidungen erschließen, haben *Ante-Hoc* Erklärungen das Ziel, ein Modell unabhängig von konkreten Entscheidungen verständlich zu machen.[4] Der zweite Pol nimmt auf die Unterscheidung nach der Anwendbarkeit von Erklärungen und damit auf jene zwischen modellspezifisch und modellagnostisch Bezug: Während modellspezifische Erklärungen Ansätze verwenden, die nur für einen bestimmten Modelltyp geeignet sind, setzen modellagnostische Erklärungen an jenen Mechanismen an, die dem Großteil der KI-Modelle zugrunde liegen. Der dritte Pol bezieht sich schließlich auf die Unterscheidung nach dem Umfang der Erklärungen und damit auf jene zwischen global und lokal: Während eine lokale Erklärung bloß eine einzelne Entscheidung in den Blick nimmt, behandelt eine globale Erklärung das gesamte Modell.

Im Folgenden werden die verschiedenen Dimensionen von Erklärungen für KI-Systeme anhand einer Unterteilung in White Box Modelle und Black Box Modelle beleuchtet. Während White Box Modelle als interpretierbar gelten, da ihre Funktionsweisen prinzipiell einsehbar sind, verbergen Black Box Modelle ihre Funktionsweisen und gelten deshalb als intransparent. Daraus resultieren wiederum maßgebliche Unterschiede in der rechtlichen Betrachtung des Einsatzes der jeweiligen Modellkategorien.

## 3. White Box Modelle

KI-Modelle, die einen Blick in die internen Vorgänge des Algorithmus zulassen und dabei unmittelbar verständlich, also inhärent interpretierbar, sind, werden als White Box Modelle bezeichnet.[5] Die Interpretierbarkeit dieser Modelle kann durch formale Attribute beschrieben werden. Demnach ist ein Modell grds interpretierbar, wenn es einen linearen Zusammenhang zwischen Input und Output abbildet (Linearität), eine Veränderung im Input immer entweder eine Zunahme oder Abnahme im Output bedeutet

1 *European Commission (Joint Research Centre)*, AI Watch, Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU (2020) 43.

2 *Ali et al*, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion 2023, 1 (2).

3 *Speith*, A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods, ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 21.-24. Juni 2022, 2239 (2240 ff).

4 Diese Begriffstrennung überschneidet sich mit der sogleich noch näher auszuführenden Unterscheidung in intransparente Black Box Modelle (Post-Hoc) und transparente White Box Modelle (eher Ante-Hoc).

5 *Cheng et al,* Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders, Proceedings of the ACM Conference on Human Factors in Computing Systems, 4.-9. Mai 2019, 1.

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

(Monotonie) und die Interaktion zwischen einzelnen Inputfeatures auf ein Minimum beschränkt wird.[6] Allerdings kann die Zahl und Verständlichkeit der Inputfeatures auch die Interpretierbarkeit eines grds transparenten Modells einschränken.[7]

### 3.1. Beispiel: Entscheidungsbäume

Entscheidungsbäume beschreiben eine Gruppe von Algorithmen, die auf der Grundlage eines Datensatzes baumartige Strukturen erstellen, mithilfe derer sich neue Datenpunkte klassifizieren bzw vorhersagen lassen. Durch die grundsätzlich leicht verständliche Baumstruktur gelten sie damit als typische Beispiele interpretierbarer Modelle. Algorithmen, die Entscheidungsbäume konstruieren, unterteilen einen Datensatz anhand bestimmter Schwellenwerte der Features in kleinere Teilsets. Dieser Prozess wird wiederholt, bis jeder Pfad des Baumes zu einer eindeutigen Klassifikation führt oder sämtliche Features für die Unterteilungen verwendet wurden.[8] Entscheidungsbäume können sowohl zur Klassifizierung (diskrete Werte) als auch zur Regression (kontinuierliche Werte) verwendet werden. Im Gegensatz zu Modellen wie etwa der linearen Regression haben Entscheidungsbäume dabei den Vorteil, auch nicht-lineare Beziehungen zwischen Inputfeatures und Output darstellen zu können.[9] Durch die Erstellung des Baumdiagramms gelten diese Modelle zudem als leicht interpretierbar, da die Klassifizierung eines Datenpunkts durch das Durchlaufen des Baums leicht nachvollzogen werden kann. Wenn allerdings zu viele Blätter (Zielklassen) oder Knotenpunkte (Schwellenwerte) existieren, kann dies die Interpretierbarkeit schnell verringern.[10]

