Die Nachvollziehbarkeit von KI-Anwendungen in der Medizin

Eine Betrachtung aus juristischer Perspektive mit Beispielszenarien

Stefanie Hänold (Institut für Rechtsinformatik, Gottfried Wilhelm Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, haenold@iri.uni-hannover.de),

Nelli Schlee (Institut für Rechtsinformatik, Gottfried Wilhelm Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, schlee@iri.uni-hannover.de),

Dario Antweiler (Fraunhofer IAIS, Fraunhofer Center for Machine Learning, Schloss Birlinghoven, 53757 Sankt Augustin, dario.antweiler@iais.fhg.de) und

Katharina Beckh (Fraunhofer IAIS, Competence Center for Machine Learning Rhein-Ruhr, Schloss Birlinghoven, 53757 Sankt Augustin, katharina.beckh@iais.fhg.de)

Es bestehen keine Interessenskonflikte.

Abstract:

KI hat ein hohes Potenzial, die medizinische Versorgung zu verbessern. Eine große Rolle wird dabei die Nachvollziehbarkeit von KI-Anwendungen spielen, denn diese ermöglicht, die algorithmischen Prozesse besser zu verstehen. Der Beitrag befasst sich mit den verschiedenen Formen der Nachvollziehbarkeit von KI-Systemen und setzt diese für potenzielle KI-Anwendungen in der Notfallversorgung aus dem Projekt LOTTE in einen spezifischen Kontext. Das Verständnis für das Thema „Explainable AI“ ist aus juristischer Perspektive von besonderer Relevanz, da die Nachvollziehbarkeit von KI-Technologien für die Einhaltung gesetzlicher Vorgaben des Medizinprodukterechts, des Datenschutzrechts sowie des AGG einen bedeutenden Faktor darstellt. Zwar wird die Nachvollziehbarkeit nicht direkt gesetzlich gefordert, das Erfordernis ergibt sich jedoch indirekt bzw. wird die Erfüllung gesetzlicher Vorgaben erheblich erleichtern.

I. Einleitung

Künstliche Intelligenz (KI) gilt als eine der großen Hoffnungsträgerinnen in der Medizin. Sie verspricht neue Möglichkeiten in der Diagnose und der Therapie von Krankheiten sowie die Beschleunigung und Effektivierung von Arbeitsabläufen. Das Projekt LOTTE* (Leitsystem zur Optimierung der Therapie traumatisierter Patientinnen und Patienten bei der Erstbehandlung) hat konkrete Anwendungsszenarien für mögliche KI-Anwendungen in der Notfallversorgung ausgelotet. Intensiver Erörterung unterlag dabei die Nachvollziehbarkeit der von KI-Anwendungen generierten Entscheidungen, wie konkreten Behandlungsvorschlägen oder Scores, sowie die damit einhergehenden rechtlichen und ethischen Implikationen. Dieser Beitrag hat zum Ziel, anhand von Beispielszenarien darzulegen, inwiefern Entscheidungen eines KI-Systems nachvollziehbar sein können und welche Bedeutung die Nachvollziehbarkeit im Hinblick auf die rechtlichen Anforderungen an den Einsatz von KI-unterstützten Anwendungen im medizinischen Sektor hat.

II. Das Projekt LOTTE

1. Gesamtziel

Die Versorgung von schwerverletzten Personen zählt mit zu den komplexesten Aufgaben in der Unfallchirurgie. Dies gilt insbesondere für den Zeitraum bis zur Aufnahme der verletzten Person auf eine Intensivstation oder Verlegung in ein spezialisiertes Zentrum. Vor allem das plötzliche Auftreten der Unfallsituation, die Heterogenität der Patientinnen und Patienten und die oft hohe zeitliche Dynamik des Krankheitsbildes stellen das ärztliche Personal vor große Herausforderungen. Ziel des Projekts war es, konkrete Anwendungsmöglichkeiten für intelligente Technologien zur Unterstützung klinischer Entscheidungen bei der Erstbehandlung von schwerverletzten Personen zu identifizieren und diese mit Blick auf die Implementierung in der Praxis aus medizinischer, technologischer, ökonomischer sowie ethischer und rechtlicher Perspektive zu bewerten.

* Förderung durch den Bund, Bundesministerium für Gesundheit, Kennzeichen ZMVI1-2519DAT703. Die Forschungsarbeiten wurden zudem unterstützt durch das Bundesministerium für Bildung und Forschung im Rahmen des Kompetenzzentrums für maschinelles Lernen ML2R, Förderkennzeichen 01|S18038B. Diese Arbeit ist zum Teil im Forschungszentrum Maschinelles Lernen im Fraunhofer Cluster of Excellence „Cognitive Internet Technologies“ entstanden.

2. Potenzielle KI-Anwendungen in der Notfallversorgung und zugrundeliegende Modelltechnologien

Es gelang innerhalb des Projekts eine Sammlung möglicher Einsatzszenarien von KI-Anwendungen zu erstellen.[1] Im Folgenden werden drei zentrale Szenarien vorgestellt: die Trajektorien-Klassifikation, die OP-Risikoprognose sowie das Intelligente Leitlinien-Interface.

Die Trajektorien-Klassifikation setzt zu Beginn der Behandlung im Schockraum an. Das System berechnet eine objektive und quantitative Einschätzung der Fallkomplexität und des erwarteten Verlaufs. Das Ergebnis wird durch die Schockraum-Leitung abgerufen und dem gesamten Schockraum-Team digital visualisiert zur Verfügung gestellt. Ärztinnen bzw. Ärzte sehen sich entweder in ihrer bisherigen Einschätzung bestätigt oder müssen diese überdenken und je nach Ergebnis eine Nachjustierung der Behandlungsstrategie vornehmen. Wird z.B. ein besonders kritischer Verlauf prognostiziert, können frühzeitig Maßnahmen – wie eine Notoperation oder eine Massentransfusion – ergriffen werden. Ziel dieses Szenarios ist die Verbesserung der Behandlungsqualität.

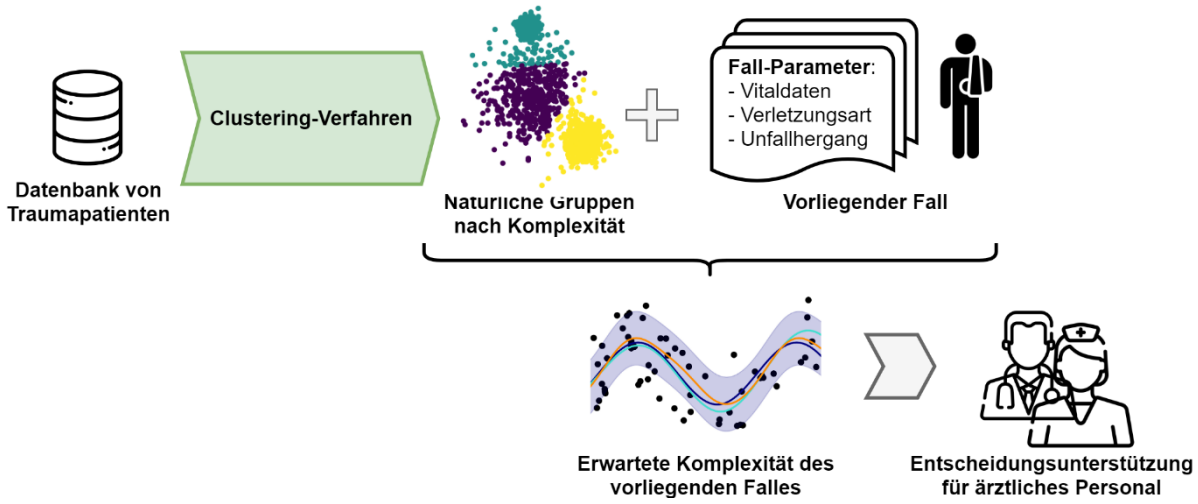

Bild 1: Die Trajektorien-Klassifikation beruht auf einer Analyse historischer Falldaten, deren Einteilung in natürliche Fallgruppen und der Berechnung der zu erwartenden Komplexität

