Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme

Technologieprogramm des Bundesministeriums für Wirtschaft und Energie

ERKLÄRBARE KI

Anforderungen, Anwendungsfälle und Lösungen

Studie im Auftrag des Bundesministeriums für Wirtschaft und Energie (BMWi) im Rahmen der Begleitforschung zum Technologieprogramm „Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme“ (KI-Innovationswettbewerb)

www.ki-innovationen.de

IMPRESSUM

Die Studie „Erklärbare KI -Anforderungen, Anwendungsfälle und Lösungen“ wurde durch die Begleitforschung zum KI-Innovationswettbewerb im Auftrag des Bundesministeriums für Wirtschaft und Energie erstellt und im April 2021 veröffentlicht.

Herausgeber

Technologieprogramm KI-Innovationswettbewerb
des Bundesministeriums für Wirtschaft und Energie
Begleitforschung
iit-Institut für Innovation und Technik in der VDI/VDE Innovation + Technik GmbH
Dr. Steffen Wischmann
Steinplatz 1
10623 Berlin
wischmann@iit-berlin.de

Autor:innen

Dr. Tom Kraus
Lene Ganschow
Marlene Eisenträger
Dr. Steffen Wischmann

Gestaltung

LoeschHundLiepold Kommunikation GmbH
Hauptstraße 28
10827 Berlin
KI-Innovationswettbewerb@lhlk.de

Stand

April 2021

Bilder

peshkov (Titel, S. 6), Yucel Yilmaz (S. 12, 17, 19) – stock.adobe.com

EXECUTIVE SUMMARY

Allein für Deutschland wird erwartet, dass mit Dienstleistungen und Produkten, die auf dem Einsatz von Künstlicher Intelligenz (KI) basieren, im Jahr 2025 Umsätze in Höhe von 488 Milliarden Euro generiert werden – damit würde ein Anteil von 13 Prozent am Bruttoinlandsprodukt erreicht. Dabei ist die Erklärbarkeit von Entscheidungen, die durch KI getroffen werden, in wichtigen Anwendungsbranchen eine Voraussetzung für die Akzeptanz bei den Nutzenden, für Zulassungs- und Zertifizierungsverfahren oder das Einhalten der durch die DSGVO geforderten Transparenzpflichten. Die Erklärbarkeit von KI-Produkten gehört damit, zumindest im europäischen Kontext, zu den wichtigen Markterfolgsfaktoren.

Die vorliegende Studie wurde durch die Begleitforschung zum Innovationswettbewerb „Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme“ (KI-Innovationswettbewerb) im Auftrag des Bundesministeriums für Wirtschaft und Energie erstellt. Die Studie basiert auf den Ergebnissen einer Online-Umfrage sowie Tiefeninterviews mit KI-Expert:innen aus Wirtschaft und Wissenschaft. Die Studie fasst den aktuellen Stand der Technik und zum Einsatz von erklärbarer KI (Explainable Artificial Intelligence, XAI) zusammen und erläutert ihn anhand praxisnaher Use Cases.

Den Kern von KI-basierten Anwendungen – womit hier im Wesentlichen Anwendungen des maschinellen Lernens gemeint sind – bilden immer die jeweils zugrundeliegenden KI-Modelle. Diese lassen sich in zwei Klassen einteilen: White- und Black-Box-Modelle. White-Box-Modelle, wie bspw. auf nachvollziehbaren Eingangsgrößen basierende Entscheidungsbäume, erlauben das grundsätzliche Nachvollziehen ihrer algorithmischen Zusammenhänge; sie sind somit selbsterklärend in Bezug auf ihre Wirkmechanismen und die von ihnen getroffenen Entscheidungen. Bei Black-Box-Modellen wie neuronalen Netzen ist es aufgrund ihrer Verflechtung und Vielschichtigkeit in der Regel nicht mehr möglich, die innere Funktionsweise des Modells nachzuvollziehen. Zumindest für die Erklärung von Einzelentscheidungen (lokale Erklärbarkeit) können dann jedoch zusätzliche Erklärungswerkzeuge eingesetzt werden, um nachträglich die Nachvollziehbarkeit zu erhöhen. KI-Entwickler:innen können für Entscheidungserklärungen je nach den konkreten Anforderungen auf etablierte Erklärungswerkzeuge zurückgreifen, bspw. LIME, SHAP, Integrated Gradients, LRP, DeepLift oder GradCAM, die allerdings Expertenwissen voraussetzen. Für die Nutzenden existieren bislang nur wenig gute Werkzeuge, die intuitiv verständliche Entscheidungserklärungen liefern (Saliency Maps, Counterfactual Explanations, Prototypen oder Surrogat-Modelle).

Die Teilnehmenden der im Rahmen dieser Studie durchgeführten Umfrage verwenden populäre Vertreter von White-Box-Modellen (statistische/probabilistische Modelle, Entscheidungsbäume) und Black-Box-Modellen (neuronale Netze) heute ungefähr in gleichem Umfang. Für die Zukunft wird jedoch gemäß der Umfrage eine stärkere Nutzung von Black-Box-Modellen erwartet, insbesondere von neuronalen Netzen. Damit wird die Bedeutung von Erklärungsstrategien in Zukunft weiter zunehmen, während sie schon heute essenzieller Bestandteil vieler KI-Anwendungen sind. Die Bedeutung von Erklärbarkeit ist dabei branchenabhängig sehr verschieden. Als mit Abstand am wichtigsten gilt sie im Gesundheitsbereich, gefolgt von der Finanzwirtschaft, dem Produktionssektor, der Bauwirtschaft und der Prozessindustrie.

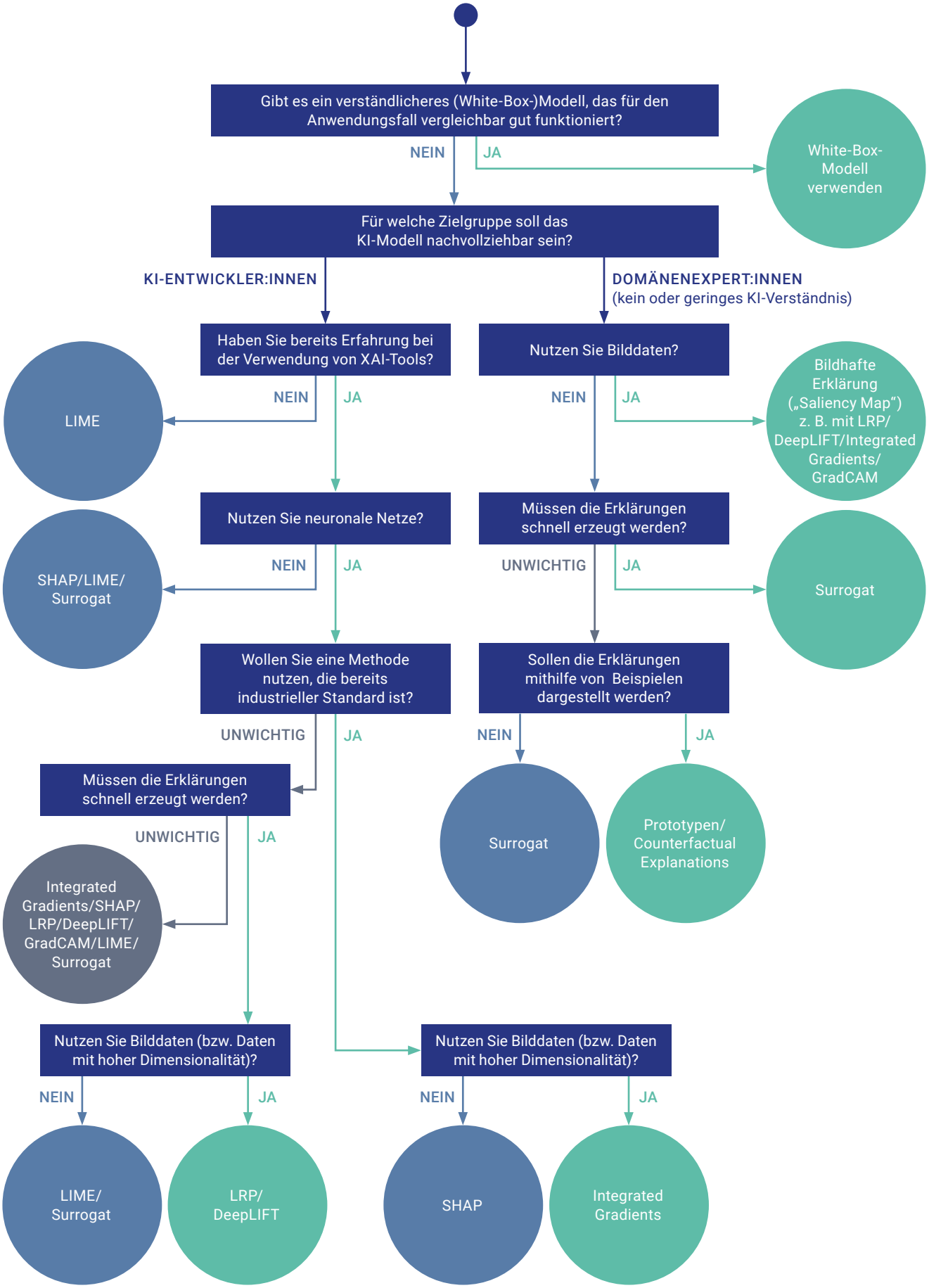
Durch Tiefeninterviews mit ausgewiesenen Expert:innen wurden vier Use-Cases genauer analysiert: die Bildanalyse histologischer Gewebeschnitte und die Textanalyse von Arztbriefen aus der Gesundheitswirtschaft, die Maschinenzustandsüberwachung im Produktionssektor sowie die Prozessführung in der Prozessindustrie. Modellerklärungen, die modellinterne Wirkmechanismen nachvollziehbar machen (globale Erklärbarkeit), sind lediglich bei der Prozessführung als strikte Zulassungsvoraussetzung unverzichtbar. In den andern Use-Cases reicht lokale Erklärbarkeit als Minimalanforderung aus. Globale Erklärbarkeit spielt jedoch in beiden betrachteten Use Cases der Produktionswirtschaft eine Schlüsselrolle für die Akzeptanz von KI-gestützten Produkten.

Die Use-Case-Analysen zeigen darüber hinaus, dass die Auswahl einer geeigneten Erklärungsstrategie von den Zielgruppen der Erklärung, den verwendeten Datentypen und dem verwendeten KI-Modell abhängt. Die Studie analysiert die Vor- und Nachteile der etablierten Werkzeuge entlang dieser Kriterien und bietet eine daraus abgeleitete Entscheidungshilfe an (siehe Abbildung 1). Da White-Box-Modelle selbsterklärend sind in Bezug auf Modellwirkmechanismen und Einzelentscheidungen, sollten sie bei allen Anwendungen, die hohe Anforderungen an Nachvollziehbarkeit stellen – wenn immer möglich – bevorzugt werden. Vor allem dann, wenn sie im Vergleich zu Black-Box-Modellen ähnlich gut, oder zumindest hinreichend gut, funktionieren.

Es ist davon auszugehen, dass mit dem zunehmenden Einsatz von KI in der Wirtschaft auch der Bedarf an verlässlichen und intuitiven Erklärungsstrategien in Zukunft stark steigen wird. Um diesen Bedarf zu bedienen, gibt es aktuell folgende technische und nichttechnische Herausforderungen, die es zu bewältigen gilt:

- Neu- und Weiterentwicklung geeigneter „hybrider" Ansätze, die daten- und wissensgetriebene Ansätze, bzw. White- und Black-Box-Modellierungsansätze, kombinieren
- Berücksichtigung von verhaltens- bzw. kognitionswissenschaftlichen Aspekten für erklärbare KI, wie Messbarkeit der Erklärung, Erklärbarkeit ganzheitlicher KI-Systeme, automatisierte Erklärungsanpassungen an Nutzer:innen
- Festlegung von Anwendungs- und Risikoklassen, aus denen das grundsätzliche Erfordernis einer Erklärung ableitbar ist
- Definition einheitlicher Anforderungen an die Erklärbarkeit von KI und damit das Schaffen klarer regulatorischer Vorgaben und Zulassungsrichtlinien entlang der zu definierenden Anwendungs- und Risikoklassen
- Schaffung von Zulassungs- und (Re-)Zertifizierungsrahmen für kontinuierlich lernende Systeme im produktiven Einsatz
- Bereitstellung und Umsetzung umfassender Aus- und Weiterbildungsmaßnahmen für Prüfer:innen zur Überprüfung der Erklärbarkeit von KI

Orientierungshilfe für den Einsatz der gängigsten Strategien und Werkzeuge für erklärbare KI („XAI-Tools“).



INHALT

1 EINLEITUNG

1 EINLEITUNG

Anwendungen Künstlicher Intelligenz (KI) liegen heute zumeist Algorithmen, Verfahren und Modelle zugrunde, die vielschichtig und verflochten sind. Dies hat zur Folge, dass die Entscheidungsfindung der KI in vielen Fällen für den Menschen nicht mehr nachvollziehbar ist – die Entwickler:innen der KI eingeschlossen.

Während dieser Umstand in manchen Anwendungsgebieten, wie etwa bei Produktempfehlungen im Unterhaltungsbereich, als nicht störend empfunden wird, kann die Nachvollziehbarkeit anderenorts sehr entscheidend sein, wenn es darum geht, KI-Produkte in der Praxis einzusetzen: Zum einem immer dann, wenn ein gewisses Maß an „Erklärbarkeit" algorithmischer Systeme für zuständige Zulassungsbehörden unverzichtbar ist, z. B. in der Gesundheitswirtschaft. Zum anderen, wenn das KI-Produkt ohne ein Mindestmaß an Erklärbarkeit von den jeweiligen Zielkund:innen nicht akzeptiert wird, etwa beim automatisierten Wertpapierhandel in der Finanzwirtschaft.

In welchem Ausmaß Erklärbarkeit für eine individuelle Zulassung oder Zertifizierung eines KI-gestützten Systems in Deutschland und Europa erforderlich ist, ist heute in vielen Anwendungsbranchen noch nicht abschließend geklärt – was für Unternehmen mit diesen Zielmärkten ein Innovationshemmnis darstellt. Angesichts prognostizierter Umsätze mit KI-basierten Dienstleistungen und Produkten in Höhe von 488 Milliarden Euro für das Jahr 2025 (eco - Verband der Internetwirtschaft e.V. 2019) ist dies auch von volkswirtschaftlicher Relevanz. Die Europäische Kommission, die mit Blick auf einen zukünftigen Rechtsrahmen für KI einen „risikobasierten Ansatz" verfolgt, vertritt die Sichtweise, gesetzliche Erklärbarkeitsanforderungen sollten vorrangig von der Kritikalität der Anwendung abhängen (European Commission 2020). Eine konkrete Bestimmung dessen, was KI-Systeme mit hohem Risikopotenzial genau charakterisiert und welcher Grad der Erklärbarkeit angemessen ist, steht jedoch von Seite der EU-Kommission noch aus.