### 3.2. Use Case I: Trelleborg-Modell

Die schwedische Gemeinde Trelleborg setzt seit 2017 vollautomatisierte Entscheidungen für die Vergabe von Sozialleistungen ein. Das algorithmische Modell, Trelleborg-Modell genannt, basiert auf einem Entscheidungsbaum, der bestimmte Variablen eines Antrags mit Datenbanken anderer Behörden, etwa der Steuerbehörde oder der Krankenversicherung, abgleicht.[11] Während die Erstanträge der Antragsteller*innen noch manuell durch Sachbearbeiter*innen behandelt werden, werden Entscheidungen über Folgeanträge direkt durch dieses Modell übernommen. Dafür lädt das Modell Daten über das Einkommen, Steuerinformationen und Studienkredite automatisiert aus einer Datenbank. Unter-

6 *Molnar*, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, github.io 21.8.2023, https://christophm.github.io/interpretable-ml-book/ (19.2.2024) 49.

7 *Fisher,* Iris data set. UCI machine learning repository, archive.ics.uci.edu 30.6.1988, https://archive.ics.uci.edu/dataset/53/iris (19.2.2024).

8 *Knuth*, Lernende Entscheidungsbäume. Informatik Spektrum 2021, 364 (366).

9 *Molnar*, github.io, 102.

10 *Molnar*, github.io, 108.

11 *Kaun*, Suing the Algorithm: The Mundanization of Automated Decision-Making in Public Services Through Litigation, Information, Communication & Society 2022, 2046 (2048).

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

scheiden sich diese Daten signifikant von denen des Vormonats, stoppt das Tool den Antragsprozess und leitet diesen für die Entscheidung an die zuständigen Sachbearbeiter*innen weiter. Im Jahr 2020 wurde die Gemeinde Trelleborg zur Offenlegung des Quellcodes aufgefordert. Bis heute wurde das System noch nicht vollständig veröffentlicht. Dennoch kann angenommen werden, dass grds alle Voraussetzungen erfüllt sind, um die Entscheidungen dieses Modells durch Analyse des baumbasierten Entscheidungssystems zu erklären.[12]

### 3.3. Rechtliche Dimension

Das Trelleborg-Modell ist ein Bsp für den Einsatz eines KI-Systems im Rahmen der Hoheitsverwaltung. Geht man davon aus, dass Entscheidungen über Folgeanträge zur Aufrechterhaltung des Sozialhilfeanspruchs je einen neuen Bescheid darstellen, stellt das Trelleborg-Modell einen Fall eines vollautomatisierten Verwaltungsverfahrens dar. Derartige Bescheide weisen eine besondere Natur auf. Im Unterschied zum Regelfall der Rechtsanwendung zeichnen sich Verfahren wie das gegenständliche nämlich gerade dadurch aus, dass sie einen besonders hohen Grad an Standardisierung aufweisen. Dies macht sie einer Automatisierung durch KI-Systeme überhaupt erst zugänglich.[13]
Unterschiedliche Rechtsordnungen haben dies erkannt und vor diesem Hintergrund der Verwaltung die Möglichkeit eingeräumt, Verfahren mit einem besonders hohen Grad an Standardisierung (auch) auf KI-Systeme zu übertragen. Besonders bekannt ist dahingehend der deutsche § 35a VwVfG[14], der normiert, dass ein Verwaltungsakt „vollständig durch automatische Einrichtungen erlassen werden [kann], sofern dies durch Rechtsvorschrift zugelassen ist und weder ein Ermessen noch ein Beurteilungsspielraum besteht".[15] Entscheidungen über sozialrechtliche Ansprüche stellen dafür Paradebeispiele dar. Insofern verwundert es nicht, dass das deutsche Sozialgesetzbuch von der Ermächtigung des § 35a VwVfG Gebrauch gemacht hat. Dessen § 31a SGB X[16] normiert, dass im Rahmen der Sozialverwaltung ein Verwaltungsakt grds vollständig mit automatisierten Einrichtungen erlassen werden darf, „sofern kein Anlass besteht, den Einzelfall durch Amtsträger zu bearbeiten".