[1] *Antweiler/Beckh/Sander/Rüping*, https://www.iais.fraunhofer.de/lotte, Zugriff am 11.11.2020.

Die OP-Risikoprognose setzt bei dem Problem an, dass schwerverletzte Personen ein hohes Risiko für Komplikationen bei Operationen haben. Die Risiken einer OP übersteigen oft den potenziellen Nutzen. Das System berechnet das individuelle Risiko mehrerer Komplikationstypen und unterstützt die Entscheidungsfindung zum Ergreifen einer operativen Maßnahme. Ziel ist die Reduktion lebensgefährlicher Komplikationen, wie einer Sepsis oder akutem Nierenversagen.

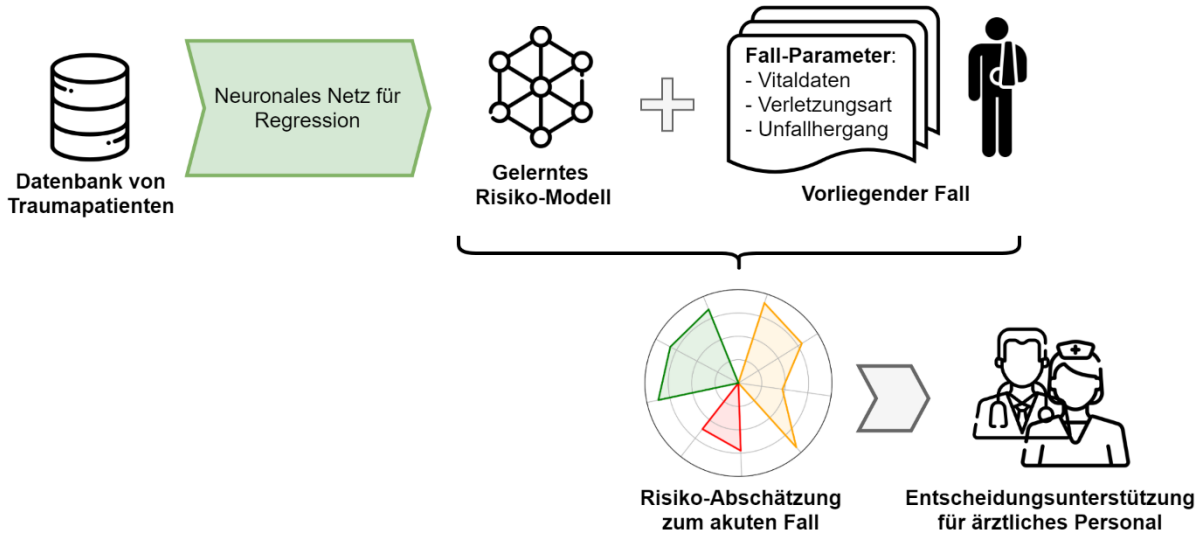

Bild 2: Die OP-Risikoprognose soll aus historischen Daten Zusammenhänge zwischen Fallparametern und Komplikationsrisiken erlernen und diese auf den vorliegenden Fall übertragen

Das Leitlinien-Interface adressiert den Umstand, dass Umfang und Komplexität der für die Schwerverletztenversorgung genutzten S3-Leitlinie eine alltägliche und patientenspezifische Nutzung erschweren. Es erstellt nach Eingabe der Fallparameter eine Visualisierung des leitlinienkonformen Behandlungsablaufs sowie fallspezifische Behandlungsempfehlungen. Dadurch wird die Priorisierung von Maßnahmen unterstützt und begründete Leitlinienabweichungen werden dokumentiert.

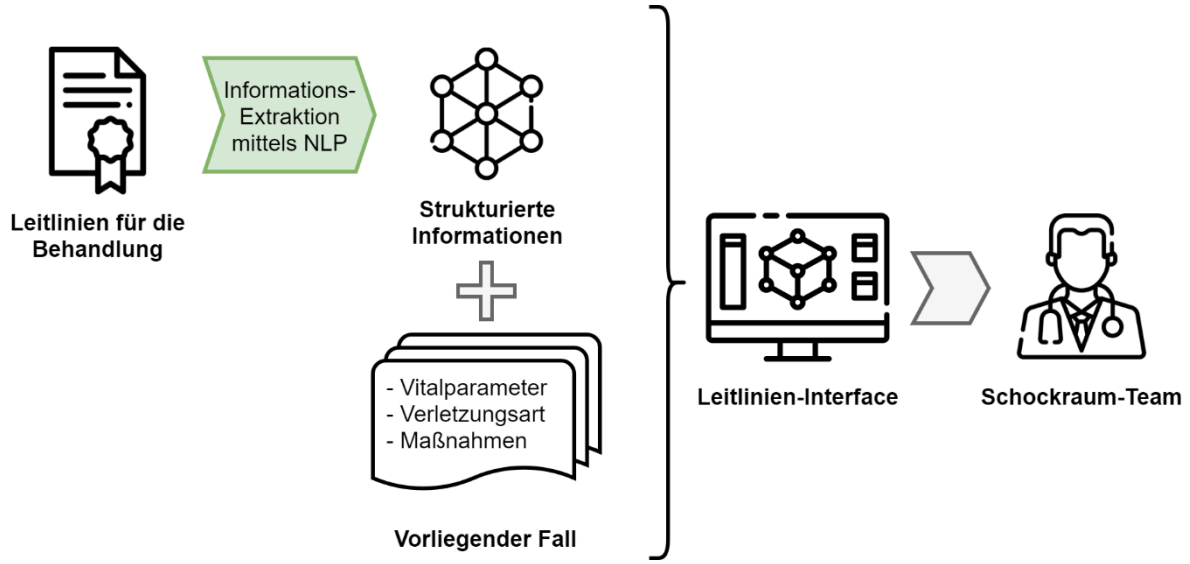

Bild 3: Das Intelligente Leitlinien-Interface bereitet die Inhalte von Leitlinien mittels Natural Language Processing (NLP) strukturiert auf und schlägt fallspezifische Behandlungsabläufe vor

Alle drei Szenarien beruhen auf der Auswertung von historischen Patientendaten und sollen ermöglichen, den akuten Fall mit hunderttausenden vorangegangenen Fällen zu vergleichen. Ziel ist es, innerhalb einer großen Population von abgeschlossenen Fällen statistische Zusammenhänge zu identifizieren, die einen Teilaspekt des Behandlungsverlaufs annähernd gut beschreiben können und im besten Falle prädiktive Aussagekraft besitzen. Die verfügbaren Daten werden dabei kombiniert mit Wissen – also Informationen, Leitlinien und Erfahrungswerten von medizinischem Fachpersonal.