In welchem Ausmaß Erklärbarkeit für eine individuelle Zulassung oder Zertifizierung eines KI-gestützten Systems in Deutschland und Europa erforderlich ist, ist heute in vielen Anwendungsbranchen noch nicht abschließend geklärt – was für Unternehmen mit diesen Zielmärkten ein Innovationshemmnis darstellt.

Wenn mögliche Fehler eines KI-Systems mit potenziell schweren oder fatalen Konsequenzen für Leib und Leben von Personen verbunden sind, wie z. B. im Gesundheitsbereich, dann muss eine „gewisse" Erklärbarkeit faktisch bereits heute sichergestellt werden, um die Grundanforderungen für eine Zulassung KI-gestützter Produkte zu erfüllen. Jedoch liegt diesbezüglich vieles im Ermessensspielraum der zulassenden Behörden, da bisher keine eindeutigen Anforderungen an Erklärbarkeit aus den Gesetzen ableitbar sind. Die Medizinprodukteverordnung formuliert zwar Anforderungen etwa zum „Risikomanagement", macht aber gleichzeitig keine Angaben, was darunter in Bezug auf Erklärbarkeit konkret zu verstehen ist. Im Gesundheitsbereich und in vielen anderen Anwendungsbranchen, z. B. beim autonomen Fahren und in der Finanzwirtschaft, besteht daher aktuell ein hoher Bedarf an Konkretisierungen, die einerseits technologieneutral formuliert sein sollten und andererseits offene Fragen in Bezug auf lernende Systeme klären müssen.

Für weniger kritische KI-Anwendungen, etwa Musik- oder Filmempfehlungen auf Unterhaltungsplattformen, bestehen keine zulassungsseitigen Vorgaben bezüglich der Erklärbarkeit. Hier entscheidet allein die Akzeptanz der Kund:innen darüber, ob ein KI-Produkt genutzt wird. Jedoch wird eine gewisse Nachvollziehbarkeit verstärkt auch vom Markt bzw. den Anwender:innen gefordert – wenn auch bislang weniger im Consumer-Bereich als im B2B[1]-Segment. Der Bedarf für erklärbare KI auf Unternehmensseite zeigt sich hier, wenn durch Fehlentscheidungen von KI-Systemen potenziell hoher wirtschaftlicher Schaden droht (z. B. bei der Instandhaltungsplanung von teuren Maschinen oder Anlagen). Im europäischen bzw. deutschen Consumer-Markt ist eine mittel- bis langfristige Erhöhung der Nachfrage für erklärbare KI grundsätzlich auch vorstellbar. Entsprechende technische Fortschritte

1 „Business-to-Business": Geschäftsbeziehungen zwischen zwei oder mehr Unternehmen

in Kombination mit der Tatsache, dass die Datenschutzgrundverordnung Transparenzpflichten gesetzlich verankert hat, können das Bürgerbewusstsein dahingehend verändern oder stärken. Die vorliegende Studie wurde durch die Begleitforschung zum Innovationswettbewerb „Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme" (KI-Innovationswettbewerb) im Auftrag des Bundesministeriums für Wirtschaft und Energie erstellt. Die Studie richtet sich an Anbieter:innen und Entwickler:innen von Systemen, die Produkte auf Basis von KI bereitstellen möchten und heute vor der Frage stehen, welche Anforderungen an die Erklärbarkeit eines Systems bestehen und wie diese adressiert werden können.

Ziele der Studie sind die Einordnung und Begriffsbestimmung einer erklärbaren KI, die Bereitstellung der Vor- und Nachteile von etablierten Erklärungsstrategien, die Analyse des aktuellen Einsatzes von erklärbarer KI in Wirtschaft und Wissenschaft, die Veranschaulichung anhand praktischer Use Cases, die Bereitstellung einer Orientierungshilfe zur Auswahl von Erklärungsstrategien sowie schließlich die Identifikation von Herausforderungen und Handlungsbedarfen für die Realisierung erklärbarer KI.

Methodik der Studie

Die Grundlage der Studie bildet eine Umfrage unter 209 Vertreter:innen aus Wirtschaft und Wissenschaft mit Bezug zum Thema Künstliche Intelligenz, eine Reihe durchgeführter Interviews mit Expert:innen sowie eine ausführliche Literaturrecherche. Die Teilnehmenden wurden aus den Reihen der Mitglieder des KI-Bundesverbands e.V. und aus den Projekten der BMWi-Technologieprogramme KI-Innovationswettbewerb, Smarte Datenwirtschaft, PAiCE sowie Smart Service Welten rekrutiert.

An der Umfrage, die als online ausfüllbarer Multiple-Choice-Fragebogen umgesetzt wurde, nahmen von Juli bis Oktober 2020 insgesamt 209 Personen teil (72 Prozent Unternehmensangehörige, 26 Prozent Vertreter:innen aus der Wissenschaft und zwei Prozent „Sonstige"). Von den Unternehmensvertreter:innen ordneten sich 70 Prozent einem kleinen oder mittleren Unternehmen zu (KMU, mit höchstens 250 Mitarbeitenden) und 30 Prozent einem Großunternehmen. Als KI-Anbieter:innen bzw. KI-Entwickler:innen bezeichneten sich 77 Prozent. Etwa 23 Prozent gaben an, KI-Anwender:innen oder KI-Nutzende zu sein.

Zuordnung der Teilnehmenden nach Ziel- und Anwendungsbranchen (Mehrfachnennung war möglich)*

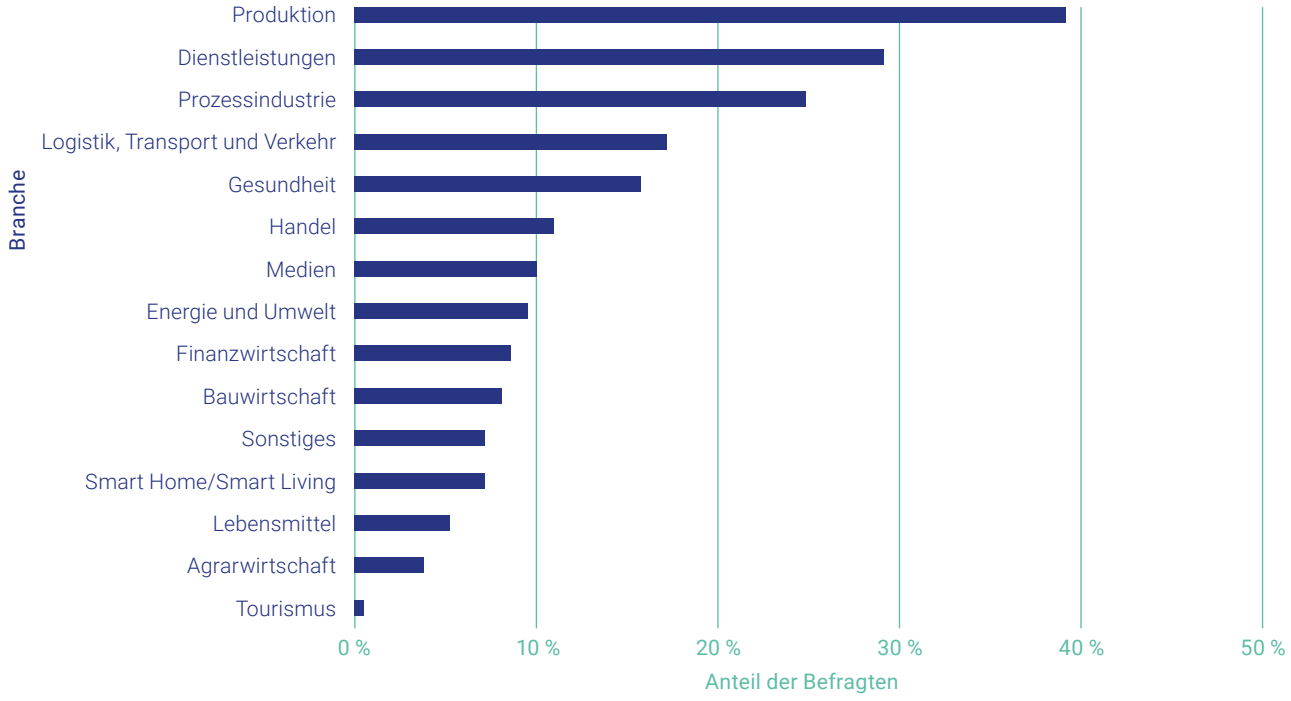
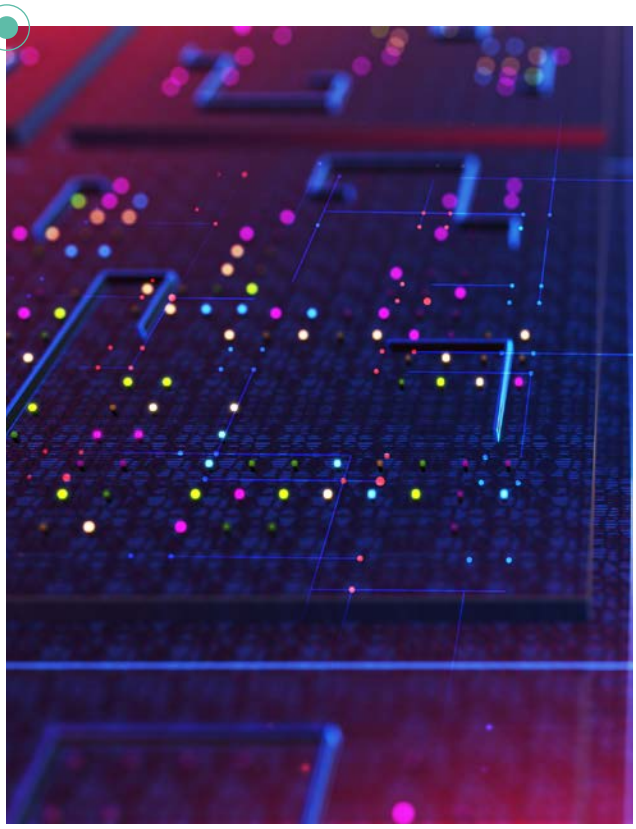


Abbildung 2 – Zuordnung der Teilnehmenden nach Ziel- und Anwendungsbranchen; n=209

* Einige nicht im Fragebogen auswählbare Branchen, z. B. IT/Software oder Tätigkeitsfelder im öffentlichen Sektor, wurden von mehreren Personen als Ziel- und Anwendungsbranche angegeben und in der Abbildung unter „Sonstiges" eingeordnet.

Mehrere Fragen der Umfrage wurden in Abhängigkeit von der bzw. den Ziel- bzw. Anwendungsbranchen gestellt, die die Teilnehmenden angegeben haben (siehe Abbildung 2). Dieser Ansatz wurde gewählt, weil für die befragten Personen der Bezug für eine Beurteilung so klarer wird und bei Branchenvergleichen nur Einschätzungen von Personen mit Domäneneinsicht berücksichtigt werden. Bei Branchen mit weniger Teilnehmer:innen in der Befragung sind die Ergebnisse wegen der kleinen Fallzahl weniger belastbar.

Die Interviews mit den Expert:innen wurden von November 2020 bis Februar 2021 durchgeführt. Die Interviewpartner:innen haben alle einen professionellen Bezug zur Künstlichen Intelligenz und decken die Bereiche Forschung, Wirtschaft, Standardisierung/Normung und Zulassung ab. In individuellen Gesprächen per Webkonferenz wurden mit den Fachleuten leitfragengestützte offene Interviews zu den Themen „technische Umsetzung von erklärbarer KI", „etablierte Erklärungswerkzeuge" sowie „technische und regulatorische Herausforderungen" geführt.



Übersicht über die Studie

Die Studie gliedert sich wie folgt:

- Aus der einschlägigen Literatur werden die bedeutendsten Konzepte zusammengetragen, um den Zugang zum fachlichen Diskurs zu erleichtern. Kapitel 2 bestimmt die zentralen Begriffe wie Transparenz, White-Box- und Black-Box-Modelle sowie Entscheidungs- und Modellerklärungen bzw. lokale und globale Erklärbarkeit.

- Die etablierten Erklärungsstrategien und -werkzeuge, die den Stand der Technik repräsentieren, werden in Kapitel 3 vorgestellt und bezüglich ihrer Einsatzmöglichkeiten und ihres praktischen Nutzen diskutiert.

- Wichtige Anwendungsfelder in Bezug auf Anwendungsbranchen, Modell- und Verfahrenskategorien, Zielgruppen sowie Datentypen und Umsetzungsmöglichkeiten wurden im Rahmen der durchgeführten Umfrage ermittelt. Die Ergebnisse werden in Kapitel 4 vorgestellt und diskutiert.

- Anhand von vier Use Cases aus der Gesundheitswirtschaft (Bildanalyse histologischer Gewebeschnitte, Textanalyse von Arztbriefen), Produktion (Maschinenzustandsüberwachung) und Prozessindustrie (Prozessführung) werden übergeordnete Ziele und konkrete Erklärbarkeitsanforderungen aus Sicht relevanter Zielgruppen identifiziert, miteinander verglichen und entsprechende Lösungswege beschrieben. Die Beschreibungen der Use Cases findet sich in Kapitel 5.

- Im Rahmen der Expert:inneninterviews wurden Vor- und Nachteile sowie Anwendungsfelder von etablierten Erklärungsstrategien und -werkzeugen besprochen. Daraus wurde eine kompakte Orientierungshilfe in Form eines Entscheidungsbaums generiert, der sich in Kapitel 6 findet.

- Wesentliche technische und regulatorische Herausforderungen und Handlungsbedarfe für die Realisierung erklärbarer KI-Systeme wurden im Rahmen leitfragengestützter Interviews mit Expert:innen ermittelt und in Kapitel 7 diskutiert.

Das Autor:innenteam möchte sich hier herzlich bei den Expertinnen und Experten bedanken, die sich für die Interviews zur Verfügung gestellt haben. Gleichzeitig danken wir dem Bundesverband KI e.V., und hier insbesondere seinem Geschäftsführer Daniel Abbou, für die Kooperation bei der Ansprache der Mitglieder sowie allen Teilnehmenden der Umfrage. Die Verantwortung für sämtliche inhaltliche Aussagen in dieser Studie liegt ausschließlich beim Team der Autorinnen und Autoren.