12 *Algorithmwatch*, Central Authorities Slow to React as Sweden's Cities Embrace Automation of Welfare Management, algorithmwatch.org 17.3.2020, https://algorithmwatch.org/en/trelleborg-sweden-algorithm/ (19.2.2024).

13 Grundlegend dazu bereits *G. Holzinger*, Der „Computerbescheid" in der Judikatur der Gerichtshöfe öffentlichen Rechts, in FS Rosenzweig (1988) 193.

14 Verwaltungsverfahrensgesetz (VwVfG), dBGBl I S 102 idF dBGBl 2023 I Nr 344.

15 Die entscheidende Qualifizierung als automatische Einrichtung erfolgt über die Art und Weise des Einsatzes: Der Erlass eines VAs muss „vollständig durch" die automatische Einrichtung erfolgen, dh diese muss funktional die eigentliche Entscheidungsfindung bewerkstelligen, also ohne das Mitwirken von Verwaltungsmitarbeiter*innen. S zB *Hornung* in *Schoch/Schneider*, VwVfG[3] (2022) § 35a Rz 66. Vgl auch *Siegel*, Automatisierung des Verwaltungsverfahrens – zugleich eine Anmerkung zu §§ 35a, 24 I 3, 41 IIa VwVfG –, DVBl 2017, 24 (26).

16 Das Zehnte Buch Sozialgesetzbuch – Sozialverwaltungsverfahren und Sozialdatenschutz – (SGB X), dBGBl I S 130 idF dBGBl 2023 I Nr 408.

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

Wendet man sich nun der hier im Zentrum stehenden Frage der Erklärung zu, so hat dies aus rechtlicher Perspektive eine zweifache Stoßrichtung: Zunächst sind staatliche Entscheidungen, und damit auch Bescheide, grds zu begründen. In Hinblick auf den soeben genannten § 31a SGB X ergibt sich dies aus der allgemeinen Anforderung des § 35 SGB X.[17] Freilich ist es denkbar, dass vollautomatisierte Bescheide von ebendieser Begründungspflicht ausgenommen werden; § 35 Abs 2 Nr 3 SGB X könnte auch so gedeutet werden, da er anordnet, dass es im Fall von Entscheidungen, die „mit Hilfe automatischer Einrichtungen" erlassen werden, keiner Begründung bedarf.[18] Doch selbst wenn man annimmt, dass Entscheidungen nach § 31a SGB X stets Entscheidungen, die „mit Hilfe automatischer Einrichtungen" erlassen werden, darstellen, kann ein*e Antragsteller*in dennoch eine Begründung für einen Bescheid verlangen (§ 35 Abs 3 SGB X). Aus einer solchen Begründung muss wiederum ersichtlich werden, welche tatsächlichen und rechtlichen Gründe für die Entscheidung wesentlich waren.[19] Der Umfang der Begründungspflicht richtet sich dabei nach den konkreten Verhältnissen des Einzelfalls.[20] Es ist nicht erforderlich, dass die Begründung sich mit allen Einzelheiten des Sachverhalts und Argumenten der Betroffenen auseinandersetzt.[21] Formelhafte Floskeln oder Wiederholungen des gesetzlichen Tatbestandes genügen dabei allerdings nicht.[22] Auch muss die Begründung jedenfalls so weit gehen, dass eine Nachprüfung der Entscheidung möglich ist.[23] Kommt ein KI-System wie das Trelleborg-Modell zum Einsatz, so muss dieses technologisch also auch in der Lage sein, eine derartige Begründung zu liefern.