Den KI-Methoden in den theoretisch entwickelten Anwendungsszenarien liegen unterschiedliche Modelltechnologien zu Grunde. Die Trajektorien-Klassifikation teilt den vorliegenden Fall auf Basis der vorhandenen Daten zu Unfallhergang und Vitalzustand in eine geringe Anzahl von Komplexitätsklassen ein. Dazu nimmt zunächst das ärztliche Fachpersonal eine Einschätzung der Komplexität von vergangenen Fällen vor, auf dessen Basis der Algorithmus einen Zusammenhang zwischen eingehenden Parametern und Komplexitätsniveau erlernt. Für die Identifikation natürlicher Komplexitäts-Gruppen innerhalb der Population wird ein sogenannter „Clustering“-Algorithmus eingesetzt, der Ähnlichkeiten in den Behandlungsverläufen identifiziert.

Bei der OP-Risikoprognose handelt es sich um ein Regressionsproblem. Dabei lernt ein Modell aus historischen Daten von Patientinnen und Patienten den Zusammenhang zwischen eingehenden Fall-Parametern (z.B. Unfallhergang, Labor- und Vitalwerte) und der Eintrittswahrscheinlichkeit von OP-Komplikationen. Hierfür müssen Protokolle und Patientendaten vergangener Operationen in strukturierter Form vorliegen.

Beim Intelligenten Leitlinien-Interface wird die Modellklasse der Klassifikation eingesetzt. Die Modellarchitektur ist ein Entscheidungsbaum, der anhand der vorliegenden Fallparameter sowohl die akut notwendigen Entscheidungen entlang des Behandlungspfades identifiziert als auch die sinnvollsten nächsten Schritte bestimmt. Der Aufbau des Entscheidungsbaumes erfolgt durch statistische Auswertung und auf Basis von Informationsextraktionsmethoden.

III. Einleitende allgemeine Ausführungen zur Bedeutung der Nachvollziehbarkeit von KI-Anwendungen in der Medizin

Methoden der KI identifizieren und extrahieren statistische Zusammenhänge in vorhandenen Datensätzen und übertragen diese auf neue Datenpunkte. Die Systeme basieren dabei zunächst nur auf Korrelationen. Sie erkennen keine Kausalitäten, d.h. konkrete Beziehungen zwischen Ursache und Wirkung. So ist es möglich, dass sich Entscheidungen im Trainingsprozess im Ergebnis als richtig erweisen, das System aber auf ein sachfremdes Merkmal abgestellt hat und deshalb in einem neuen Anwendungsszenario eine nicht sachgemäße Entscheidung vom System getroffen wird. Um dies zu erkennen, bedarf es der Nachvollziehbarkeit algorithmischer Prozesse.

Die Nachvollziehbarkeit ermöglicht auch, im Nachhinein aufgrund der festgestellten Korrelationen vorhandene kausale Zusammenhänge zwischen Datenpunkten herzustellen. Die Wechselwirkungen zwischen den eingehenden Parametern können jedoch selbst für Expertinnen und Experten äußert komplex sein. Sind beispielsweise Ergebnisse auf Basis eines Entscheidungsbaumes noch relativ gut nachvollziehbar, können dagegen die Millionen von lernbaren Parametern innerhalb eines neuralen

Netzes lediglich mit extremem Aufwand (z.B. mittels Aktivierungskarten für Faltungsnetzwerke)[2] interpretiert werden.

Während sich bei einem Algorithmus, welcher Kaufempfehlungen für Besuchende einer Webseite erstellen soll, regelmäßig keine erheblichen Nachteile bei unsachgemäßen Vorschlägen ergeben, gibt es Situationen, in denen eine Nachvollziehbarkeit der KI-Anwendung kritisch ist, da das Befolgen der Entscheidung zu irreparablen Schäden führen könnte. Wenn KI-Systeme im medizinischen Bereich eingesetzt werden, kann die Gesundheit der Patientinnen und Patienten negativ beeinträchtigt werden. Daher besteht das Erfordernis sachlicher Erwägungen für eine Diagnose oder einen Therapievorschlag[3]. Zudem sollte es möglich sein, dass das System eine Erklärung für eine Entscheidung bereitstellt, die auf die jeweilige Sachkenntnis des Nutzers, z.B. der Ärztin oder des Arztes, zugeschnitten ist[4]. Wenn KI-Systeme selbstständige Entscheidungen treffen sollen und eventuell korrigierende Einschätzungen einer Ärztin oder eines Arztes wegfallen, wird als Bedingung für Vertrauen in das System vollständige Nachvollziehbarkeit des Entscheidungsprozesses genannt[5].

IV. Definition von Transparenz, Interpretierbarkeit, Erklärbarkeit

Die Begriffe Transparenz, Interpretierbarkeit und Erklärbarkeit werden im Zusammenhang mit der Nachvollziehbarkeit von algorithmischen Entscheidungen sowohl in der Praxis als auch in der wissenschaftlichen Literatur nicht einheitlich verwendet und oft auch synonym gebraucht. Es gibt keine verbindlichen bzw. standardmäßigen Definitionen. Im Folgenden wird klargestellt, wie die Begriffe innerhalb dieses Beitrags zu verstehen sind. Transparenz beschreibt den Umstand, dass ein trainiertes Machine-Learning-Modell der Anwenderin bzw. dem Anwender als White-Box- oder als Black-Box-Modell vorliegen kann. Bei ersterem liegt die vollständige mathematische Formulierung des Modells vor und ermöglicht eine detaillierte Validierung der Güte und Robustheit des Modells durch Inspektion aller trainierten Parameter und Modellarchitekturen. Bei letzterem hingegen besitzt die

[2] *Simonyan*/*Vedaldi*/*Zisserman*, https://arxiv.org/abs/1312.6034, Zugriff am 5.11.2020.
[3] Vgl. das Bsp. bei *Krumm*/*Dwertmann*, in: *Wittpahl*, Künstliche Intelligenz, 2019, S. 169 f.
[4] Vgl. Hochrangige Expertengruppe für künstliche Intelligenz, Ethikleitlinien für eine vertrauenswürdige KI, 2019, S. 22.
[5] *Dettling*/*Krüger*, PharmR 2018, 521.