Die Autorinnen und Autoren bedanken sich herzlich bei den Expert:innen für die Teilnahme an den Interviews:

- Dr. Tarek Besold, DEKRA Digital GmbH
- Dr. Richard Büssow, Industrial Analytics IA GmbH
- Christian Geißler, Technische Universität Berlin
- Prof. Dr. Martin Hirsch, Universität Marburg, Ada Health GmbH
- Prof. Dr. Marco Huber, Universität Stuttgart, Fraunhofer IPA
- Prof. Dr. Alexander Löser, Forschungszentrum Data Science, Beuth Hochschule für Technik Berlin
- Prof. Dr. Axel-Cyrille Ngonga Ngomo, Universität Paderborn
- Dr. Christoph Peylo, Bosch Center for Artificial Intelligence, Robert Bosch GmbH
- Prof. Dr. Philipp Rostalski, Universität zu Lübeck
- Prof. Dr. Ute Schmid, Universität Bamberg, Fraunhofer IIS
- Gerald Spyra, Ratajczak & Partner mbB
- Thomas Staufenbiel, Gestalt Robotics GmbH
- Martin Tettke, Berlin Cert GmbH
- Prof. Dr. Leon Urbas, Technische Universität Dresden
- Betty van Aken, Forschungszentrum Data Science, Beuth Hochschule für Technik Berlin

Zudem bedanken wir uns beim Bosch Center for Artificial Intelligence für bereitgestellte Informationen. Auch möchten wir uns beim Fraunhofer IPA, insbesondere bei Nina Schaaf, für die Zusammenarbeit in Bezug auf die inhaltliche Abstimmung und Abgrenzung zwischen der dort erstellten und der hier vorliegenden Studien und für die Diskussionen zum Thema Erklärbare Künstliche Intelligenz bedanken.

2 ERKLÄRBARKEIT VON KÜNSTLICHER INTELLIGENZ: ZIELE, EINORDNUNG UND BEGRIFFE

2 ERKLÄRBARKEIT VON KÜNSTLICHER INTELLIGENZ: ZIELE, EINORDNUNG UND BEGRIFFE

2.1 Übergeordnete Ziele und Einordnung erklärbarer Künstlicher Intelligenz

Die Motivation für die Entwicklung und Umsetzung von erklärbarer Künstlicher Intelligenz unterscheidet sich je nach Anwendungsfall und den Interessen der Zielgruppen einer erklärbaren KI. Dennoch können übergeordnete Ziele von erklärbarer KI formuliert werden – basierend auf (Arrieta et al. 2019) – und entweder einzeln oder in Kombination verfolgt werden:

1. **Kausalitätsbeziehungen plausibilisieren:** Mit erklärbarer KI sollen Muster, die durch KI entdeckt wurden, zusätzlich auf ihre Gültigkeit und Plausibilität geprüft werden. Eine häufige Motivation von Anwender:innen ist das „Finden" bzw. die Abbildung von kausalen Zusammenhängen.

2. **Übertragbarkeit testen:** KI-Modelle werden in der Regel trainiert, um eine spezifische Aufgabe zu lösen, z. B. „finde alle Bilder mit einer Katze". Erklärbarkeit soll helfen, die Übertragbarkeit der gefundenen Lösung auf neue Aufgaben abzuschätzen (z. B. „finde alle Hunde"). Dies hilft, den Anwendungsbereich und die Grenzen einer KI zu bestimmen.

3. **Informationsgewinn erhöhen:** Um eine Entscheidung eines KI-Systems überhaupt als Entscheidungsunterstützung nutzen zu können, sollen Informationen zur Grundlage der Entscheidung, etwa zur Funktionsweise des Modells, in verständlicher und einfacher, aber nicht zu sehr vereinfachter Form bereitgestellt werden.

4. **Konfidenz bestimmen:** Das KI-System soll auf Robustheit (Erhaltung von Gütekriterien des Systems bei Nichterfüllung von Annahmen oder statistischen Ausreißern), Stabilität (ähnliche Daten liefern ähnliche Ergebnisse) oder Reproduzierbarkeit (gleiches Ergebnis bei mehrmaliger Ausführung) überprüft werden, um Anfälligkeiten und Gültigkeitsbereiche zu identifizieren.

5. **Fairness testen:** Bei diesem Ziel soll mithilfe erklärbarer KI die Möglichkeit geschaffen werden, ein Modell auf Fairness zu überprüfen, insbesondere durch Aufdecken eines möglicherweise existierenden Bias (d.h. einen systematischen Fehler) in der Datengrundlage.

6. **Interaktionsmöglichkeiten verbessern:** Erklärbare KI soll die Nutzer:innen – insbesondere solche mit wenig KI-Expertise – dabei unterstützen, direkt mit dem KI-System zu interagieren, beispielsweise, um dessen Entscheidungsfindung oder die Verständlichkeit von Erklärungen zu verbessern, etwa über die Bereitstellung alternativer Erklärungen (Zusammenfassung von „Interactivity" und „Accessability" aus (Arrieta et al. 2019)).

7. **Privacy-Bewusstsein erhöhen:** Potenziell kann erklärbare KI der Nutzer:in einen Einblick in die erfassten und gespeicherten Daten ermöglichen und so zu einer verstärkten Wahrnehmung von Privacy-Aspekten führen.

8. **Verantwortlichkeiten klären:** Erklärbare KI kann zur Klärung von Verantwortlichkeiten und Haftungsfragen herangezogen werden. Es kann z. B. über den Weg eines oder einer Sachverständigen bzw. eines Gutachtens festgestellt werden, dass in ein System absichtlich in großem Umfang Bias-behaftete Daten eingeschleust wurden, um dieses zu beeinflussen.

Ein zusätzliches Ziel, das in der Literatur ebenfalls häufig als Motivation formuliert wird (Ribeiro et al. 2016; Arrieta et al. 2019), ist das Herstellen von Vertrauen („Trustworthiness"), das allerdings nicht mehr nur als übergeordnetes Ziel von Erklärbarkeit, sondern als generelles Entwicklungsziel von KI-Systemen gesehen wird.

Laut der Richtlinien der „High-Level Expert Group on AI" der Europäischen Union für eine vertrauenswürdige Künstliche Intelligenz („trustworthy AI") (High-Level Expert Group on AI 2019), trägt Erklärbarkeit bereits per se zur Herstellung von Vertrauen bei. Die sieben Säulen, die das Vertrauen in KI stützen sollen, sind:

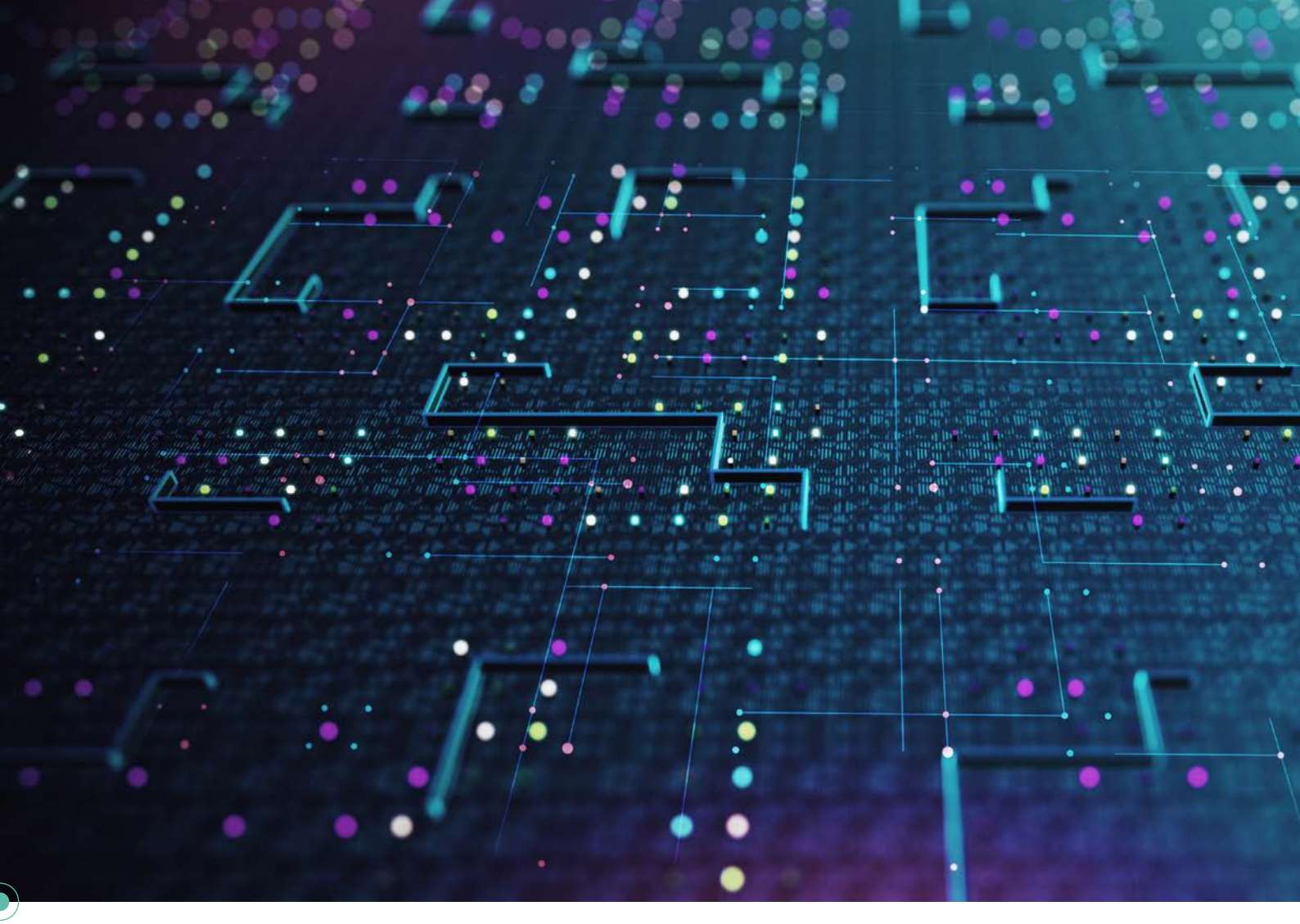
1. Vorrang menschlichen Handelns und menschlicher Aufsicht

2. Technische Robustheit und Sicherheit

3. Schutz der Privatsphäre und Datenqualitätsmanagement

4. Transparenz (z. B. Nachverfolgbarkeit, *Erklärbarkeit* und Kommunikation)

5. Vielfalt, Nichtdiskriminierung und Fairness



6. Gesellschaftliches und ökologisches Wohlergehen

7. Rechenschaftspflicht

Im Kontext der sieben Säulen bezeichnet „Transparenz" die Eigenschaft einer KI, über eine nachverfolgbare und vor allem erklärbare Entscheidungsfindung zu verfügen, die dem Nutzer bzw. der Nutzerin über entsprechende Informationen kommuniziert wird.

Gleichzeitig sind die Wirkmechanismen innerhalb der sieben Säulen nicht gänzlich unabhängig voneinander. Wenn Erklärbarkeit als einer der Kernaspekte der „Transparenz" gegeben ist, dann wirkt sich dies für viele Anwendungen auch unmittelbar auf die Punkte 1, 2 und 5 aus:

- Wenn beispielsweise KI-Systeme anfällig gegenüber Datenbias[1] in Realdaten sind, kann die Transparenz bei (teil-)autonomen Systemen eine menschliche Aufsicht enorm erleichtern. Auch im Falle reiner Entscheidungsunterstützungssysteme kann indirekt die Gefahr reduziert werden, dass Entscheidungen des Systems ohne ausreichende menschliche Prüfung umgesetzt werden.
- Aus Sicht von Entwickler:innen kann die Systemanfälligkeit gegenüber Datenbias durch eine entsprechende Transparenz besser adressiert und damit die technische Robustheit und Sicherheit von Systemen verbessert werden.
- Auch ist eine solche Transparenz eine Grundvoraussetzung, um die Gleichbehandlung von Personen zu ermöglichen und möglicherweise diskriminierende Entscheidungen algorithmischer Systeme zu identifizieren (Nichtdiskriminierung und Fairness[2]).

Erklärbarkeit ist darüber hinaus ein wichtiger Aspekt, der zur Akzeptanz von KI beitragen kann. Grundsätzlich bestimmt sich die Akzeptanz einer Technologie über deren freiwillige, aktive und zielgerichtete Nutzung.

Es existiert bis heute noch keine allgemein anerkannte, folglich auch keine einheitliche Taxonomie für erklärbare Künstliche Intelligenz, weshalb im folgenden Abschnitt genauer darauf eingegangen wird, welche Konzepte und Begrifflichkeiten in der Studie zugrunde gelegt werden.

1 Hier und im Folgenden bezieht sich der Ausdruck „Datenbias" bzw. „Bias" auf ein anwendungsneutrales Begriffsverständnis der Statistik, d. h., es ist eine allgemeine systematische Abweichung gemeint.

2 Das Testen von Fairness stellt dennoch – besonders aufgrund der Schwierigkeit, geeignete Metriken auszuwählen – fast immer eine Herausforderung dar.

2.2 Grundlegende Begriffe erklärbarer Künstlicher Intelligenz

Trotz teilweise widersprüchlicher Bezeichnungen besteht weitreichender Konsens (Lipton 2016; Gilpin et al. 2018; Arrieta et al. 2019) über die grundsätzliche Unterscheidung zweier Konzepte, nämlich

- Transparenz[3]
- und Erklärbarkeit (zumeist in Form von Post hoc-Erklärbarkeit).

Die Begrifflichkeiten in dieser Studie orientieren sich an (Arrieta et al. 2019), die Darstellung lehnt an (Lipton 2016) an.

Liegt Transparenz eines KI-Modells vor – wobei Transparenz hier explizit als Eigenschaft zu verstehen ist –, so kann es unter der Voraussetzung nachvollziehbarer Eingangsdaten auch als „White-Box"-Modell bezeichnet werden. Bei solchen Modellen sind insbesondere die algorithmischen Mechanismen zur Generierung des Modells nachvollziehbar. Eine detaillierte Begriffsbestimmung der Transparenz folgt in Abschnitt 2.2.1.

Bei der Erklärbarkeit geht es hingegen darum, einer Zielperson eine verständliche Begründung aktiv bereitzustellen, die es ihr ermöglicht, das Ergebnis eines KI-Modells nachzuvollziehen. Die Wahrnehmung und der Wissensstand der Zielperson, aber auch die Ausrichtung der Fragestellung, sind beim Erstellen von Erklärungen notwendigerweise zu berücksichtigen. Eine detaillierte Begriffsbestimmung folgt in Abschnitt 2.2.2.

Liegt Transparenz eines KI-Modells vor, so kann es unter der Voraussetzung nachvollziehbarer Eingangsdaten auch als „White-Box"-Modell bezeichnet werden. Bei der Erklärbarkeit geht es hingegen darum, einer Zielperson eine verständliche Begründung aktiv bereitzustellen, die es ihr ermöglicht, das Ergebnis eines KI-Modells nachzuvollziehen.

2.2.1 Transparenz

Transparenz wird im Folgenden als eine Modelleigenschaft behandelt. Ist die Transparenz eines Modells gegeben, so ist es unter der Annahme nachvollziehbarer Eingangsgrößen selbsterklärend[4]. Die Eigenschaft der Transparenz lässt sich weiter unterteilen in die drei unterschiedlichen Ausprägungen der „Simulierbarkeit", der „Unterteilbarkeit" und der „Algorithmischen Transparenz" (Lipton 2016). Dabei wird in der Literatur häufig von einer hierarchischen Abhängigkeit ausgegangen (Arrieta et al. 2019), sodass die Simulierbarkeit eines Systems dessen Unterteilbarkeit und dessen algorithmische Transparenz impliziert. Entsprechend begründet die Unterteilbarkeit eines Systems auch dessen algorithmische Transparenz. Folglich gilt ein Modell – unter der Annahme erklärbarer Eingangsdaten – bereits als transparent, wenn es lediglich die Eigenschaft der algorithmischen Transparenz erfüllt. Die höchste Transparenzstufe erreicht ein Modell, wenn es die Eigenschaft der Simulierbarkeit und somit auch die beiden anderen Eigenschaften erfüllt.