In dieser Hinsicht wird vielfach nicht bloß die Begründung der in concreto getroffenen Entscheidung diskutiert, sondern auch die Frage, inwieweit darüber hinaus im Fall vollständig automatisierter Bescheide auch gewisse Anforderungen hinsichtlich der Nachvollziehbarkeit des Models, losgelöst von einer konkreten Entscheidung, bestehen. White Box Modelle, wie das Trelleborg-Modell, erscheinen dahingehend dem Grunde nach unproblematisch; zeichnen sie sich doch gerade dadurch aus, dass sie grds als leicht interpretierbar zu qualifizieren sind. Die Baumstruktur kann manuell durchlaufen werden, wodurch der Entscheidungsfindungsprozess abstrakt, aber auch konkret, nachvollziehbar wird. Doch auch bei einem solchen White Box Modell ist es – wie bereits betont – denkbar, dass zu viele Zweige oder zu viele technische nicht-interpretierbare Attribute bzw Kriterien innerhalb des Entscheidungsbaums die Interpretierbarkeit entscheidend verringern. So besteht etwa der Code im Fall Trelleborg aus 136.000 Zeilen und zahl-

17 *Mutschler* in *Rolfs et al*, KassKomm, § 31a Rz 3 (Stand 12/2020, beck.online) und *Siewert* in *Diering/Timme/Stähler*, LPK-SGB X[6] (2022) § 31a Rz 6.

18 So etwa *Mutschler* in *Rolfs et al*, KassKomm, § 31a Rz 2.

19 BSG 24.11.1983, 10 RAr 11/82, BSGE 56, 55.

20 Begründung des Regierungsentwurfs zum VwVfG, BT-Drs 7/910, 60; BSG 9.12.2004, B 6 KA 44/03 R, BSGE 94, 50; 23.10.1985, 7 RAr 32/84, BSGE 59, 30; 24.11.1983, 10 RAr 11/82, BSGE 56, 56 (55); BVerwG 14.3.1985, 5 C 145/83, BVerwGE 71, 139; BVerwG 15.5.1986, 5 C 33/84, BVerwGE 74, 196.

21 BSG 23.6.2020, B 2 U 14/18 R; LSG NRW 9.3.2021, L 18 R 306/20.

22 BSG 18.4.2000, B 2 U 19/99 R, NZA 2000, 994.

23 BVerwG 7.5.1981, 2 C 42/79, DVBl 1982, 198–199.

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

reichen Regeln, die sich auf 127 XML-Dateien verteilen. In diesem Fall ist es zwar noch möglich, aber sehr langwierig, einen allgemeinen Entscheidungsbaum zu skizzieren. Eine holistische Bewertung, welches Feature in welcher Konstellation welches Gewicht aufweist, kann für die Endanwender*innen aber bereits in diesem Fall schwierig sein. Insofern erscheint es durchaus zweifelhaft, inwieweit KI-Systeme wie das Trelleborg-Modell trotz der Eigenschaft als White Box Modell umfassend erklärbar im rechtlichen Sinn sind.

## 4. Black Box Modelle

Wendet man sich nun Black Box Modellen zu, so zeichnen sich diese dadurch aus, dass ihre innere Funktionsweise im Unterschied zu White Box Modellen nicht von außen unmittelbar zugänglich ist, sie also nicht aus sich heraus erklärbar sind.[24] Grund dafür ist die Komplexität des zugrundeliegenden Modells: Im Beispiel neuronaler Netze, auf denen bspw alle Anwendungen aktueller Sprachmodelle basieren, sind die im Modell durchgeführten Berechnungen schon bei einfachen Architekturen so zahlreich, dass ein manuelles Nachverfolgen der Operationen nicht möglich ist.[25] Das bedeutet jedoch nicht, dass Entscheidungen von Black Box Modellen in keinem Fall erklärbar sein können, wie anhand des Beispiels modellagnostischer Methoden gezeigt wird.