Anwenderin bzw. der Anwender lediglich eine Form des Modells, welches eingehende Daten verarbeitet und eine Ausgabe produziert, ohne die inneren Vorgänge zu offenbaren. Die Anwenderin bzw. der Anwender kennt auch die eingesetzten Trainingsdaten nicht. Interpretierbarkeit liegt vor, wenn die KI-Anwendung vollständig nachvollziehbar ist, d. h. die KI-Anwendung erlaubt, die Abbildung von Eingabe der Daten zur Ausgabe der Systementscheidung zu verstehen. Erklärbarkeit ist erreicht, wenn die wesentlichen Einflussfaktoren ausgemacht werden können.

V. Welches Modell ist wie nachvollziehbar?

Die Entwicklung von Methoden zur Erhöhung der Nachvollziehbarkeit von KI-Methoden ist Gegenstand aktueller Forschung. Existierende Ansätze lassen sich in intrinsisch interpretierbare Modelle und modellagnostische Methoden untergliedern. Intrinsisch interpretierbare Modelle sind Modellarchitekturen, die auf Grund ihrer algorithmischen Struktur bereits eine hohe Nachvollziehbarkeit besitzen und im hiesigen Kontext als interpretierbar eingeordnet werden können. Ein Beispiel dafür sind Entscheidungsbäume, bei denen jede einzelne Entscheidung eines KI-Systems eindeutig nachvollzogen und häufig von Expertinnen und Experten sinnvoll gedeutet werden kann. Modellagnostische Methoden bezeichnen Verfahren, welche unabhängig von der gewählten Architektur eingesetzt werden können. Diese lassen sich zusätzlich unterteilen in solche, die individuelle Entscheidungen des KI-Systems erklären, in dem sie die wesentlichen eingehenden Parameter für die Entscheidung ausfindig machen und anzeigen, sowie Verfahren, welche die Gesamtheit eines gelernten Modells analysieren können. Während erstere ermöglichen, zu verstehen, warum das System zu einer bestimmten Entscheidung gekommen ist, eröffnen letztere die Möglichkeit, systematische Fehlleistungen, z.B. unzulässige bzw. unerwünschte Diskriminierungen, zu erkennen. Zu diesen Verfahren gehören unter anderem der „Permutation Feature Importance“-Algorithmus[6] sowie die „Shapley Additive exPlanations (SHAP)“[7], die jeweils für gesamte Modelle bzw. für einzelne Modellausgaben den Einfluss der eingehenden Parameter auf die Ausgabe abschätzen.

[6] *Fisher/Rudin/Dominici*, Journal of Machine Learning Research 20 (177) 2019, 1 f.
[7] *Lundberg/Su-In Lee*, A Unified Approach to Interpreting Model Predictions, NIPS 2017, https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf, Zugriff am 16.10.2020.

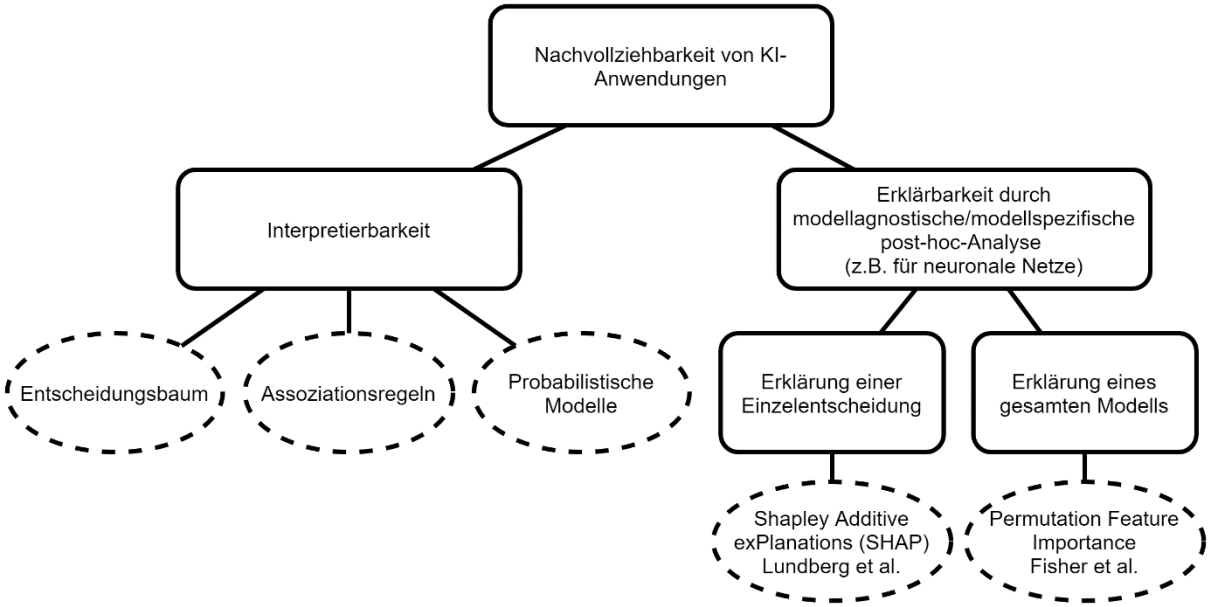

Bild 4: Formen der Nachvollziehbarkeit von KI-Anwendungen inkl. Beispielen

Eine Analyse der KI-Modelltechnologien in den vorgestellten Einsatzszenarien ergab: Der Klassifikations-Teil der Trajektorien-Klassifikation kann durch eine passende Modellarchitektur, wie z.B. einen Entscheidungsbaum, oder modellagnostische Erklärungsansätze zumindest partiell interpretierbar gestaltet werden. Die Clustering-Phase nimmt als sogenanntes „unüberwachtes“[8] Lernverfahren eine Sonderstellung ein. Im Falle des Clustering wird eine Vielzahl von Distanzen zwischen Datenpunkten in hochdimensionalen Vektorräumen berechnet, die sich einer Interpretierbarkeit typischerweise entziehen. Interpretierbare Clustering-Verfahren sind Gegenstand aktueller Forschung[9].

Die OP-Risikoprognose kann in der beschriebenen Variante als interpretierbares KI-System gestaltet werden. Dafür können sowohl für die historischen Trainingsdaten als auch für einen vorliegenden Fall die Einflussgrößen der eingehenden Parameter auf die Ausgabe des Systems berechnet werden.

Das Leitlinien-Interface kann hinsichtlich der Interpretierbarkeit unterschiedlich umgesetzt werden. Ein automatischer Abgleich der akuten Situation mit den Vorgaben

[8] Im Gegensatz zu den „überwachten" Lernverfahren, bei denen das Modell auf eine oder mehrere gegebene Zielvariablen hin angepasst wird, werden hier lediglich inhärente Zusammenhänge in den Daten identifiziert.
[9] *Bertsimas/Orfanoudaki/Wiberg*, Interpretable Clustering via Optimal Trees, Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.

durch die Leitlinie mittels einer Klassifikation in einem Entscheidungsbaum erlaubt dabei eine gewisse Interpretation.