Ein System ist simulierbar, wenn auch eine Person die Entscheidungen des zugrundeliegenden Algorithmus in angemessener Zeit nachvollziehen kann oder könnte, indem sie die einzelnen Schritte, die zur Herbeiführung einer Entscheidung nötig sind, manuell durchführt.

Beispiel: Beim manuellen Durchlaufen unterschiedlicher Pfade eines nicht allzu großen Entscheidungsbaumes, der auf nachvollziehbaren Eingangsgrößen beruht, kann eine Person in jedem Knoten selbst überprüfen, ob eine individuelle Eigenschaft von Eingangsdaten bzw. ein Attribut erfüllt ist oder nicht. Gibt es keine Attribute mehr zu prüfen, hat die Person ein „Blatt" des Entscheidungsbaums erreicht, welches das Ergebnis repräsentiert.

3 Dabei bezeichnet der Begriff der Transparenz hier ein anderes Konzept als das, was zuvor in den Richtlinien für eine vertrauenswürdige Künstliche Intelligenz (High-Level Expert Group on AI 2019) referenziert wurde.

4 In der Literatur ist in diesem Zusammenhang auch häufig von „interpretierbaren" Modellen die Rede. Allerdings wurde dieser Begriff hier vermieden, da „Interpretierbarkeit" in der einschlägigen Literatur häufig widersprüchlich verwendet wird.

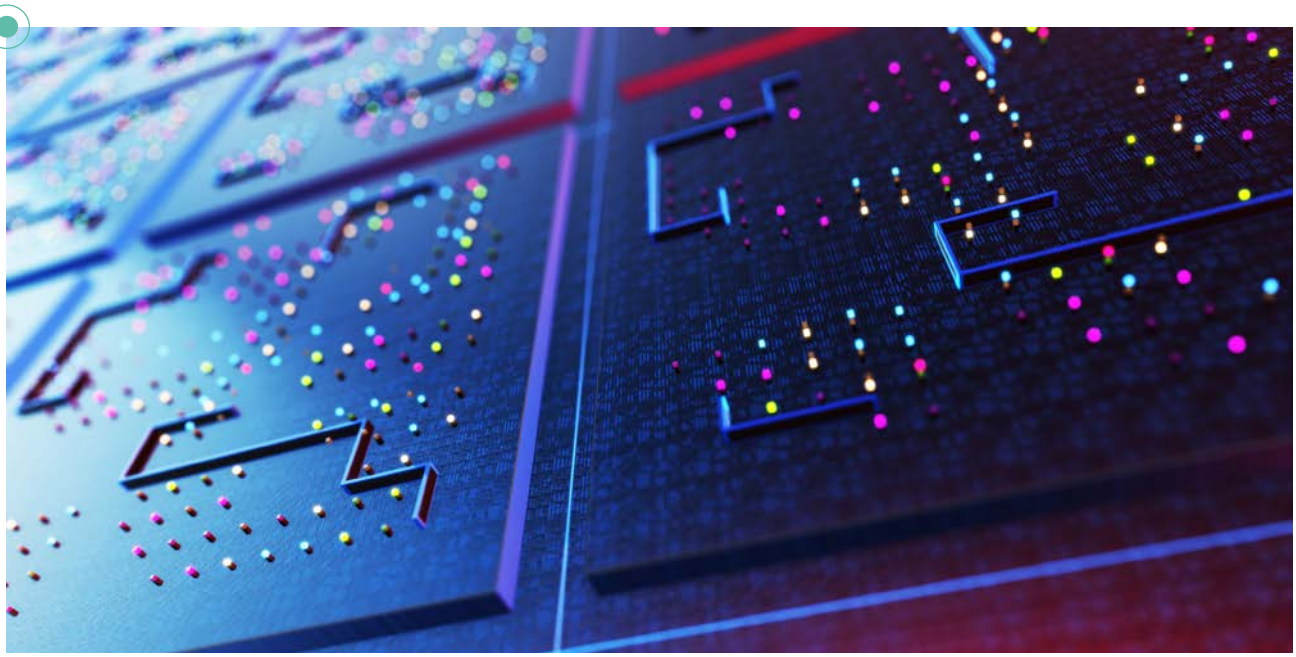
In einem unterteilbaren System können oder könnten die einzelnen Komponenten (Eingangsdaten, Parameter, Modellebenen, Berechnungen etc.) mit einer intuitiven Beschreibung versehen werden, sodass ihre Funktionen im Gesamtsystem nahvollziehbar sind.

Beispiel: In Entscheidungsbäumen ist für jeden Knoten festgelegt, welches Attribut gerade getestet wird und welche Ausprägung jeweils für die Auswahl der folgenden Teilpfade notwendig ist. Einzelne Modellebenen können über die Beschreibung der jeweils enthaltenen Knoten charakterisiert werden. Die intuitive Beschreibung und Nachvollziehbarkeit der Eingangsgrößen muss beim Systementwurf sichergestellt sein, um einen Entscheidungsbaum als unterteilbar bezeichnen zu können, d.h. die Eingangsgrößen dürfen selbst keine undurchsichtigen Konstrukte aus vielen Größen sein.

Algorithmische Transparenz bezieht sich auf den eigentlichen Lernprozess bzw. die Generierung von Modellen. Hierbei kommt es darauf an, ob nachvollzogen werden kann, wie ein Modell im Detail erzeugt wird und wie in der Trainingsphase mit möglichen Situationen umgegangen wird, mit denen der betreffende Algorithmus konfrontiert sein könnte (in Bezug auf unbekannte Input- bzw. Trainingsdaten). Dabei geht es bei der algorithmischen Transparenz tatsächlich nur um die Eigenschaften des Algorithmus und nicht um konkrete Modellausprägungen oder Daten.

Beispiel: Im Falle einer linearen Regression, bei der ein lineares Modell z. B. mittels Methode der kleinsten Quadrate (mathematisches Standardverfahren zur Ausgleichsrechnung) an eine aus Messwerten bestehende Punktwolke angepasst wird, ist im Detail nachvollziehbar, wie das eindeutige Ergebnis ermittelt wird. Bei bestimmten Annahmen zu statistischen Verteilungen kann man zusätzliche statistische Aussagen zum ermittelten Ergebnis treffen. Dabei ist das resultierende lineare Modell stets eindeutig, die Konvergenz ist verlässlich und die Grenzen sind wohlbekannt (z. B. Anfälligkeit der klassischen Methode der kleinsten Quadrate gegenüber statistischen Ausreißern).

Am Beispiel des Entscheidungsbaums und des linearen Regressionsmodells zeigt sich, dass diese KI-Modelle grundlegende Transparenzeigenschaften erfüllen. Bei anderen Modellen ist das schon für kleinere Modelle, die für praktische Anwendungen noch genutzt werden können, nicht mehr der Fall. Um etwa bei neuronalen Netzen in der Bildverarbeitung von einem transparenten System sprechen zu können, müsste für eine Unterteilbarkeit die Funktion jedes einzelnen Knotens und jeder einzelnen Schicht im Netz eindeutig beschreibbar und auf das Ergebnis beziehbar sein. So könnte ein Knoten „verantwortlich“ dafür sein, horizontale Linien im Bild zu erkennen, ein weiterer erkennt vertikale Linien etc. Diese Beschreibung müsste für das gesamte Netz vorhanden sein, sodass auch tiefere Schichten, die auf den Ergebnissen der vorherigen beruhen, erklärbar sind. Daher scheint die Eigenschaft der Unterteilbarkeit für neuronale Netze nicht erfüllbar zu sein. Entsprechend bewirken auch die Verflechtung und Vielschichtigkeit, die neuronale Netze in der Regel charakterisieren, dass



Modellvarianten neuronaler Netze – gemäß der obigen Begriffsbestimmung – als nicht simulierbar gelten. Die fehlende algorithmische Transparenz von neuronalen Netzen wird in der Literatur häufig daran festgemacht, dass gängige Trainingsmethoden für unbekannte Input- bzw. Trainingsdaten in der Regel nicht zu eindeutigen Lösungen führen (Lipton 2016). Dies liegt daran, dass die Oberflächen (engl. error oder loss surface) der Verlustfunktion angesichts der Problemstruktur zumeist schwer analysierbar sind (Arrieta et al. 2019; Kawaguchi 2016; Datta et al. 2016) und Lösungen nur approximativ mit heuristischen Optimierungsverfahren ermittelt werden können (Arrieta et al. 2019). Laut (Arrieta et al. 2019) ist generell die Zugänglichkeit eines Modells für entsprechende mathematische Analysen und Methoden das entscheidende Kriterium für die algorithmische Transparenz eines Modells.

Eine Diskussion der Transparenzeigenschaften weiterer Modelle in Bezug auf algorithmische Transparenz, Simulierbarkeit und Unterteilbarkeit findet sich in (Arrieta et al. 2019).

2.2.2 Erklärbarkeit

Da Transparenz für diverse Modelle wie z. B. neuronale Netze nicht erreichbar ist, diese folglich nicht selbsterklärend sind, kommt bei diesen das Konzept der „Erklärbarkeit“ zur Anwendung. Dabei ist zwar in der Regel festgelegt, ob die Erklärbarkeit eine Entscheidung oder ein Modell betrifft. Wie eine konkrete Erklärung dabei ausgestaltet ist oder wieviel Erkenntnis sie der Zielperson gewährt, bleibt dabei jedoch zunächst unbestimmt. Beim Beispiel der Bildverarbeitung mit neuronalen Netzen spricht man etwa bereits von einer Erklärung einer Entscheidung, wenn bestimmte Bereiche im Eingabebild, die zur Klassifikation eines konkreten Objektes geführt haben, für die anwendende Person farblich hervorgehoben werden. In diesem Fall wird nicht jeder einzelne Schritt des Algorithmus erklärt, sondern nur die für die Entscheidungsfindung bedeutsamsten Daten hervorgehoben. Alternativ kann eine Erklärung auch durch eine textliche Beschreibung repräsentiert werden, z. B. „Auf diesem Bild ist ein Hund abgebildet, da vier Beine, eine Schnauze, Fell und ein Schwanz erkannt wurden.“.

Übersicht über White-Box-/Black-Box-Charakter von Modellen, die für das maschinelle Lernen eingesetzt werden

KI-Modell	Transparenz			White-Box/ Black-Box	Post-hoc-Analyse notwendig?
	Simulierbarkeit	Unterteilbarkeit	Algorith. Transparenz		
Neuronale Netze	X	X	X	Black-Box	Notwendig: Werkzeuge in Kap. 3
Ensemble-Modelle (z. B. Tree Ensembles)	X	X	X	Black-Box	Notwendig: Werkzeuge in Kap. 3
Support Vector Machines	X	X	X	Black-Box	Notwendig: Werkzeuge in Kap. 3
Bayes-Netze	(✓)	(✓)	✓	White-Box*	Nicht notwendig
Lineare/logistische Regressionsmodelle	(✓)	(✓)	✓	White-Box*	Nicht notwendig
Entscheidungsbäume (Decision Trees)	(✓)	(✓)	✓	White-Box*	Nicht notwendig

Tabelle 1: Übersicht über White-Box-/Black-Box-Charakter von Modellen, die für das maschinelle Lernen eingesetzt werden (basierend auf (Arrieta et al. 2019)).

* Gilt im Falle nachvollziehbarer Eingangsparameter und generell bei Unterteilbarkeit.

Grundsätzlich unterscheidet man zwei Typen von Erklärungen:

- Erklärungen von Einzelentscheidungen bzw. Entscheidungserklärungen, die dabei helfen, individuelle, datenbezogene Entscheidungen konkret nachzuvollziehen (sogenannte lokale Erklärbarkeit oder Datenerklärbarkeit).

- Erklärungen von Modellen bzw. Modellerklärungen, die dabei helfen, Wirkzusammenhänge von KI-Modellen zu begreifen (sogenannte globale Erklärbarkeit oder Modellerklärbarkeit), z. B. lineare oder allgemein funktionale Zusammenhänge zwischen Eingangs- und Ausgangsgrößen.

Es handelt sich um eine „Post hoc"-Erklärung, wenn ein geeignetes Analysewerkzeug „im Nachhinein" angewandt wird, um eine Erklärung zu generieren – d. h. nach der Entscheidungsfindung im Falle von Entscheidungserklärungen oder nach dem Modelltraining (bei Modellerklärungen). Post hoc-Erklärungen können theoretisch unabhängig davon, ob es sich um ein transparentes oder ein „undurchsichtiges" Modell handelt, erzeugt werden – zumindest, wenn das eingesetzte Analysewerkzeug, von denen in Kapitel 3 eine ganze Reihe vorgestellt werden, entsprechend flexibel ist. In der Regel werden Erklärungen aber nur benötigt, um für undurchsichtige Modelle („Black-Box") trotzdem eine gewisse Nachvollziehbarkeit herstellen zu können.

2.2.3 White- und Black-Box-Modelle

Anhand des dargestellten Paradigmas der Transparenz lassen sich Modelle der Klasse der Black-Box-Modelle zuordnen, wenn keine der drei Eigenschaften bzw. Stufen der Transparenz – Simulierbarkeit, Unterteilbarkeit und algorithmische Transparenz – erfüllt ist. Umgekehrt sollen Modelle, die zumindest die niedrigste der drei genannten Transparenzstufen (algorithmische Transparenz) erfüllen und nachvollziehbare Eingangsgrößen verwenden, im Folgenden als White-Box-Modelle bezeichnet werden.

Ein Überblick darüber, ob die zurzeit häufig verwendeten KI-Modelle die Eigenschaften von Simulierbarkeit, Unterteilbarkeit und algorithmischer Transparenz erfüllen und damit entsprechend unter der Annahme nachvollziehbarer Eingangsgrößen als White- oder Black-Box zu erachten sind, findet sich in Tabelle 1. Dabei folgt die Einordnung der einzelnen Modelle im Wesentlichen dem Konzept von (Arrieta et al. 2019).

Es zeigt sich, dass die Einteilung in White-Box- und Black-Box-Modelle mithilfe der Transparenzstufen sehr gut gelingt. Dabei kann jedes nominelle White-Box-Modell zwar die beiden Eigenschaften der Simulierbarkeit und der Unterteilbarkeit bei zu hoher Dimension oder komplex konstruierten Eingangsgrößen potenziell verlieren, die algorithmische Transparenz bleibt aber in jedem Fall erhalten. Dies unterscheidet sie entscheidend von Black-Box-Modellen, die bei Modellen niedrigster Dimension, die in der Anwendung noch von praktischem Nutzen sind, keine der drei Eigenschaften erfüllen – vor allem auch nicht die niedrigste Transparenzstufe der algorithmischen Transparenz. Ein interessanter Spezialfall sind Bayes-Netze. Diese Modellklasse hat den Vorteil, dass statistische Informationen zu Trainingsdaten („Dichte" von Trainingsdaten im Datenraum) in Konfidenzwerten berücksichtigt werden können. Das bedeutet, Bayes-Netze stellen nicht nur die Entscheidung selbst bereit, sondern liefern auch quantitative Aussagen mit, z. B. darüber, wie wahrscheinlich der Eintritt eines Ereignisses ist. Diese Eigenschaft wird auch gewahrt, wenn die Anforderungen für Simulierbarkeit und für Unterteilbarkeit nicht erfüllt sind.