### 4.1. Modellagnostische Erklärungen

Jedes Machine Learning Modell, das mittels eines Datensatzes trainiert wird und basierend auf diesem Training Vorhersagen über neue Datenpunkte treffen kann, kann grds durch sogenannte modellagnostische Methoden erklärt werden. Modellagnostisch werden diese Methoden dadurch, dass sie nicht auf spezifische Charakteristika einer Modellfamilie angewiesen sind – etwa eine Beschränkung auf bestimmte Datenstrukturen oder einen Trainingsprozesses – sondern jedes Modell, das die Grundlagen überwachten Lernens erfüllt, erklären können (dh, es spielt keine Rolle, ob etwa ein neuronales Netzwerk oder eine Regression erklärt wird).[26] Modellagnostische Erklärungen sind damit – zumindest ihrer Konzeption nach – unabhängig vom Modell und somit universell anwendbar. Ihr größter Vorteil ist ihre Flexibilität. Einerseits können durch die Austauschbarkeit des zugrundeliegenden Modells mehrere Modelle für dieselbe Aufgabe trainiert und mithilfe dieser Erklärungen verglichen werden, andererseits können auch der spezifische Erkläransatz und dessen Darstellung beliebig variiert werden (dh je nach Bedarf

24 *Rudin*, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence 2019, 206.

25 *Molnar*, github.io, 317.

26 *Arrieta et al*, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 2020, 82 (115).

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

kann die Erklärung als Text, Diagramm, oder in anderem Format dargestellt werden).[27] Gleichzeitig stehen modellagnostische Erklärungen aber auch dafür in der Kritik, etwa weniger eine gewissenhafte Erklärung des eigentlichen Black Box Modells, als eher eine statistische Zusammenfassung einzelner Vorhersagen zu liefern.[28]

### 4.2. Beispiel: LIME

Eine bekannte modellagnostische Methode sind Local Interpretable Model-Agnostic Explanations, kurz: LIME.[29] LIME verwendet eine spezifische Vorhersage eines Black Box Modells als Basis, um die Relevanz der einzelnen Inputfeature für diese Vorhersage zu berechnen (dh, das Gewicht der Features). Um diese Gewichtung zu bestimmen, verändert LIME geringfügig die Werte des Inputs (perturbiert diese) und zeichnet die Unterschiede auf, die diese Veränderungen in der Vorhersage des Modells auslösen. Führt etwa eine Veränderung in Features A zu einer anderen Vorhersage als dieselbe Veränderung in Feature B, hat Feature A in diesem Fall ein größeres Gewicht.

### 4.3. Use Case II: LIME zur Erklärung von Risikofaktoren bei Herzoperationen

Ein zentraler Einsatzbereich von XAI-Ansätzen wie LIME sind medizinische Settings. So trainierten Forscher*innen in einer hypothetischen Studie verschiedene ML-Modelle, um das Verhältnis zwischen Risikofaktoren und Sterblichkeitsrate bei Herzoperationen abzubilden.[30] Von anfänglich 38 Features (bestehend aus Variablen zu Begleiterkrankungen, Laborwerten und Demografie), wurden in dieser Studie durch den Vergleich zwischen verschiedenen ML-Modellen 12 Features als signifikant für die Vorhersage der Sterblichkeit identifiziert. Im Anschluss setzten die Autor*innen LIME für eine Analyse des Modells mit der besten Vorhersagegenauigkeit ein, um die jeweils fünf wichtigsten Features auf individueller Basis je Patient*in zu bestimmen.
Im Ergebnis erhielten die Autor*innen durch dieses Vorgehen je Patient*in eine durch das ML-Modell berechnete Überlebenswahrscheinlichkeit sowie eine Übersicht der Features, die für die jeweilige Vorhersage den größten Einfluss hatten. Zur Illustration führen die Autor*innen beispielhaft einen spezifischen Fall an, in dem ua die Einordnung als „Notfall“ (+ 8 %), der vorherige Einsatz von künstlicher Beatmung (+ 8 %), sowie auch die Ethnizität des oder der Patient*in (+ 7 %) relevant für die Vorhersage der Überlebenswahrscheinlichkeit (43 %) waren. Die Autor*innen argumentieren, dass durch die