Titel	**KI-Technologie**	**Nachvollziehbarkeit**
Trajektorien-Klassifikation	Clustering und Klassifikation	Clustering-Verfahren weisen typischerweise geringe Interpretierbarkeit auf. Werden die identifizierten Gruppen durch Expertinnen und Experten charakterisiert, so kann der darauffolgende Klassifikationsschritt z.B. mit Hilfe von Entscheidungsbäumen interpretierbar gestaltet werden.
OP-Risikoprognose	Regression mit neuronalem Netz	Architektur mit geringer Modell-Interpretierbarkeit – mit modellagnostischen Verfahren können zumindest Einflussfaktoren für eine Entscheidung priorisiert werden.
Leitlinien-Interface	Klassifikation mit Entscheidungsbaum	Entscheidungsbäume sind äquivalent zu einer Menge von Regeln, die bei geringer Länge und Anzahl für hohe Interpretierbarkeit sorgen.

Tabelle 1: Bewertung der Nachvollziehbarkeit für die Trajektorien-Klassifikation, die OP-Risikoprognose und das Leitlinien-Interface

VI. Rechtliche Implikationen der Nachvollziehbarkeit von KI-Anwendungen in der Medizin

Vorliegend werden drei Rechtsgebiete aufgegriffen, bei denen die Nachvollziehbarkeit von KI-Anwendungen im medizinischen Bereich eine bedeutende Rolle einnimmt: das Medizinprodukterecht, das Datenschutzrecht sowie Regelungen des Antidiskriminierungsrechts. Während ersteres die Zuverlässigkeit und Sicherheit von Medizinprodukten zum Regelungsziel hat, sind die hier relevanten Vorschriften des Datenschutzrechts und des AGG auf Selbstbestimmung und Fairness ausgerichtet.

1. Medizinprodukterecht

Medizinische Expertensysteme, wie die drei vorgestellten KI-Anwendungen, sind gem. Art. 2 Nr. 1 VO (EU) 2017/745 (Medizinprodukte-VO) als Medizinprodukt einzuordnen, da sie der Diagnoseunterstützung, der Therapieunterstützung oder dem Monitoring dienen. Medizinprodukte dürfen gem. Art. 5 Abs. 1 Medizinprodukte-VO nur in den Verkehr gebracht oder in Betrieb genommen werden, wenn sie bei sachgemäßer Lieferung, korrekter Installation und Instandhaltung und ihrer Zweckbestimmung gemäßen Verwendung der Medizinprodukte-VO entsprechen. Sie müssen dafür unter Berücksichtigung ihrer Zweckbestimmung den in Anhang I Medizinprodukte-VO festgelegten für das Produkt geltenden Sicherheits- und Leistungsanforderungen genügen (Abs. 2). Ein Nachweis der Einhaltung der Anforderungen umfasst eine klinische Bewertung (Abs. 3).

Nach Ziff. 1 Anhang I Medizinprodukte-VO müssen sich die Produkte für ihre normale Zweckbestimmung eignen. Sie müssen sicher und wirksam sein und dürfen weder den klinischen Zustand und die Sicherheit der Patientinnen und Patienten noch die Sicherheit und die Gesundheit der Anwenderinnen und Anwender oder Dritter gefährden. Bestehende Risiken müssen gemessen am Nutzen für die Patientin oder den Patienten vertretbar und mit einem hohen Maß an Gesundheitsschutz und Sicherheit vereinbar sein. Hier ist der allgemein anerkannte Stand der Technik zugrunde zu legen. Für Software sieht Ziff. 17.1 Anhang I Medizinprodukte-VO zusätzlich vor, dass Wiederholbarkeit, Zuverlässigkeit und Leistung entsprechend ihrer bestimmungsgemäßen Verwendung zu gewährleisten sind. Nach Ziff. 17.2 Anhang I Medizinprodukte-VO ist Software entsprechend dem Stand der Technik zu entwickeln, wobei die Grundsätze des Software-Lebenszyklus, des Risikomanagements einschließlich der Informationssicherheit, der Verifizierung und der Validierung zu berücksichtigen sind.

Bei der Verwendung von KI-Technologien ergeben sich besondere Herausforderungen in Bezug auf die Risikovorhersage und damit die Einschätzung der Sicherheit und Zuverlässigkeit, da innere Vorgänge des Systems teilweise schwer zu ergründen sind[10]. Entsprechende Standards oder gar eine harmonisierte Norm i.S.d.

[10] *Johner*, in: *Böttinger*/*Putlitz*, Regulatorische Anforderungen an Medizinprodukte, 2019, S. 367.

Art. 8 Medizinprodukte-VO für die Zuverlässigkeitsprüfung von Medizinprodukten mit KI-Komponenten gibt es noch nicht. Allerdings lässt sich schon jetzt feststellen, dass Transparenz, Interpretierbarkeit und Erklärbarkeit von KI-Systemen für die Bewertung der Sicherheit und Zuverlässigkeit eine erhebliche Rolle spielen. Hersteller müssen darlegen können, dass ausreichende und geeignete Trainingsdaten verwendet wurden und ein Bias in den Daten vermieden wurde. Sie müssen begründen können, warum die Ergebnisse nicht nur zufällig richtig sind. Die Interpretierbarkeit oder Erklärbarkeit eines Modells ermöglichen, zu klären, ob es lediglich „right for the wrong reasons" agiert, ohne sinnvolle Zusammenhänge zwischen Eingabe und Ausgabe identifiziert zu haben[11]. Besonders wenn, wie bei unüberwachten Lernverfahren, wie dem Clustering aus der Trajektorien-Klassifikation, quantifizierbare Gütekriterien schwierig zu erstellen sind und die Modellgüte nur mit entsprechender Fachexpertise einzuschätzen ist, erweisen sich Interpretationsmethoden als besonders nützlich. Ist eine KI-Anwendung zur Behandlung von Patientinnen und Patienten zumindest erklärbar, erhöht dies auch die Anwenderfreundlichkeit, da dem ärztlichen Fachpersonal ermöglicht wird, zu verstehen, welche Faktoren die Entscheidung des Systems herbeigeführt haben.

Die drei vorgestellten potenziellen Anwendungen müssten sich – bevor sie zur Behandlung von Patientinnen und Patienten eingesetzt werden können – aus medizinprodukterechtlicher Sicht als zuverlässig erweisen, wobei sich ein höherer Grad der Nachvollziehbarkeit als hilfreich bei der Risikoprognose erweisen wird[12]. Besonders das Leitlinien-Interface könnte durch Verwendung der vorgesehenen Modellarchitektur – Klassifikation mit Entscheidungsbäumen – als interpretierbares KI-System gestaltet werden, weshalb die Zuverlässigkeitsprüfung im Vergleich zu den zwei anderen Anwendungen mit weniger Herausforderungen verbunden sein wird.

[11] Zum sog. Kluger-Hans-Phänomen s. z.B. *Lapuschkin/Wäldchen/Binder/Montavon/Samek/Müller*, Nature Communications 10 (1096) 2019, https://www.nature.com/articles/s41467-019-08987-4, Zugriff am 16.10.2020.

[12] Detaillierte Anforderungen an die Überprüfung der Sicherheit finden sich im Leitfaden zu KI bei Medizinprodukten, Zugang über *Johner*, https://www.johner-institut.de/blog/regulatory-affairs/kuenstliche-intelligenz-in-der-medizin/, Zugriff am 17.8.2020.