Im Gegensatz zu selbsterklärenden White-Box-Modellen muss für Black-Box-Modelle – beispielsweise für neuronale Netze – eine zusätzliche Strategie genutzt werden, um das Modell nachvollziehbar zu gestalten bzw. es zu erklären. Dabei handelt es sich um eine „Post hoc"-Analyse, wenn ein geeignetes Erklärungswerkzeug im Nachhinein, also nach der Entscheidungsfindung bzw. nach dem Trainieren des KI-Modells, auf dieses angewendet wird. Im folgenden Kapitel werden verschiedene Erklärungsstrategien vorgestellt, von denen die meisten als Post hoc-Analysewerkzeuge bezeichnet werden können. Weitere diskutierte Ansätze ergänzen die KI-Modelle selbst um bestimmte Komponenten, die ermöglichen, dass aus den erweiterten Modellen Erklärungen extrahiert werden können.

3 VOR- UND NACHTEILE ETABLIERTER STRATEGIEN UND WERKZEUGE FÜR ERKLÄRBARE KI

3 VOR- UND NACHTEILE ETABLIERTER STRATEGIEN UND WERKZEUGE FÜR ERKLÄRBARE KI

In Hinblick auf Erklärungsstrategien kann zwischen Ansätzen unterschieden werden, die Modellerklärungen liefern, und Ansätzen, die Entscheidungserklärungen bereitstellen. Eine Modellerklärung gibt Aufschluss über die konkrete Funktionsweise des Modells. Eine Entscheidungserklärung liefert Gründe, die zu einer einzelnen Entscheidung des KI-Modells geführt haben. Ein KI-Modell kann per se selbsterklärend sein (White Box) oder aber – häufig aufgrund entsprechender Erweiterungen von Black-Box-Modellen – Erklärungen gleichzeitig mit der Entscheidung generieren. Alternativ können Erklärungen auch nach der Entscheidung (Post hoc) durch ein zusätzliches (Post hoc-)Analysewerkzeug bereitgestellt werden. Der letztgenannte Ansatz adressiert gezielt Black-Box-Modelle und die Verbesserung ihrer Nachvollziehbarkeit.

White-Box-Modelle – beispielsweise lineare und logistische Regressionsmodelle, Entscheidungsbäume oder Bayes-Netze – sind selbsterklärend in Bezug auf Modellwirkmechanismen (aufgrund ihrer direkt nachvollziehbaren Funktionsweise) und hinsichtlich ihrer Entscheidungen. Das White-Box-Modell kann folglich sowohl für die konkrete Aufgabe (Klassifikation, Regression oder Clustering) als auch für die Erklärungsbereitstellung verwendet werden.

Bei den im Folgenden vorgestellten Erklärungsstrategien handelt es sich nur um eine begrenzte Auswahl, in der Forschung und Praxis werden noch viele weitere Methoden eingesetzt und diskutiert. Die Auflistung umfasst die zehn Erklärungswerkzeuge, deren zugehörige wissenschaftliche Erstveröffentlichung laut Google Scholar mindestens 500 Zitierungen aufweist (Stand Dezember 2020), sowie etablierte Methoden, die von den Expertinnen und Experten im Rahmen der Interviews zusätzlich benannt wurden. Für jede hier vorgestellte Erklärungsstrategie wird eine kurze Erläuterung geliefert, die ein leicht verständliches Beispiel, eine kurze Beschreibung des technischen Hintergrunds, wichtige Vor- und Nachteile sowie Verweise auf weiterführende Literatur beinhaltet.

In Hinblick auf Erklärungsstrategien kann zwischen Ansätzen unterschieden werden, die Modellerklärungen liefern, und Ansätzen, die Entscheidungserklärungen bereitstellen. Eine Modellerklärung gibt Aufschluss über die konkrete Funktionsweise des Modells. Eine Entscheidungserklärung liefert Gründe, die zu einer einzelnen Entscheidung des KI-Modells geführt haben.

3.1 Einbindung von Prototypen

Art der Erklärungen:
Entscheidungserklärungen; ein Modell liefert sowohl Entscheidung als auch Erklärung

Anwendbar auf:
alle Modelle, unabhängig von deren konkreter Implementierung (aber mit Fokus auf Klassifikationsproblemen); Bild- und Textdaten sowie numerische Daten

Beispiel:
Ein KI-Modell soll Patient:innen anhand ihrer Symptome einem Krankheitsbild zuordnen. Für jedes Krankheitsbild, z. B. Erkältung, Grippe oder Lungenentzündung, wird ein Prototyp erstellt. Dieser Prototyp fungiert als eine Art Steckbrief, der die jeweils häufig auftretenden Symptome zusammenfasst. Die Prototypen können auf Grundlage der Symptome vieler verschiedener Patient:innen, die unter der entsprechenden Erkrankung leiden, erstellt werden. Für jeden zu klassifizierenden Erkrankten wird nun ebenfalls ein Steckbrief erstellt, der die Symptome enthält, z. B. Husten, Fieber und Schnupfen. Dieser wird dann mit den Repräsentationen der einzelnen Klassen (Krankheitsbilder) verglichen und die ähnlichste herausgesucht.

Technischer Hintergrund:
Bei der Verwendung von Prototypen geht es darum, Repräsentationen einzelner Klassen zu erzeugen. Diese Repräsentationen können beispielsweise Datenpunkte aus der Trainingsgrundlage sein, die die jeweilige Klasse gut beschreiben, oder künstlich erzeugte Repräsentationen, die die für die jeweilige Klasse charakteristischen Merkmale umfassen. Diese künstlichen Repräsentationen können beispielsweise mit generativen Netzen wie

Generative Adversarial Networks oder Variational Autoencoders erzeugt werden. Es ist zudem möglich, dass eine Klasse durch mehrere Prototypen charakterisiert wird. Um eine Erklärung zu liefern, muss für ein Klassifikationsergebnis der ähnlichste Prototyp gefunden werden. Dazu kann beispielsweise eine K-Nearest-Neighbor-Suche[5] eingesetzt werden. Der bzw. die Nutzer:in kann schließlich den verwendeten Prototyp für die konkrete Klasse sowie die eingegebenen Datenwerte vergleichen und so nachvollziehen, auf welcher Grundlage, also gemäß welcher Gemeinsamkeiten, die Entscheidung durch das KI-Modell getroffen wurde.

Vorteile:
- Zahl der Prototypen kann frei gewählt werden
- Intuitiv und leicht verständlich
- Unabhängig vom KI-Modell und Datentypen

Nachteile:
- Unter Umständen Zahl der benötigten Prototypen unklar
- Bei künstlich erzeugten Prototypen: eventuell nicht realistisch

(Molnar 2019; Barbalau et al. 2020; Li et al. 2017)

3.2 Einbindung von externen Wissensbasen

Art der Erklärungen:
Entscheidungserklärungen; ein Modell liefert sowohl Entscheidung als auch Erklärung

Anwendbar auf:
Alle Modelle, unabhängig von deren konkreter Implementierung (Fokus auf Klassifikationsprobleme); nur Textdaten (Wissensbasis erforderlich)

Beispiel:
Die Datenbank PubMed enthält zahlreiche medizinische Artikel, die unter anderem konkrete Krankheiten und deren Symptome beschreiben. Durch Nutzung dieser Wissensbasis kann ein KI-Modell die Zusammenhänge zwischen Symptomen und Krankheiten lernen. Werden dem Modell anschließend Symptome von Patient:innen präsentiert, etwa eine starke Schwellung und ein Bluterguss am Knöchel sowie Schmerzen bei der Belastung, so könnte das Ergebnis des Modells „Verdacht auf Bänderriss" sein. Gleichzeitig erfolgt der Verweis auf einen oder mehrere Artikel aus PubMed, in denen genau diese Symptome und die Ableitung des entsprechenden Krankheitsbilds beschrieben sind und für die Anwender:in klar ersichtlich hervorgehoben werden.

Technischer Hintergrund:
Bei diesem Ansatz wird das KI-Modell (z. B. neuronale Netze) mit externen Wissensbasen kombiniert. Die Wissensbasis wird bereits während des Trainings des KI-Modells genutzt, um das Modell zu erstellen. Wissensbasen können beispielsweise Online-Publikationen zu einem bestimmten Thema, Fachbücher oder Wikipedia sein. Ziel ist hierbei, Zusammenhänge aus den Wissensbasen zu lernen und vom KI-Modell getroffene Entscheidungen mit konkreten Einträgen in der Wissensbasis begründen zu können.

Vorteile:
- Leicht nachvollziehbar: Verlässlichkeit der den Entscheidungen zugrundeliegenden Publikationen kann einfach geprüft werden
- Kombination mehrerer Wissensbasen möglich

Nachteile:
- Abhängig von Qualität (und Vorhandensein) der Wissensbasis
- Selbstständiger Aufbau einer qualitativen Wissensbasis ist sehr zeitaufwendig

(van Aken et al. 2021; Holzinger et al. 2017)

3.3 Surrogat-Modelle (Stellvertreter-Modelle)

Art der Erklärungen:
Bezogen auf das Ursprungsmodell weder Modell- noch Entscheidungserklärungen, da ein neues Modell erstellt wird; Post hoc

Anwendbar auf:
Alle Modelle, unabhängig von deren konkreter Implementierung; Bild- und Textdaten sowie numerische Daten

Beispiel:
Es wird ein nicht leicht verständliches KI-Modell trainiert, z. B. eine Support Vector Machine, um die täglichen

5 K-Nearest-Neighbor ist eine Methode, mit der einem gegebenen Datenpunkt ähnliche weitere Datenpunkte zugeordnet werden. Die Ähnlichkeit kann auf unterschiedliche Weise bestimmt werden.

Verleihzahlen für Wintersportausrüstung zu prognostizieren. In das Modell fließen viele Faktoren ein, beispielsweise die Jahreszeit, der Wetterbericht, die Zeiten der Schulferien und der Wochentag. Nun wird ein einfacheres Modell, etwa ein Entscheidungsbaum, trainiert, dessen Entscheidungen nachvollziehbarer sind, der aber nicht die komplette Komplexität des ursprünglichen Modells abbilden kann. So sind die Prognosen des zweiten Modells oft ungenauer, doch es lassen sich allgemeingültige Regelmäßigkeiten ableiten wie: „Bei Nebel werden weniger Skier ausgeliehen" oder „Montags ist die Zahl der Ausleihen deutlich geringer als Sonntags".

Technischer Hintergrund:
Bei der Erstellung von Surrogat-Modellen geht es darum, auf Grundlage eines Black-Box-Modells ein zweites Modell zu erstellen, z. B. ein lineares Modell oder einen Entscheidungsbaum, das besser nachvollziehbar ist und zur Erklärung der Entscheidungen genutzt werden kann. Aufbauend auf den Ein- und Ausgaben des ursprünglichen Modells, wird die Vorhersagefunktion des Surrogat-Modells abgeleitet. So können Regeln aus trainierten neuronalen Netzen extrahiert und, darauf aufbauend, nachvollziehbare Entscheidungsbäume erstellt werden, etwa mit dem Algorithmus TREPAN (Touretzky 1996). Insbesondere für Bilddaten ist die Erstellung derartiger Bäume nicht trivial.

Vorteile:
- Sehr flexibel: Ursprüngliches und Surrogat-Modell können frei gewählt werden

Nachteile:
- Nur Approximation: Repräsentativität des Surrogat-Modells schwer messbar
- Surrogat-Modell selbst kann sehr komplex und weniger nachvollziehbar werden
- Nicht leicht auf Bilddaten anwendbar

(Adadi und Berrada 2018; Danilevsky et al. 2020; Molnar 2019)

3.4 Counterfactual Explanations

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Alle Modelle, unabhängig von deren konkreter Implementierung; Bild- und Textdaten sowie numerische Daten

Beispiel:
Bei der Prüfung der Bewerbungen für eine Mietwohnung werden die Interessierten hinsichtlich mehrerer Kriterien bewertet und der geeignetste Interessent oder die geeignetste Interessentin ausgewählt. Beispielhaft werden hier die drei Faktoren Einkommen, Haustierbesitz und Schufa-Auskunft betrachtet. Wird vom KI-System nun einer Interessentin oder einem Interessenten mit einem Jahreseinkommen von 40.000 Euro, keinen Haustieren und einer positiven Schufa-Auskunft eine Absage erteilt, so wäre eine mögliche Erklärung: Bei einem Einkommen von 45.000 Euro kommt die Bewerberin oder der Bewerber für eine Anmietung in Frage. Genauso kann das Konzept bei einer positiven Entscheidung genutzt werden. Beispielhafte Erklärungen wären dafür: „Hätte die Bewerberin oder der Bewerber eine Katze, käme sie/er nicht als Interessent:in in Frage" oder „Wäre die Schufa-Auskunft negativ, so kann er die Wohnung nicht mieten".

Technischer Hintergrund:
Counterfactual Explanations ist ein Konzept, das darauf abzielt, für ein konkretes Klassifikationsergebnis eine möglichst kleine Änderung in den Eingabewerten zu identifizieren, die zu einer Einteilung in eine andere Klasse führen würde. Es können mehrere dieser Änderungsmöglichkeiten existieren, die das Ergebnis gut erklären. Die folgenden vier Eigenschaften machen gute Counterfactual Explanations aus:

- Die ursprüngliche Entscheidung und die neu erzeugte sind sehr ähnlich
- So wenig Features wie möglich sollten verändert werden
- Mehrere unterschiedliche Erklärungen können hilfreich sein
- Die veränderten Features sollten realistisch sein

Praktische Umsetzungen adressieren nicht unbedingt alle dieser Anforderungen, sodass geprüft werden muss, welche dieser Punkte für den eigenen Anwendungsfall notwendig sind. Die Implementierung von Wachter et al. fokussiert beispielsweise nur auf die ersten beiden Anforderungen (Wachter et al. 2017). Dazu wird eine entsprechende Verlustfunktion aufgestellt, die hinsichtlich eines oder mehrerer Ziele optimiert wird.