27 *Molnar*, github.io, 145.
28 Vgl dazu etwa, am Bsp der Analyse des COMPAS-Modells durch die Rechercheplattform ProPublica, *Rudin*, Nature Machine Intelligence 2019, 206.
29 *Ribeiro/Singh/Guestrin*, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 1135.
30 *Rogers et al*, A Machine Learning Approach to High-Risk Cardiac Surgery Risk Scoring, Journal of Cardiac Surgery 2022, 4612 (4618).

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

Verwendung von LIME auch Nutzer*innen ohne detaillierte Kenntnis der algorithmischen Grundlagen die entsprechenden Modelle zur Risikoeinschätzung verwenden könnten. Unerwähnt bleibt jedoch, wie Personen ohne algorithmische Kenntnis Fehler in den Modellen erkennen können, und wie der Einsatz dieser sowie von LIME an die Patient*innen kommuniziert werden würde.[31] Weiterhin rücken durch die Darstellung der Feature-Gewichtungen in diesem Fall auch gesellschaftliche Fragen jenseits der direkten medizinischen Anwendung in den Vordergrund, wie durch den hohen Einfluss der Patient*innen-Ethnizität auf die Sterblichkeitsvorhersage gezeigt wird.

### 4.4. Rechtliche Dimension

Bei dem soeben dargestellten Use Case handelt es sich um eine KI-Anwendung, die in medizinischen Settings zum Einsatz kommt. Aus rechtlicher Perspektive ist somit das Medizinrecht der zentrale Anknüpfungspunkt. Für die Frage, ob und inwieweit ein KI-System aus medizinrechtlicher Sicht als erklärbar anzusehen ist, sind nicht bloß die Ausgestaltung des konkreten KI-Systems, sondern auch die Anwender*innen, also Ärzt*innen und Patient*innen in Betracht zu ziehen. Im Zentrum steht dabei die Frage, ob und inwieweit diese jeweils die Funktionsweise des KI-Systems nachvollziehen können und welche Informationen dafür von wem zur Verfügung gestellt werden müssen.
Bei Use Case II, also LIME als ein modellagnostischer Erklärungsansatz, steht die Offenlegung der Art, wie unterschiedliche ML-Modelle jeweils die Gewichtung von Risikofaktoren für die Sterblichkeit bei Herzoperationen bestimmen, im Zentrum. Im Verhältnis zu Ärzt*innen ist anzunehmen, dass derartige ML-Modelle diese im Rahmen einer Diagnose sowie der Festlegung weiterer Behandlungsschritte unterstützen, nicht aber ersetzen sollten.[32] Für diese Konstellation ist allen voran die allgemeine ärztliche Sorgfaltspflicht nach § 49 ÄrzteG von Relevanz. Sie normiert die Notwendigkeit einer Plausibilitätsprüfung durch die Ärzt*innen und verbietet damit im Umkehrschluss eine ungeprüfte, „blinde“ Übernahme von KI-Ergebnissen. Die KI-Anwendung und ihr Entscheidungsfindungsprozess müssen also jedenfalls insoweit erklärbar sein, dass Ärzt*innen eine derartige Plausibilitätsprüfung durchführen können.
In Hinblick auf Patient*innen steht aus rechtlicher Sicht die Debatte rund um Art und Ausmaß eines sogenannten „informed consent“ im Zentrum.[33] Dahinter steckt der Gedanke, dass Patient*innen die Möglichkeit haben müssen, Risiken, die mit einer bestimmten Behandlung einhergehen, selbst abzuwägen, bevor sie eine Entscheidung treffen. Im