2. Datenschutzrecht

a) Informations- und Auskunftspflichten bei automatisierten Einzelentscheidungen

Aus dem Datenschutzrecht ergeben sich ebenfalls Anforderungen, die mit Blick auf die Nachvollziehbarkeit von KI-Systemen relevant sind. Die Art. 13 Abs. 2 lit. f bzw. Art. 14 Abs. 2 lit. g VO (EU) 2016/679 (DS-GVO) sehen Informationspflichten des Verantwortlichen[13] vor, wenn eine automatisierte Einzelentscheidung i.S.d. Art. 22 DS-GVO ergeht. In diesen Fällen hat der Verantwortliche, um eine faire und transparente Verarbeitung zu gewährleisten, aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen der Verarbeitung für die betroffene Person zur Verfügung zu stellen. Der Art. 15 Abs. 1 lit. h DS-GVO enthält ein korrespondierendes Auskunftsrecht. Auch aus Art. 22 Abs. 3 DS-GVO könnten sich bestimmte Informationserfordernisse bei automatisierten Einzelentscheidungen ergeben.

aa) Vorliegen einer automatisierten Einzelentscheidung

Zunächst ist zu untersuchen, ob überhaupt eine automatisierte Einzelentscheidung i.S.d. Art. 22 DS-GVO vorliegt, da nur dann die in 2. a) genannten Regelungen relevant werden. Eine automatisierte Einzelentscheidung liegt gem. Art. 22 Abs. 1 DS-GVO vor, wenn eine ausschließlich auf einer automatisierten Datenverarbeitung beruhende Entscheidung ergeht, die der betroffenen Person gegenüber rechtliche Nachteile entfalten oder sie in ähnlicher Weise erheblich beeinträchtigen könnte. In Bezug auf die hier vorgestellten Anwendungen ist die entscheidende Frage, ob die Wahl der Behandlungsmaßnahmen allein auf Grund der vom KI-System generierten Entscheidung ergeht oder ob die Ärztin bzw. der Arzt eine inhaltliche Entscheidungsbefugnis hat und diese auch ausübt, sich also nicht vollständig auf das System verlässt[14]. Bei der Trajektorien-Klassifikation und der OP-Risikoprognose erfolgt eine Wahrscheinlichkeitsprognose, welche die Ärztin bzw. der Arzt bei seiner Entscheidung mitberücksichtigen soll. Es ist vorgesehen, dass die Ärztin bzw. der Arzt die vorhandenen Informationen und die Prognose des Systems prüft und eine eigene Behandlungsentscheidung trifft. Das Intelligente Leitlinien-Interface bietet

[13] „Verantwortlicher“ ist die Person oder Einrichtung, die über die Zwecke und Mittel der Verarbeitung von personenbezogenen Daten entscheidet (Art. 4 Nr. 7 DS-GVO).
[14] *Buchner*, in: *Kühling*/*Buchner*, DS-GVO BDSG, 2. Aufl. 2018, Art. 22 Rn. 15.; *Schulz*, in: *Gola*, DS-GVO, 2. Aufl. 2018, Art. 22 Rn. 14 f.

Empfehlungen für die zukünftigen Behandlungsschritte an, welche aber auch von der behandelnden Person zu überprüfen sind. Es erfolgt somit grundsätzlich keine ausschließlich auf der Prognose bzw. Empfehlung beruhende Behandlungsentscheidung. Die fachliche Richtigkeits- und Endkontrolle liegt noch bei der Ärztin bzw. dem Arzt.[15] Es sind jedoch Anwendungen denkbar, bei denen es keines Dazwischentretens einer Ärztin oder eines Arztes mehr bedarf, und klar eine automatisierte Einzelentscheidung vorliegt. Als Beispiel ist ein von der FDA als Medizinprodukt zugelassenes KI-gestütztes Diagnosesystem für diabetische Retinopathie zu nennen[16].

bb) Umfang der Informationspflichten bzw. Auskunftspflichten

Nur wenn eine automatisierte Einzelentscheidung vorliegt, wird die Frage relevant, welche Informationen nach den Informationspflichten gem. Art. 13 Abs. 2 lit. f bzw. Art. 14 Abs. 2 lit. g DS-GVO bzw. nach Ausübung des Auskunftsrechts nach Art. 15 Abs. 1 lit. h DS-GVO zur Verfügung zu stellen sind. Grundsätzlich haben sowohl die Informationspflichten als auch das Auskunftsrecht zum Gegenstand, dass aussagekräftige Informationen über die involvierte Logik zur Verfügung zu stellen sind. Aufgrund der unterschiedlichen Natur der Informationspflichten und des Auskunftsrechts könnten die Erklärungsinhalte in diesem Punkt jedoch abweichen. So sind die Informationen gem. Art. 13 Abs. 2 lit. h und Art. 14 Abs. 2 lit. g DS-GVO vorab zu erteilen, sodass es hier nahe liegt, dass nur generelle Informationen über die Funktionsweise des Systems darzulegen sind, wohingegen bei Erfüllung der Auskunftspflicht bereits Informationen zu der konkreten Entscheidung vorliegen und eine Erläuterung der konkreten Gründe erfolgen könnte[17]. Die Art. 29 Datenschutzgruppe, bestätigt durch den Europäischen Datenschutzausschuss, und Teile der Literatur gehen jedoch davon aus, dass die Informationspflichten und das

[15] Eine Erforschung von Automation Bias-Effekten durch die extreme Drucksituation in der Notfallversorgung sowie Haftungsangst wird angeregt.
[16] Einzelheiten bei *Dettling*, PharmR 2019, 636. Gegebenenfalls wäre für einen Einsatz in der Praxis auch eine Lockerung des teilweise gesetzlich geregelten Arztvorbehalts erforderlich *(Frost*, MPR 2019, 121 f.).
[17] *Kumkar/Roth-Isigkeit*, JZ 2020, 283; vgl. auch *Kaminski*, Berkeley Technology Law Journal 2019, 199 f.

Auskunftsrecht gleichlaufen, also nur generelle Informationen über die Funktionsweise zu erteilen sind[18].

Welche Informationen konkret über die Funktionsweise des KI-Systems bereitzustellen sind, ist umstritten. Nach der Art. 29 Datenschutzgruppe, sind die Kriterien, welche für die zu treffende Entscheidung ausschlaggebend sind, in verständlicher Art und Weise offenzulegen. Eine Offenlegung des Algorithmus ist nicht erforderlich. Es genügt, wenn generelle Informationen hinsichtlich der Faktoren, die eine Rolle spielen, und deren Gewichtung mitgeteilt werden[19]. Die Informationen sind nach Art. 12 Abs. 1 DS-GVO in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln. Gerade bei komplexen KI-Anwendungen erweist es sich mitunter als schwierig, die Informationen entsprechend aufzubereiten. Die Fülle an Informationen in Datenschutzerklärungen könnte zudem zu einer „Informationsüberlastung" der betroffenen Personen führen[20].