Vorteile:
- Gut verständlich
- Kein Daten- oder Modellzugriff erforderlich

Nachteile:
- Vorgeschlagene Änderung kann in der Praxis nicht realistisch oder gar unmöglich sein
- Mehrere, sich widersprechende Erklärungen möglich

(Wachter et al. 2017; Stepin et al. 2021; Molnar 2019)

3.5 LIME (Local Interpretable Model-Agnostic Explanations)

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Alle Modelle, unabhängig von deren konkreter Implementierung; Bild- und Textdaten sowie numerische Daten

Beispiel:
Als konkretes Beispiel wird die Sterbewahrscheinlichkeit einer Krebspatientin abhängig vom Alter berechnet. Für eine 25-jährige Patientin wird als Ergebnis des KI-Modells eine Wahrscheinlichkeit von 45 Prozent angenommen. Nun werden die Sterbewahrscheinlichkeiten für Patient:innen mit einem ähnlichen Alter, beispielsweise 24 (44 Prozent) und 26 Jahre (46 Prozent), berechnet. Anhand dieser drei Werte – in der Praxis werden üblicherweise mehr verwendet – ist es möglich, das Verhalten des Modells in einem begrenzten Rahmen einzuschätzen: z. B. ein leichter (linearer) Anstieg der Sterbewahrscheinlichkeit mit zunehmendem Alter. Für Patient:innen im Alter von 74, 75 und 76 Jahren kann sich das Modell anders verhalten – beispielweise einen deutlich stärkeren Anstieg der Sterbewahrscheinlichkeit mit zunehmendem Alter ausweisen. So können mithilfe von LIME einzelne „Abschnitte" des eigentlich stark verflochtenen und vielschichtigen Modells vereinfacht und so vom Nutzer oder der Nutzerin besser nachvollzogen werden (Nguyen 2020).

Technischer Hintergrund:
Grundidee von LIME ist das Erlernen eines lokal approximierten, interpretierbaren Modells für ein konkretes Klassifikations- oder Regressionsergebnis. Dadurch kann mithilfe eines einfacheren, oft linearen Modells ein konkretes Ergebnis nachvollzogen werden, obwohl das ursprüngliche Modell nur schwer nachvollziehbar ist. LIME „sampelt" mehrere Ergebnisse (bzw. Entscheidungen) und gewichtet diese entsprechend ihrer Nähe zum zu erklärenden Ergebnis. Auf dieser Basis kann ein lokales Modell entwickelt werden, das mit den betrachteten Samples gut funktioniert und nachvollziehbar ist.

Vorteile:
- Intuitiv und im Allgemeinen gut interpretierbar
- Schnell und einfach in bestehende Implementierungen integrierbar (entsprechendes Framework vorhanden)

Nachteile:
- Problematisch bei ausgeprägt nichtlinearen Modellen
- Unter Umständen hohe Rechenzeit bei mehrdimensionalen Daten, z. B. Bilddaten
- Kaum reproduzierbar aufgrund Datensampling (eine mehrmals ausgeführte Klassifikation könnte unterschiedlich erklärt werden)

(Nguyen 2020; Ribeiro et al. 2016)

3.6 SHAP (SHapley Additive exPlanations)

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Alle Modelle, es existieren Optimierungen für einzelne Modelle (z. B. TreeSHAP für Random Forests); Bild- und Textdaten sowie numerische Daten

Beispiel:
Als Beispiel soll die Prognose des Einkommens anhand der drei Faktoren Alter, Geschlecht und Beruf betrachtet werden. Es wird der Einfluss jedes einzelnen Faktors auf ein konkretes Ergebnis des KI-Systems bestimmt. Um herauszufinden, wie wichtig das Alter für die Einkommensprognose ist, wird zuerst eine „normale" Prognose unter Berücksichtigung von Alter, Geschlecht und Beruf berechnet. Anschließend wird erneut eine Prognose erstellt, die jedoch nur die zwei Faktoren Geschlecht und Beruf nutzt. So kann im Nachhinein der Unterschied zwischen den beiden Ergebnissen – einmal unter Berücksichtigung des Alters, einmal ohne Berücksichtigung des Alters – berechnet und der Einfluss des Faktors „Alter" bestimmt werden. Dieser Vorgang wird für die anderen beiden Faktoren wiederholt (Mazzanti 2020).

Technischer Hintergrund:
Bei SHAP handelt es sich um einen Ansatz aus der Spieltheorie. Bei der Anwendung der Methode wird jedes Feature bzw. jeder Inputwert im Hinblick auf ein konkretes Klassifikationsergebnis gewichtet. Diese Gewichte werden auch als Shapley Values bezeichnet. Die Idee dahinter ist, dass alle möglichen Kombinatio-

nen von Features beachtet werden, um die Wichtigkeit eines einzelnen Features zu bestimmen. Jedem Eingabefeature wird so ein positiver oder negativer Wert zugeordnet, der den Einfluss des einzelnen Features auf das Ergebnis angibt. Die Methode kann genutzt werden, um Entscheidungserklärungen zu generieren. TreeSHAP ist eine Variante von SHAP, die besonders effizient auf baumbasierte Modelle angewendet werden kann.

Vorteile:
- Modellagnostisch (Variationen sorgen für hohe Effizienz)
- Sehr präzise Erklärungen möglich
- Gilt als industrieller Standard

Nachteile:
- Erklärungen nicht immer intuitiv
- Eventuell hohe Rechenzeiten (vor allem bei Modellen mit hoher Parameterzahl)

(Molnar 2019; Lundberg und Lee 2017; Mangalathu et al. 2020; Mazzanti 2020; Bhatt et al. 2019)

3.7 Attribution Methods

Mithilfe sogenannter Attribution Methods wird der negative oder positive Einfluss von Teilen oder Bereichen der Eingabe eines KI-Modells auf dessen Ausgabe betrachtet (Sundararajan et al. 2017). Dieser Gruppe können die folgenden konkreten Methoden zugeordnet werden: Sensitivitätsanalyse, LRP, DeepLIFT, Integrated Gradients, Grad-CAM, Guided Backpropagation und Deconvolution. Anhand eines gemeinsamen Beispiels wird die Funktionsweise erläutert. Die Unterschiede sind in den nachfolgenden technischen Einzelheiten beschrieben.

Beispiel:
Bei der Bildklassifizierung geht es darum, Bildbereiche zu identifizieren, die ausschlaggebend für das Klassifikationsergebnis sind. Beispielhaft sollen mithilfe eines neuronalen Netzes Objekte auf einem Bild erkannt werden. Die beiden möglichen Klassen sind „Katze“ oder „Hund“. Nachdem das KI-Modell eine Entscheidung, z. B. „Katze“, geliefert hat, wird der Einfluss einzelner Pixel und Bildbereiche auf die konkrete Entscheidung untersucht. Dafür werden die einzelnen Bestandteile des neuronalen Netzes – Einheiten („Units“) und Schichten – betrachtet, um so die Ausgaben auf das Eingabebild zu „mappen“. Es entsteht eine sogenannte Saliency Map, in der die Pixel und Bildbereiche, die einen besonders großen Einfluss darauf hatten, dass das Tier als Katze erkannt wurde, hervorgehoben sind.

3.7.1 CAM / Grad-CAM / Grad-CAM++ (Gradient-weighted Class Activation Mapping)

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Neuronale Netze, insbesondere Convolutional Neural Networks (für CAM ist u. U. das Hinzufügen spezieller Schichten notwendig); Bilddaten

Technischer Hintergrund:
CAM ist eine Methode zur Visualisierung von ausschlaggebenden Regionen für ein konkretes Klassifizierungsergebnis eines neuronalen Netzes, insbesondere Convolutional Neural Network (CNN). Das Ergebnis ist eine Saliency Map, die über das ursprüngliche Bild gelegt werden kann und die betreffenden Regionen hervorhebt. Für die Erstellung der Saliency Map werden jeweils nur die letzten Schichten (Layer) des Netzes betrachtet. CAM ist nicht für jede Netzwerkarchitektur direkt anwendbar; unter Umständen muss diese durch Hinzufügen weiterer Schichten zuvor angepasst und dann das Netz neu trainiert werden.

Grad-CAM ist eine Generalisierung der CAM-Methode, erfordert kein erneutes Training des Modells und ist auf mehr Netzwerkarchitekturen anwendbar. Ein Nachteil von Grad-CAM ist jedoch, dass nicht mehrere Vorkommen eines Objekts in einem Bild erkannt werden können. Grad-CAM++ löst dieses Problem, sodass das Erkennen von mehreren Objektinstanzen in einem Bild möglich wird.

Vorteile:
- Visualisierungen korrelieren mit menschlicher Aufmerksamkeit → leicht verständlich
- Gute Ergebnisse bei Aufgaben, in denen Bildobjekte lokalisiert werden müssen

Nachteile:
- Visualisierungen oft zu grob für kleine Bildobjekte → nur Grobvalidierung (Qualität stark von konkreter Anwendung abhängig)
- CAM: zusätzliche Layer müssen trainiert werden

(Zhou et al. 2015; Selvaraju et al. 2019; Chattopadhyay et al. 2017)

3.7.2 LRP (Layer-Wise Relevance Propagation)

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Neuronale Netze; Fokus auf Bilddaten

Technischer Hintergrund:
Durch LRP wird der Einfluss einzelner Eingaben auf das Ergebnis einer Klassifikation betrachtet. Der Fokus liegt hierbei auf nicht linearen Klassifizierern wie neuronalen Netzen. Betrachtet man die Bildklassifikation, so ist das Ziel, für einzelne Bilder herauszufinden, welche Pixel in welchem Umfang das Klassifizierungsergebnis positiv oder negativ beeinflussen. Jedem Inputwert (hier: Pixel) wird ein Relevanzwert zugeordnet. Der Wert der „Relevanz" gibt an, wie groß der Einfluss eines Eingabewerts oder einer Unit des Netzes auf das Klassifikationsergebnis ist. Der Relevanzwert der Ausgabe setzt sich aus der Summe der Relevanzwerte der Eingabewerte zusammen. Der Ausgabewert des Netzes wird also „zerlegt" in die jeweiligen Beiträge (bzw. den Einfluss) der Eingabewerte (= Dekomposition). Die Berechnung der Relevanz der Eingabewerte wird iterativ von hinten (letzter Layer) nach vorn (Input-Layer) ausgeführt.

Vorteile:
- Gute Qualität der Erklärungen auch bei vielschichtigen Modellen mit hoher Parameterzahl
- Erklärungen können sehr schnell erzeugt werden (bezogen auf die Laufzeit)

Nachteile:
- Numerische Probleme bei Dekomposition möglich → möglicherweise irreführende Visualisierungen

(Bach et al.; Samek et al. 2019; Shiebler 2017)

3.7.3 IG (Integrated Gradients)

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Neuronale Netze; Bild- und Textdaten sowie numerische Daten

Technischer Hintergrund:
Diese Methode ist ebenfalls zur Verbesserung der Erklärbarkeit von neuronalen Netzen durch Visualisierung gedacht. Ein Vorteil ist, dass die Struktur des Netzes nicht verändert werden muss, wie u. U. bei CAM. Für die beispielhafte Betrachtung von Bilddaten wird bei der Anwendung von IG ein Bild als Baseline gewählt, etwa ein komplett schwarzes Bild. Anschließend wird eine Reihe interpolierter Bilder „zwischen" der Baseline und dem originalen Input erstellt, die sich jeweils nur wenig voneinander unterscheiden. Auf dieser Grundlage werden einzelne Gradienten berechnet, die wiederum genutzt werden, um interessante Bereiche – also für die Klassifikation ausschlaggebende – im Eingabebild zu identifizieren.

Vorteile:
- Skaliert gut für Bildverarbeitung
- Positiver und negativer Einfluss einzelner Eingabewerte separat darstellbar
- Verwendung Baseline: intuitiver Ansatz
- Gilt als industrieller Standard

Nachteile:
- Korrekte Wahl der Baseline unklar → stark variierende Ergebnisse
- Erklärungen nicht immer intuitiv

(Sundararajan et al. 2017; Bhatt et al. 2019; Google 2020)

3.7.4 DeepLIFT (Deep Learning Important FeaTures)

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Neuronale Netze; Fokus auf Bilddaten

Technischer Hintergrund:
DeepLIFT ist ein Erklärungswerkzeug, das zur Verbesserung der Nachvollziehbarkeit von neuronalen Netzen genutzt wird. Bei der Methode wird einzelnen Units des neuronalen Netzes, bezogen auf einen konkreten Output (Klassifikations- oder Regressionsergebnis), ein Score zugeordnet. Wie bei der Methode Integrated Gradients wird eine Baseline genutzt: Es wird ein neutraler Input gewählt (abhängig vom konkreten Anwendungsfall), für den die Aktivierungen der einzelnen Units bzw. Neuronen des Netzes berechnet werden. Es werden also Referenzwerte bestimmt. Anschließend wird die Abweichung – der „Score" – von diesen Referenzwerten für eine konkrete Eingabe pro Unit berechnet. Die Wahl des neutralen Inputs ist kritisch und sollte unter Nutzung von Domänenwissen erfolgen. In einigen Fällen ist

es sinnvoll, mehrere neutrale Inputs zu bestimmen und die einzelnen Scores auf Grundlage mehrerer Werte zu berechnen.

Vorteile:
- Positiver und negativer Einfluss einzelner Eingabewerte separat darstellbar
- Verwendung Baseline: intuitiver Ansatz
- Ermöglicht schnelle Approximation für Integrated Gradients

Nachteile:
- Korrekte Wahl der Baseline unklar → stark variierende Ergebnisse

(Shrikumar et al. 2016; Shrikumar et al. 2017; Salehi 2020)

3.7.5 Guided Backpropagation und Deconvolution / DeconvNet

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Neuronale Netze, insbesondere Convolutional Neural Networks; Bilddaten

Technischer Hintergrund:
Mit Guided Backpropagation bzw. DeconvNet (Deconvolution) können wichtige Features der Eingabe sowie einzelne Layer eines neuronalen Netzes visualisiert werden. Bei beiden Methoden werden die Aktivierungswerte der einzelnen Units durch das neuronale Netz zurück auf den jeweiligen Input gemappt, um mit einer Saliency Map die Inputwerte zu identifizieren, die für eine konkrete Klassifizierung ausschlaggebend sind. Es werden die gleichen Komponenten wie bei einem Convolutional Neural Network verwendet – z. B. pooling –, jedoch „umgekehrt". Der Prozess des Durchgehens des Netzes von hinten nach vorn wird auch als Backpropagation bezeichnet. Die beiden Methoden Guided Backpropagation und Deconvolution bzw. DeconvNet unterscheiden sich nur in den konkreten Berechnungen der Backpropagation-Schritte.

Vorteile:
- Schnelle Berechnung, nur zwei „Netzdurchläufe" – vorwärts/rückwärts – nötig
- Motivation hinter den Methoden sehr intuitiv
- Guided Backpropagation: „trennschärfere" Visualisierungen im Vergleich zu DeconvNet

Nachteile:
- Starker Fokus auf Convolutional Neural Networks → für andere Architekturen weniger gut geeignet

(Springenberg et al. 2014; Zeiler und Fergus 2013; Zeiler et al. 2011)

3.7.6 Activation Maximization

Art der Erklärungen:
Modellerklärungen; Post hoc

Anwendbar auf:
Neuronale Netze; Fokus auf Bilddaten

Technischer Hintergrund:
Durch Activation Maximization sollen Erkenntnisse über die von einem neuronalen Netz gelernten Strukturen zur Erkennung verschiedener Klassen gewonnen werden. Ziel dabei ist es, Inputdaten zu finden, die dazu führen, dass die Entscheidung des neuronalen Netzes mit größtmöglicher Konfidenz einer bestimmten Klasse entspricht. Anschließend kann der so erzeugte „perfekte" Input auf Plausibilität überprüft werden. Bezogen auf das gesamte Netz, kann jede einzelne Unit betrachtet und die Aktivierung dieser durch einen bestimmten Input maximiert werden. So können einzelne Units und Layer innerhalb des Netzes untersucht und damit Modellerklärungen bereitgestellt werden.