31 *Rogers et al*, Journal of Cardiac Surgery 2022, 4612 (4617).
32 Würde ein KI-System demgegenüber gänzlich eine Aufgabe übernehmen, die eigentlich von Ärzt*innen auszuführen wäre, wäre dies dem Grunde nach an § 2 Abs 1 und Abs 2 Ärztegesetz 1998, BGBl I 1998/169 (im Folgenden: ÄrzteG) zu messen.
33 Es handelt sich dabei um ein primär von der Judikatur herausgebildetes Prinzip des Medizinrechts und korrespondiert mit der ärztlichen Aufklärungspflicht. Grundlegend etwa OHG 15.12.1964, 8 Ob 342/64, EvBl 1965/217. Vgl weiters OGH 18.4.1973, 1 Ob 66/73, RZ 1973, 167.

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

Fall des Use Case II wird die KI-Anwendung im Rahmen der Diagnose durch Ärzt*innen eingesetzt. Vor diesem Hintergrund ist anzunehmen, dass Patient*innen jedenfalls von diesen darüber aufgeklärt werden müssen, dass KI zum Einsatz kommt. Auch scheint es erforderlich zu sein, Patient*innen über die Treffsicherheit des Modells sowie darüber, ob der Einsatz der KI-Anwendung dem Stand der Wissenschaft entspricht, zu informieren. Wie weitreichend Patient*innen über die konkrete Entscheidung des KI-Systems oder aber auch über die abstrakte Funktionsweise des Tools aufzuklären sind, ist im Einzelnen umstritten. Dem Grunde nach wird wohl gerade vor dem Hintergrund, dass die Informationen für die Patient*innen verständlich sein sollen, kein allzu strenger Maßstab an die technische Detailtiefe anzusetzen sein.[34] Die Erklärung, die LIME gegenüber dem*der Ärzt*in abgibt, sollte – jedenfalls in Form einer aufbereiteten Weitergabe durch den*die Ärzt*in an Patienten*innen – die dahingehenden Anforderungen erfüllen.

Jenseits der unmittelbaren „Interaktion" durch Ärzt*innen oder Patient*innen mit dem KI-System, sei in Hinblick auf den hier behandelten Use Case noch angemerkt, dass für diesen aufgrund des Einsatzes im medizinischen Bereich das Regime des Medizinprodukterechts maßgeblich ist. Die Qualifikation als Medizinprodukt iSd Art 2 Z 1 MPVO[35] iVm MPG[36] (einschlägig ist hier „Software")[37] ist insoweit von Bedeutung, als Medizinprodukte ein sogenanntes Konformitätsbewertungsverfahren nach Art 52 MPVO iVm Anhang IX bis XI durchlaufen müssen.[38] In diesem spielt wiederum die Frage, inwieweit die medizinische KI-Anwendung erklärbar ist, eine grundlegende Rolle. Handelt es sich nämlich um eine Black Box, die sich auch nicht mit Hilfe von XAI zumindest teils öffnen lässt, so ist es gerade für Außenstehende kaum möglich, eine Bewertung, ob und inwieweit das System als sicher qualifiziert werden kann, vorzunehmen.

Schließlich bildet die Qualifikation als Medizinprodukt den zentralen Anknüpfungspunkt für den in der Endphase der Entstehung befindlichen AIA der EU. Aktuell geht insb aus Art 6 AIA[39] hervor, dass KI-Systeme, die Medizinprodukte darstellen, grds einer besonders strengen Regulierung unterliegen, da es sich um Hochrisiko-Systeme handelt. In Hinblick auf die hier im Zentrum stehende Erklärbarkeit ist besonders das Erfordernis, dass das System „hinreichend transparent" sein muss und Nutzer die Ergebnisse angemessen „interpretieren und verwenden können" müssen, von Bedeutung. Ebenso der Aspekt der menschlichen Aufsicht, um allfällige Risiken bestmöglichen hintanzuhalten, lässt sich gerade im Kontext eines KI-Systems, das in medizinischen Settings zum Ein-

34 *Stöger*, Explainability und »informed consent« im Medizinrecht, in *Leyens/Eisenberger/Niemann*, Smart Regulation (2021) 143; *Paar/Stöger*, Medizinische KI, in *Fritz/Tomaschek*, Konnektivität (2021) 85.