Soweit auch Betriebs- oder Geschäftsgeheimnisse oder Rechte des geistigen Eigentums betroffen sind, kann ein Spannungsverhältnis mit den Informationspflichten bzw. dem Auskunftsrecht entstehen[21]. Der Art. 15 Abs. 4 DS-GVO lässt in Bezug auf das Auskunftsrecht Raum für die Berücksichtigung der rechtlich geschützten Interessen des Verantwortlichen. Im EG 63 S. 6 DS-GVO heißt es dazu, dass das Auskunftsrecht Geschäftsgeheimnisse oder Rechte des geistigen Eigentums und insbesondere das Urheberrecht an Software, nicht beeinträchtigen sollte. Allerdings dürfe dies nicht dazu führen, dass der betroffenen Person jegliche Auskunft verweigert wird. Somit ist ein Interessenausgleich herbeizuführen, wobei in der DS-GVO unbeantwortet bleibt, ob der Verantwortliche auch über die Gewichtung der Merkmale

---

[18] Art. 29 Data Protection Working Party, Guidelines on Automated Individual Decision Making for the Purposes of Regulation 2016/679, WP 251 Rev. 01, S. 26 f., https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, Zugriff am 4.6.2020; Europäischer Datenschutzausschuss, Endorsement 1/2018, Januar 2018 https://edpb.europa.eu/sites/edpb/files/files/news/endorsement_of_wp29_documents_en_0.pdf, Zugriff am 16.10.2020; *Kumkar/Roth-Isigkeit*, JZ 2020, 283; *Wachter/Mittelstadt/Floridi*, International Data Privacy Law 2017, 83ff.; a.A. jedoch z.B. *Bäcker*, in: *Kühling/Buchner*, DS-GVO BDSG, Art. 15 Rn. 27, hält bei Art. 15 DS-GVO eine Ex-post-Erklärung der Entscheidung für erforderlich.

[19] Art. 29 Data Protection Working Party, Guidelines on Automated Individual Decision Making for the Purposes of Regulation 2016/679, WP 251 Rev. 01, S. 24 ff., https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, Zugriff am 4.6.2020.

[20] *Strassemeyer*, in: *Taeger*, Die Macht der Daten und der Algorithmen - Regulierung von IT, IoT und KI, DSRITB 2019, S. 32.

[21] *Bäcker*, in *Kühling*/*Buchner*, DS-GVO BDSG, 2. Aufl. Art. 13 Rn. 54; *Dix*, in: *Simitis*/*Hornung*/*Spiecker*, Datenschutzrecht, 1. Aufl. 2019, Art. 13 Rn. 16.

oder Vergleichsgruppen Angaben machen muss. Dies wird von einem Teil der Literatur befürwortet[22]. Ein anderer Teil geht davon aus, dass die Rechtsprechung des BGH, der eine Offenbarungspflicht hinsichtlich Vergleichsgruppen und Gewichtungen von Merkmalen verneinte[23], unter der DS-GVO fortgeführt werden wird[24]. Im Hinblick auf die Informationspflichten nach Art. 13 bzw. 14 DS-GO könnte eine Beschränkungsregelung i.S.d. Art. 23 DS-GVO das Spannungsverhältnis auflösen[25].

Auch wenn hier aus rechtlicher Perspektive noch viel umstritten ist, wird klar, dass die Nachvollziehbarkeit von KI-Entscheidungen je nachdem, welche Ansicht sich in der Praxis durchsetzen wird bzw. welche Regelung gegebenenfalls erlassen wird, eine wesentliche Rolle für die Erfüllung der Pflichten des Verantwortlichen spielen kann. Müssen Gewichtungen angegeben werden, bedarf es zumindest einer Erklärbarkeit des gesamten Modells. Unter Umständen werden durch die Offenlegung allgemeiner Informationen hinsichtlich der relevanten Faktoren und deren Gewichtung auch gar keine Geschäftsgeheimnisse berührt, da mit den zu erteilenden Informationen ein Reverse Engineering der vollständigen Modellarchitektur nicht ermöglicht wird. Oft sind es eher die Trainingsdatensätze, die für die Unternehmen besonders wertvoll sind.

cc) Recht auf Erklärung gem. Art. 22 Abs. 3 DS-GVO

Nach Art. 22 Abs. 3 DS-GVO ist der Verantwortliche außerdem verpflichtet, angemessene Maßnahmen zu treffen, um die Rechte und Freiheiten sowie die berechtigten Interessen der betroffenen Person zu wahren. Umstritten ist, ob ein Recht auf Erläuterung der Entscheidung (Ex-post-Erklärung) eine solche angemessene Maßnahme darstellt[26]. Die Art. 29 Datenschutzgruppe befürwortet dies, da die betroffene Person die Gründe für die Entscheidung benötigt, um ihren Standpunkt vorzutragen und die Entscheidung anzufechten[27]. Diesem Ansatz folgend müssten

---

[22] *Dix*, in: *Simitis/Hornung/Spiecker*, Datenschutzrecht, 1. Aufl. 2019, Art. 13 Rn. 16 m.w.N.
[23] BGH, Urteil vom 28.1.2014 – VI ZR 156/13.
[24] *Buchner*, in *Kühling/Buchner*, DS-GVO BDSG, 2. Aufl. 2018, Art. 22 Rn. 35; *Franck*, in: *Gola*, DS-GVO, 2. Aufl. 2018, Art. 13, Rn. 28.
[25] *Bäcker*, in: *Kühling/Buchner*, DS-GVO BDSG, 2. Aufl. 2018, Art. 13 Rn. 54.
[26] *Kumkar/Roth-Isigkeit*, JZ 2020, 281; *Wachter/Mittelstadt/Floridi,* International Data Privacy Law 2017, 81 ff.
[27] Art. 29 Data Protection Working Party, Guidelines on Automated Individual Decision Making for the Purposes of Regulation 2016/679, WP 251 Rev. 01, S. 27, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, Zugriff am 4.6.2020; a.A. z.B. *Kumkar/Roth-Isigkeit*, JZ 2020, 281; *Wachter/Mittelstadt/Floridi,* International Data Privacy Law 2017, 81 ff.

Erklärungen einer Einzelentscheidung, z.B. mittels modellagnostischer Post-hoc-Analysen, möglich sein.

b) Nachvollziehbarkeitserfordernisse aus dem Fairness-Grundsatz

Unabhängig vom Vorliegen einer automatisierten Einzelentscheidung können sich Nachvollziehbarkeitserfordernisse aus dem Fairnessgrundsatz gem. Art. 5 Abs. 1 lit. a DS-GVO ergeben, wonach die personenbezogenen Daten nach Treu und Glauben verarbeitet werden müssen. Generell versteht man darunter eine Rücksichtnahmepflicht, wobei auf die „vernünftigen Erwartungen" der betroffenen Person abzustellen ist. So ist nach EG 71 S. 6 a.E. DS-GVO durch technische und organisatorische Maßnahmen sicherzustellen, dass es aufgrund von Rasse, ethnischer Herkunft, politischer Meinung, Religion oder Weltanschauung, Gewerkschaftszugehörigkeit, genetischer Anlagen oder Gesundheitszustand sowie sexueller Orientierung nicht zu diskriminierenden Wirkungen oder zu Maßnahmen mit gleicher Wirkung kommt. Noch ist unklar, ob und wie sich der Fairness-Grundsatz aus Art. 5 Abs. 1 lit. a DS-GVO als Anker für ein Verbot diskriminierender Maßnahmen herauskristallisieren wird. Ungerechtfertigte Benachteiligungen sind jedoch schon aus ethischen Gründen zu vermeiden. Bestimmte Benachteiligungen sind nach dem Allgemeinen Gleichbehandlungsgesetz (AGG) sogar illegal. Auf das AGG und die Bedeutung der Nachvollziehbarkeit von KI-Anwendungen zur Erkennung rechtswidriger Diskriminierungen wird im folgenden Abschnitt eingegangen.