Vorteile:
- Erklärungen sehr feingranular möglich, z. B. für einzelne Layer oder Units
- Liefert Modellerklärungen und Einblick in Funktionsweise des Modells

Nachteile:
- Ergebnisse sind rein qualitativ
- Interpretation schwierig und sehr subjektiv (insbesondere für tiefe Layer)

(Erhan et al. 2009; Ye 2020)

3.7.7 Sensitivitätsanalyse

Art der Erklärungen:
Entscheidungserklärungen; Post hoc

Anwendbar auf:
Alle Modelle, unabhängig von deren konkreter Implementierung; Bild- und Textdaten sowie numerische Daten

Technischer Hintergrund:
Die Sensitivitätsanalyse ist ein Konzept, das disziplinübergreifend für die Analyse von Systemen angewendet wird. Bei der Sensitivitätsanalyse werden einzelne Eingabeparameterwerte eines Modells systematisch variiert (im jeweils zulässigen Bereich). Durch diese systematischen Variationen, auch Pertubationen genannt, kann ermittelt werden, welche Eingabeparameter bzw. Features den größten Einfluss auf z. B. ein Klassifikationsergebnis haben. Relevante Features können als Grundlage einer entsprechenden Erklärung herangezogen werden. Die Sensitivitätsanalyse ist modellagnostisch und liefert auf sehr einfache Weise Entscheidungserklärungen im Sinne einer Feature-Wichtigkeit. Bei der eindimensionalen Sensitivitätsanalyse wird immer nur ein einzelner Inputwert variiert, bei mehrdimensionalen Varianten kann auch der Einfluss von mehreren variierten Eingabeparametern gleichzeitig untersucht werden.

Vorteile:
- Sehr schnell und einfach für differenzierbare Modelle

Nachteile:
- Nicht geeignet für nicht differenzierbare Modelle

(Cortez und Embrechts 2011; Baehrens et al. 2009)

Bemerkung

Vor- und Nachteile wurden auf Grundlage der Expert:inneninterviews, auf vom Bosch Center for Artificial Intelligence zur Verfügung gestellten Informationen und unter Nutzung der folgenden zusätzlichen Quellen zusammengestellt: (Bhatt et al. 2019; Sundararajan et al. 2017; Google 2020; Gondal et al. 2017; Montavon et al. 2019; Shrikumar et al. 2017; Tjoa und Guan 2020).

4 DER AKTUELLE EINSATZ VON ERKLÄRBARER KI IN WIRTSCHAFT UND WISSENSCHAFT

4 DER AKTUELLE EINSATZ VON ERKLÄRBARER KI IN WIRTSCHAFT UND WISSENSCHAFT

Im Rahmen dieses Kapitels werden die Ergebnisse der mit Vertreter:innen aus Unternehmen mit KI-Bezug sowie wissenschaftlichen Einrichtungen durchgeführten Umfrage vorgestellt (Anzahl der Teilnehmenden: n = 209). Die Ergebnisse spiegeln die Angaben von KI-Entwickler:innen (etwa 75 Prozent der Befragten) und KI-Anwender:innen wider (etwa 25 Prozent der Befragten) hinsichtlich

- der Nutzung spezifischer Datentypen, KI-Verfahren und -Modelle,
- der Erklärbarkeit ausgewählter KI-Modelle,
- branchenspezifischer Anforderungen in Bezug auf Erklärbarkeit sowie
- konkreter Zielgruppen und Umsetzungsmöglichkeiten für Erklärungen.

Die im vorherigen Kapitel vorgestellten Erklärungsstrategien verdeutlichen, dass für die Auswahl eines geeigneten Werkzeugs die vorhandene Datengrundlage berücksichtigt werden muss. Einige Ansätze sind besonders für die Erstellung von Erklärungen für bilddatenverarbeitende KI-Systeme geeignet – z. B. Grad-CAM oder LRP –, für andere werden textbasierte Wissensbasen benötigt.

Bei der Betrachtung der Datentypen, mit denen Entwickler:innen und Anwender:innen laut Befragung am häufigsten arbeiten (siehe Abbildung 3), wird deutlich, dass insbesondere numerische Daten, Bild- bzw. Videodaten sowie Textdaten eine wichtige Rolle spielen. Numerische Daten belegen in der Umfrage den ersten Platz und werden von etwa drei Vierteln der Befragten verwendet.

Deutlich weniger häufig genutzt werden Audiodaten. Von ca. zehn Prozent der Befragten wurden zusätzliche Datentypen benannt, diese umfassen z. B. 3D-, CAD- und Geo-Daten und wurden für die Auswertung in der Kategorie „Sonstige“ zusammengefasst.

Unter der hypothetischen Annahme, dass bei einer Einteilung von Anwendungsproblemen nach Datentypen die Anwendungskritikalität, die Verteilung des KI-Modelleinsatzes und gleichzeitig die zielgruppenbezogenen Bedarfe für erklärbare KI je nach Kategorie nicht variieren würden, lässt sich aus den Umfrageergebnissen Folgendes schließen: Entweder müssen Erklärungsstrategien unterschiedliche Datentypen unterstützen, oder sie sollten für einen dieser drei Datentypen – numerisch, text- oder bildbasiert – besonders geeignet sein. Dieser Zusammenhang impliziert gleichzeitig, dass auch die eingesetzten KI-Modelle die Auswahl geeigneter Erklärungsstrategien beeinflussen. Einige der in Kapitel 3 diskutierten Ansätze, z. B. Integrated Gradients oder DeepLIFT, sind etwa nur auf neuronale Netze als zugrundeliegenden Modelltyp anwendbar. Surrogat-Modelle oder Counterfactual Explanations können sowohl auf neuronale Netze als auch auf andere KI-Modelle angewendet werden.

Die Umfrageergebnisse zum Einsatz von KI-Modellen und -Verfahren zeigen, dass die in der öffentlichen und in der Fachdebatte häufig im Vordergrund stehenden neuronalen Netze derzeit keineswegs die allein eingesetzten KI-Modelle sind. Genauso häufig werden die einfacher zu interpretierenden Entscheidungsbäume (Decision Trees) verwendet. Statistische und probabilis-

Von den Befragten verwendete Datentypen (Mehrfachnennung war möglich)*

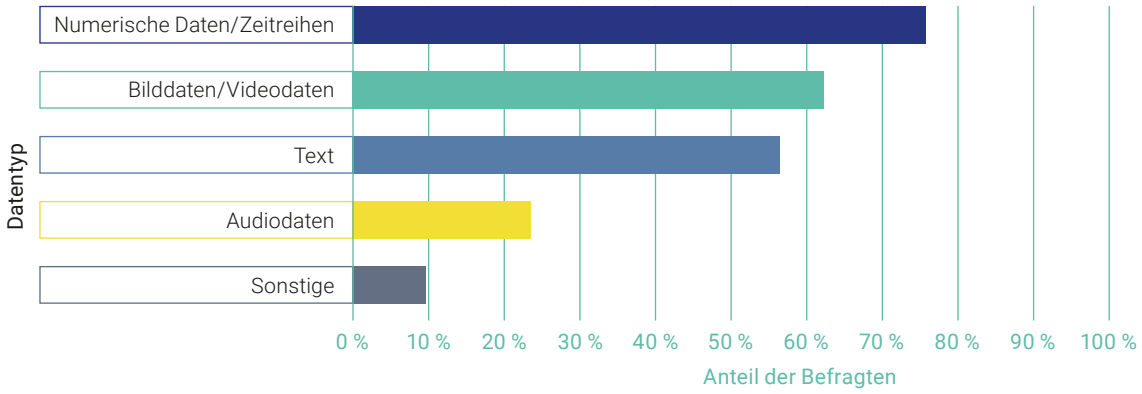


Abbildung 3 – Umfrageergebnis: Numerische Daten sind mit ca. 75 Prozent der am häufigsten verwendete Datentyp.

* Einige nicht im Fragebogen auswählbare Datentypen wie z. B. 3D-, CAD- oder auch Geo-Daten wurden von mehreren Personen angegeben und in der Abbildung unter „Sonstige“ eingeordnet.

Heutige und zukünftige Anwendung ausgewählter Modelle und Verfahren laut Befragten über alle Ziel-/Anwendungsbranchen (Mehrfachnennung möglich)*

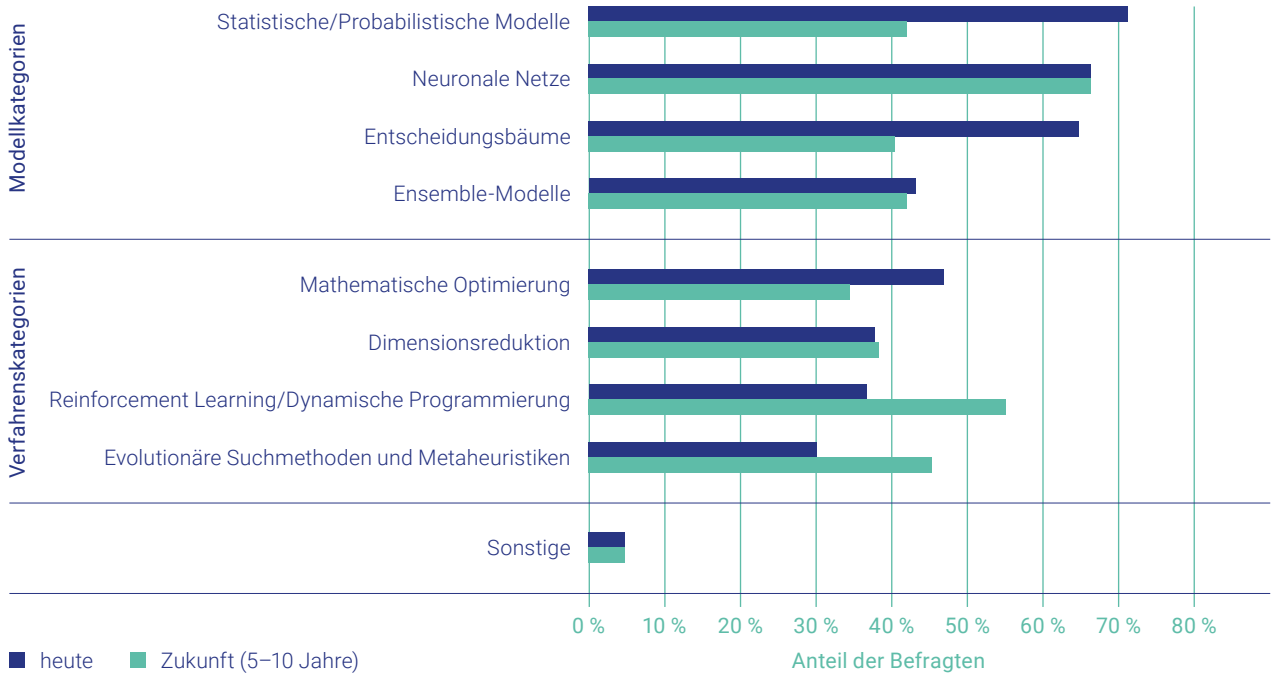


Abbildung 4 – Umfrageergebnis: Neuronale Netze zukünftig wichtigster Modelltypus, wesentliche Steigerung im Anwendungsbereich Reinforcement Learning erwartet

* Im Fragebogen auswählbar war nur eine begrenzte Anzahl von Kategorien typischer KI-Modelle und -Verfahren, die für den Studiengegenstand als relevant erachtet wurden. Aus der Grafik ist ablesbar, welcher Anteil der Befragten das jeweilige KI-Modell bzw. KI-Verfahren einsetzt. Mehrere Teilnehmende ergänzten unter „Sonstiges" die Kategorien „Expertensysteme", „Wissensgraphen" und „Semantic Web".

tische Modelle stellen die derzeit am weitesten verbreitete Modellkategorie dar. Auch die Kombination bzw. Hintereinanderschaltung mehrerer redundanter Modelle (Ensemble-Modelle) stellt für einen großen Kreis der Befragten eine häufig genutzte Modellkategorie dar, siehe Abbildung 4.

Von den Verfahrenskategorien, die zur Auswahl standen, sind Ansätze der mathematischen Optimierung bei den befragten Personen am weitesten verbreitet. Verfahren zur Dimensionsreduktion und Reinforcement Learning bzw. dynamische Programmierung werden nur geringfügig seltener eingesetzt. Evolutionäre Suchmethoden und Metaheuristiken bilden heutzutage bei den betrachteten Ansätzen das „Schlusslicht", kommen aber dennoch bei 30 Prozent der befragten Personen zum Einsatz, was letztlich eine gewisse Relevanz aller wählbaren Verfahrens- und Modellkategorien unterstreicht.

Der Blick in die Zukunft zeigt, dass die Befragten bei gleich zwei White-Box-Modellkategorien, nämlich den statistischen/probabilistischen Modellen und den Entscheidungsbäumen, die stärkste rückläufige Entwicklung sehen. Aufgrund der laut Umfrage perspektivisch gleichbleibenden Bedeutung der neuronalen Netze (für ca. 66 Prozent der Befragten) könnte somit eine Black-Box-Modellkategorie in fünf bis zehn Jahren den wichtigsten Typus eines Modells darstellen. Damit erhalten implizit auch die Erklärungsstrategien eine immer größere Wichtigkeit.

Aus der Umfrage geht andererseits hervor, dass die Bedeutung von Reinforcement-Learning bzw. evolutionären Suchmethoden und Metaheuristiken künftig laut Umfrageteilnehmenden zunehmen wird und somit vermehrt „on-the-job" lernende bzw. nichtdeterministische Verfahren zum Einsatz kommen werden. Diese Studie setzt sich nur im Rahmen eines speziellen Use Cases teilweise mit dem „Modelltyp" der Regelstrategie (Control-Policy) auseinander (im Abschnitt 5.2.2 des folgenden Kapitels). Es sei daher hier auch nur am Rande erwähnt, dass beide Verfahrenskategorien in diesem Anwendungsgebiet je nach methodischer Umsetzung und Einbettung in übergeordnete Steuerungsprozesse potenzielle Herausforderungen in Bezug auf die Nachvollziehbarkeit und funktionale Sicherheit darstellen können. Mögliche „Explorationsphasen" bzw. nichtdeterministische Funktionsweisen von autonomen System stellen häufig ein Ausschlusskriterium für die Zulassung dar, etwa für Steuer- und Regelsysteme in der Produktionswirtschaft (siehe hierzu auch Abschnitt 5.2.3).