35 VO (EU) 2017/745 des EP und des Rates v 5.4.2017 über Medizinprodukte, ABl L 2017/117, 1.

36 Medizinproduktegesetz (MPG) 2021 BGBl I 2021/122.

37 Weiters ist diese KI-Anwendung im Lichte des Art 51 MPO den Risikoklassen zuzuweisen; für Software ist allen voran Anhang VIII Regel 11 maßgeblich.

38 Vgl dazu auch Art 61 MPVO.

39 Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act, AIA) and amending certain union legislative acts, COM(2021) 206 final.

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout

satz gelangt, fruchtbar machen, und zwar dahingehend, dass den Ärzt*innen eine bedeutsame Funktion zukommt, ein entsprechendes Verständnis für die verwendete KI-Anwendung, ihre Funktionsweise, ihre Leistungsfähigkeit, aber auch ihre Grenzen mitzubringen. Gegenüber Patient*innen bedarf es wiederum einer entsprechenden Aufklärung seitens der Ärzt*innen. Ausgenommen von der Qualifikation als Hochrisiko-Systeme und damit auch den strengen rechtlichen Anforderungen sind nur jene KI-Systeme mit limitiertem bzw geringem Risiko; in diesem Fall werden bloße Transparenzanforderungen gestellt.[40]

## 5. Conclusio

Der Beitrag beschreibt ausgewählte Anwendungsfälle von XAI und illustriert einige der aus technologischer und rechtlicher Perspektive auftretenden Herausforderungen. Dabei wird deutlich, dass nicht nur der Begriff der KI, sondern auch die Ansätze, um konkrete KI-Systeme erklärbar zu machen, heterogen sind: Nicht nur ist eine Erklärung abhängig vom zugrundeliegenden Modell, sondern auch von der Zugänglichkeit zu dessen Funktionsweisen sowie dem Interesse, Wissen und den Anforderungen des Zielpublikums. Inhalt, Form und Nutzen der Erklärung müssen deshalb dem jeweiligen Kontext angepasst werden. Der jeweils spezifische rechtliche Rahmen, der sodann an den konkreten Anwendungsbereich eines KI-Systems anknüpft, potenziert die Komplexität dieses Themenbereichs. Aus rechtlicher Perspektive sind nämlich uU selbst jene KI-Systeme, die aus technologischer Sicht als erklärbar gelten, nicht notwendigerweise als (ausreichend) erklärbar anzusehen. Erklärungsansätze zu entwickeln, die diesem Erfordernis größter Flexibilität zur Erfüllung faktischer wie normativer Anforderungen gerecht werden, ist eine in vielerlei Hinsicht noch nicht bewältigte Herausforderung der XAI-Forschung. Eine stärkere interdisziplinäre Zusammenarbeit zwischen Rechtswissenschaften und Informatik könnte einen großen Beitrag dazu leisten, Erklärungsansätze auszugestalten und umzusetzen, die diesen hohen Anforderungen genügen.

Dr.in Elisabeth Paar, LL.M. (Yale) ist Universitätsassistentin PostDoc am Institut für Öffentliches Recht und Politikwissenschaft der Universität Graz; elisabeth.paar@uni-graz.at

Timothée Schmude, M.A. ist Doktorand an der Fakultät der Informatik der Universität Wien; timothee.schmude@univie.ac.at

Dipl.-Jur.in Cansu Cinar ist Universitätsassistentin am Institut für Rechts- und Verfassungsgeschichte der Universität Wien; cansu.cinar@univie.ac.at

40 Vgl dazu Art 52 AIA.

Lizenziert für Universität Wien Hauptbibliothek am 19.07.2024 um 17:38 Uhr

© Verlag Österreich 2024

BiblioScout