3. Allgemeines Gleichbehandlungsgesetz

Bei der Erbringung von Gesundheitsdienstleistungen ist eine Benachteiligung gemäß § 19 Abs. 2 AGG aus Gründen der Rasse oder wegen der ethnischen Herkunft i.R.v. medizinischen Behandlungsverhältnissen unzulässig[28]. Das in § 19 Abs. 1 Nr. 1 AGG enthaltene umfassendere Benachteiligungsverbot, das auch Benachteiligungen aus Gründen des Geschlechts, der Religion, einer Behinderung, des Alters oder der sexuellen Identität untersagt, ist nicht anwendbar, da es sich bei personalisierten Gesundheitsdienstleistungen nicht um Massengeschäfte oder

[28] Arzt- und Behandlungsverträge werden von § 2 Abs. 1 Nr. 5 AGG erfasst (*Schlachter*, in: Erfurter Kommentar zum Arbeitsrecht, 20. Aufl. 2020, § 2 AGG Rn. 13).

massengeschäftsähnliche Schuldverhältnisse i.S.d. Norm handelt[29]. Unmittelbare Benachteiligungen i.S.d. § 19 Abs. 2 AGG sind nach dem AGG in keinem Fall zu rechtfertigen. Wird dem Anschein nach auf ein neutrales Merkmal abgestellt, welches aber mittelbar bzw. versteckt eine diskriminierende Wirkung aufweist, kommt es darauf an, ob das Abstellen auf das Merkmal sachlich gerechtfertigt ist und die Mittel angemessen und erforderlich sind (§ 3 Abs. 2 AGG). Auf eine Benachteiligungsabsicht kommt es nicht an[30]. Das diskriminierende Merkmal muss lediglich mitursächlich für die Benachteiligung sein[31]. Daher bedarf es einer Möglichkeit nachzuvollziehen, welche Merkmale die Entscheidung eines KI-Systems beeinflusst haben.

Unzulässige Diskriminierungen können sich durch den Einsatz von KI-Anwendungen im medizinischen Bereich z.B. dadurch ergeben, dass ethnische Minderheiten aufgrund ihres geringeren Bevölkerungsanteils seltener in den Trainingsdaten vertreten sind. Dies kann zur Folge haben, dass Personen, die einer ethnischen Minderheit angehören, vom Algorithmus fehlerhaft einer Risikoklasse zugeordnet werden und sie somit nicht von einer verbesserten Behandlung profitieren können.

Die hier vorgestellten Modellarchitekturen bzw. modellagnostischen Methoden machen es auf Entwicklungsebene möglich, eine Überprüfung von KI-Systemen auf deren diskriminierende Wirkungen vorzunehmen, insofern, als dass sichtbar wird, ob und welchen Einfluss bestimmte Attribute auf die Entscheidung haben. Unzulässige Diskriminierungen aufgrund eines bestimmten Merkmals, wie der ethnischen Zugehörigkeit, kann man mit Interpretationsmethoden ermitteln, z.B. in dem überprüft wird, ob jene Attribute einen wesentlichen Einfluss auf den Modell-Output haben[32]. Indirekte Diskriminierungen durch stark mit unzulässigen Merkmalen korrelierte Attribute lassen sich mit diesen Methoden jedoch nicht unmittelbar erkennen. Dafür müssen zusätzlich die kausalen Beziehungen zwischen eingehenden Attributen und schützenswerten gruppenspezifischen Merkmalen bekannt sein. Werden unzulässige Diskriminierungen erkannt, muss die Modellarchitektur oder der Trainingsdatensatz angepasst werden. Besonders bei kontinuierlich weiterlernenden Modellen ist die

[29] *Mörsdorf*, in: BeckOK AGG, Stand 1.9.2020, § 19 Rn. 32; *Thüsing*, in: MüKoBGB, 8. Aufl. 2018, § 2 AGG Rn. 23.
[30] *Baumgärtner*, in: BeckOK AGG, Stand 1.9.2020, § 3 Rn. 49.
[31] BAG, Urteil vom 22.1.2009 – 8 AZR 906/07.
[32] Es existieren jedoch Methoden, die modellagnostische Post-hoc-Methoden durch die Maskierung des Bias täuschen können (*Slack/Hilgard/Jia/Singh/Lakkaraju*, Proceedings of the AAAI/ACM Conference on AI, Ethics and Society 2020, S. 180-186, https://dl.acm.org/doi/abs/10.1145/3375627.3375830, Zugriff am 16.10.2020).

Nachvollziehbarkeit des gesamten Modells höchst relevant, um schleichende Qualitätsminderungen oder einfließende Diskriminierungen (sog. Concept Drift) zu erkennen und zu vermeiden.

VII. Fazit

KI spielt eine immer größere Rolle bei der Entwicklung neuer medizinischer Anwendungen. Dabei erweist sich die Nachvollziehbarkeit der zugrundeliegenden KI-Technologien als bedeutender Faktor für die Einhaltung gesetzlicher Vorgaben des Medizinprodukterechts, des Datenschutzrechts sowie des AGG. Zwar wird die Nachvollziehbarkeit nicht direkt gefordert, sie ergibt sich jedoch indirekt bzw. erleichtert sie die Erfüllung gesetzlicher Erfordernisse, wie z.B., wenn es um die Zuverlässigkeit KI-gestützter Medizinprodukte geht. Die Vorhersage des mit dem Einsatz des Medizinprodukts verbundenen Risikos gestaltet sich effektiver, wenn innere Vorgänge im Modell verstanden werden können und man sich nicht nur auf statistische Werte verlassen muss, wobei sicherlich andere Faktoren, wie die Auswahl der Trainingsdatensätze, auch eine erhebliche Rolle spielen. Die zukünftigen Entwicklungen auf dem Feld „Explainable AI“ sind deshalb auch aus juristischer Perspektive und auch über den medizinischen Bereich hinaus interessant und bedürfen stetiger Beobachtung. Die vorgestellten anvisierten KI-gestützten medizinischen Anwendungen aus dem Projekt LOTTE haben durch die vorgesehenen Modellarchitekturen bzw. durch Verwendung modellagnostischer Nachvollziehbarkeitsmethoden bei Sicherstellung einer repräsentativen Datenbasis Potential, ein ausreichendes Maß an Nachvollziehbarkeit zu erreichen, um zuverlässig und fair zu agieren.