Einschätzung der Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit), die ggf. durch Anwendung von Erklärungswerkzeugen erhöht wurde, bezogen auf ausgewählte KI-Modelle*

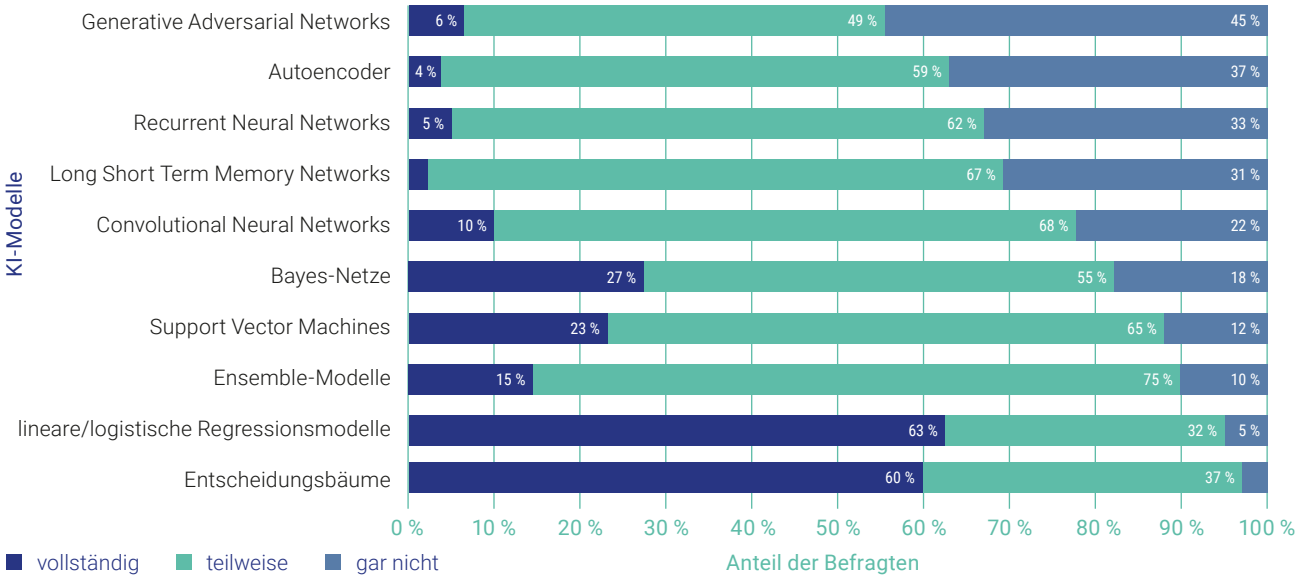


Abbildung 5 – Umfrageergebnis: Entscheidungserklärungen für neuronale Netze bereitzustellen gilt als schwierig.

* Es wurden nur Personen zu individuellen Verfahren und Modellen befragt, die zuvor angegeben hatten, Verfahren oder Modelle der zugeordneten Oberkategorie zu entwickeln oder anzuwenden. Für die Befragten gab es jeweils auch die Möglichkeit, „kann ich nicht beurteilen" als Einschätzung anzugeben. In der Darstellung wurden jedoch nur Angaben von Personen berücksichtigt, die eine entsprechende Einschätzung abgegeben haben. Dabei wurden die individuellen Verfahren von 16 bis 81 Personen und, relativ gesehen, von 50 bis 84 Prozent der jeweils befragten Personen entsprechend beurteilt.

Im Rahmen der Umfrage wurden die Teilnehmenden auch zu einer Einschätzung der Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit) bei der Nutzung unterschiedlicher KI-Modelle aufgefordert. Dabei wurde in der Fragestellung explizit darauf hingewiesen, dass Erklärungswerkzeuge gegebenenfalls mitberücksichtigt werden sollten. Auffällig an den Umfrageergebnissen ist, dass die fünf KI-Modelle, die als insgesamt am wenigsten erklärbar eingeschätzt wurden, ausschließlich aus der Modellfamilie der neuronalen Netze stammen (siehe Abbildung 5). Jedoch schätzen mehr als die Hälfte der Befragten auch diese fünf KI-Modelle bereits heute, zumindest teilweise, als lokal erklärbar ein, zumindest unter Zuhilfenahme entsprechender Erklärungswerkzeuge. Hier zeichnet sich ein Trend ab, der eine Diskrepanz zur öffentlichen Debatte aufweist, in der beispielsweise neuronale Netze oft als gar nicht erklärbar diskutiert werden.

Auffällig ist andererseits auch, dass ein hoher Anteil der befragten Personen die verschiedenen Modellvarianten, die den neuronalen Netzen zuzuordnen sind, als gar nicht erklärbar einschätzen. Hier reicht die Spanne von knapp über 20 Prozent bei Convolutional Neural Networks bis zu 45 Prozent bei Generative Adversarial Networks. Dies deutet darauf hin, dass existierende, einschlägige Erklärungswerkzeuge (Kapitel 3) derzeit einem beträchtlichen Anteil der befragten Personen noch nicht bekannt sind.

Insgesamt ist ersichtlich, dass sich die theoretische Einteilung in White- und Black-Box-Modelle, wie sie die Literatur häufig beschreibt (siehe in Abschnitt 2.2.3), bei expliziter Berücksichtigung von Erklärungswerkzeugen nicht mehr gleichermaßen in den Umfrageergebnissen widerspiegelt. Bei einem Großteil der Modelle wurde von einer Mehrheit angegeben, dass sie „teilweise" erklärbar seien. Nur Entscheidungsbäume sowie lineare und logistische Regressionsmodelle werden mehrheitlich (ca. 60 Prozent) der Kategorie „vollständig erklärbar" zugeschlagen. Die Umfrage zeigt auch, dass nominelle White-Box-Modelle wie Bayes-Netze von geringfügig mehr Personen als „gar nicht" erklärbar eingeschätzt werden als Black-Box-Modelle, wie Support Vector-Machines oder Ensemble-Modelle. Dies lässt vermuten, dass neben der Vielschichtigkeit und der Parameteranzahl von Modellen auch eine gewisse Erfahrung im Umgang mit den Modellen eine Rolle spielt.

Die Einschätzung der derzeitigen Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit) einzelner KI-Modelle steht den konkreten Anforderungen aus den einzelnen Branchen gegenüber. Die Umfrageergebnisse zeigen deutliche Unterschiede zwischen den Branchen

(siehe Abbildung 6). Gerade für Anwendungsfelder, in denen kritische Entscheidungen getroffen werden, wird der Erklärbarkeit eine besonders wichtige Rolle zugewiesen. Erklärbarkeit ist in diesen Fällen häufig sogar zwingend erforderlich, z. B. für Zertifizierung, Prüfsiegel, Normen etc. In der Gesundheitsbranche ist die Schaffung von Entscheidungserklärungen am wichtigsten, was sich gut plausibilisieren lässt, da hier Fehlentscheidungen im schlimmsten Fall fatale Folgen haben können. Weitere Branchen, in denen die Erklärbarkeit von Entscheidungen als besonders wichtig angesehen wird, sind die Finanzwirtschaft, Produktion, Bauwirtschaft und Prozessindustrie. Die Schlüsselrolle der lokalen Erklärbarkeit ergibt sich hier aus den Anforderungen von Kund:innen und Anwender:innen, die ein System, das Einzelentscheidungen nicht erklären kann, in der Regel schlicht nicht akzeptieren würden. Produktionswirtschaft und Prozessindustrie zeichnen sich durch einen hohen Automatisierungsgrad und besondere sicherheitstechnische Anforderungen aus; sie sind daher anspruchsvolle Anwendungsfelder für KI-Anwendungen im Allgemeinen und gleichzeitig von hoher Bedeutung für den deutschen Wirtschaftsstandort.

Abhängig von der jeweiligen Branche, werden KI-Produkte und -Modelle von unterschiedlichen Personen mit unterschiedlichen Eigeninteressen eingesetzt. Entsprechend müssen auch die erzeugten Erklärungen an die jeweiligen Zielgruppen angepasst werden, um einen Mehrwert bieten zu können. Die Adressaten reichen dabei von den Gruppen der KI-Expert:innen (Entwickler:innen) und Domänenexpert:innen (Nutzende) über

Bedeutung Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit) nach Anwendungsbranchen*

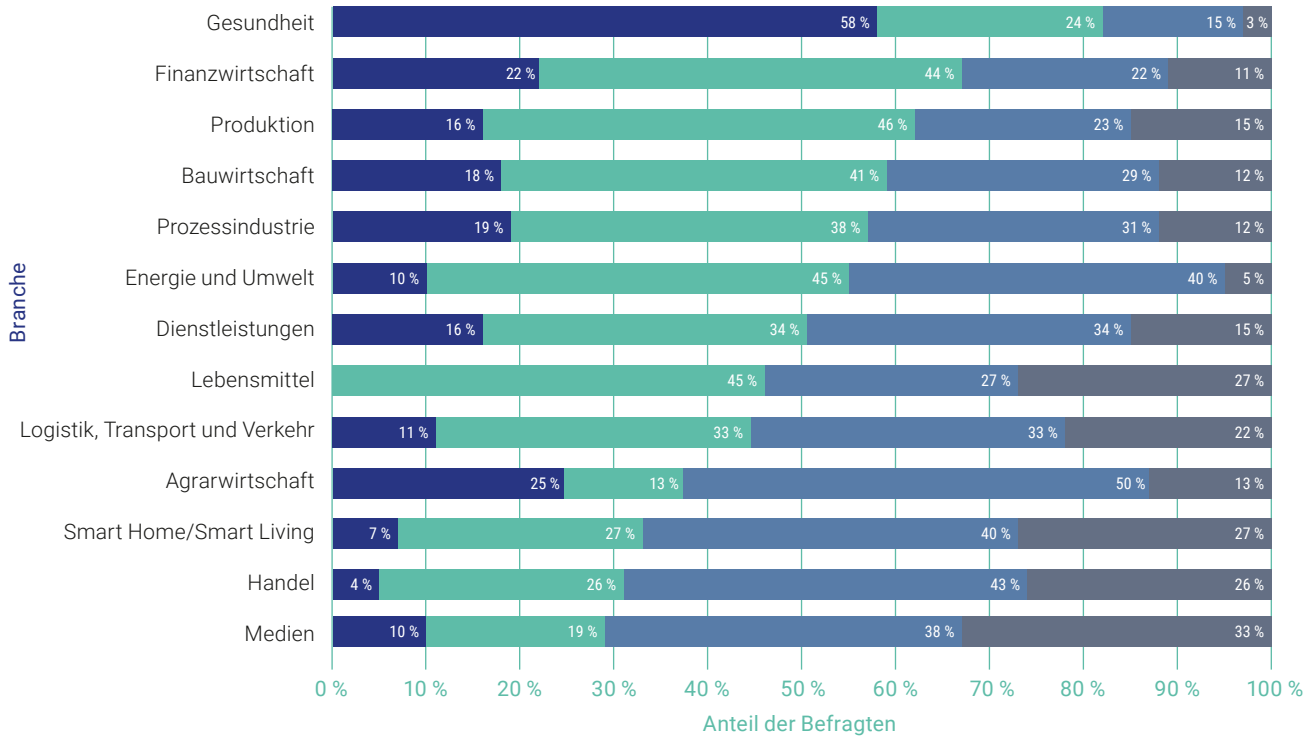


Abbildung 6 – Umfrageergebnis: lokale Erklärbarkeit in Branchen Gesundheit, Finanzwirtschaft, Produktion am stärksten gefordert.

* Es wurden nur Personen zu einzelnen Anwendungsbranchen befragt, die zuvor angegeben hatten, in der jeweiligen bzw. für die jeweilige Anwendungsbranche KI-Systeme anzuwenden bzw. zu entwickeln. Die Branchenreihenfolge in der Abbildung wurde gemäß der empfundenen Bedeutung von Entscheidungserklärungen bzw. lokaler Erklärbarkeit sortiert; hier die Kategorien „zwingend erforderlich" oder „sehr wünschenswert". Bei sonstigen Anwendungsbranchen, die von mehreren Personen angegeben wurden, wurde vor allem für das vergleichsweise unspezifische Anwendungsgebiet „IT / Software" die lokale Erklärbarkeit mehrfach als sehr wünschenswert oder zwingend erforderlich eingestuft. Für weitere einzelne Ergänzungen wie „Legal Tech", Personalwesen und Öffentliche Sicherheit, die auch unter „Sonstiges" eingeordnet wurden, galt die lokale Erklärbarkeit für die betreffenden Personen als zwingend erforderlich.

interne oder externe Prüfer:innen bis zum Management und möglichen Endkund:innen. Im Gesundheitsbereich stellen z. B. Patient:innen die Endkundschaft dar. Die KI-Systeme – zumeist Entscheidungsunterstützungssysteme – werden hier in der Regel vom medizinischen Personal genutzt, das in diesem Kontext als eine Gruppe von Domänenexpert:innen zu klassifizieren ist.

Die Umfrageergebnisse legen nahe, dass die Erklärbarkeit heute vor allem für KI-Entwickler:innen und Domänenexpert:innen bedeutsam ist (siehe Abbildung 7). Dies ist auch bei der Betrachtung aller Einzelbranchen so zu beobachten. Außerdem erwarten die Befragten, dass sich die Bedeutung von Erklärbarkeit für die meisten Zielgruppen in fünf bis zehn Jahren mehr und mehr angleichen wird: Das heißt, Erklärbarkeit könnte generell auch für Endkund:innen, die Management-Ebene sowie interne und externe Prüfer:innen eine größere Bedeutung entfalten, während gleichzeitig ihre Bedeutung für KI-Expert:innen sinkt und für Domänenexpert:innen unverändert bleibt.

Dass die größte Veränderung für Endkund:innen und externe Prüfer:innen prognostiziert wird, lässt sich so interpretieren: Die Teilnehmenden der Umfrage erwarten eine Steigerung der Bedeutung von KI-Zertifizierung sowie der generellen Nachfrage von Kund:innen nach erklärbarer KI infolge eines größeren Produktangebots. Etwas schwieriger nachzuvollziehen ist der erwartete Bedeutungsverlust von Entscheidungserklärungen aus Sicht der KI-Entwickler:innen. Dieser Trend könnte von den beiden sehr gegensätzlichen Annahmen der Umfrageteilnehmer:innen getrieben sein, dass der Erklärbarkeit zukünftig von regulatorischer Seite ein weniger großer Stellenwert eingeräumt werde oder dass die Nutzung von Black-Box-Modellen in gewissen Anwendungsbereichen („High-Risk“) zukünftig generell untersagt werde. Auch ist denkbar, dass viele wissenschaftliche Vertreter:innen, die sich zurecht selbst als KI-Entwickler:innen sehen, im Forschungsfeld in fünf bis zehn Jahren signifikante Fortschritte bzw. einen geminderten Forschungsbedarf vermuten. Nimmt man aber diesen leicht rückläufigen Trend bei den KI-Entwickler:innen von der Betrachtung aus, wird die wachsende Bedeutung von Erklärbarkeit für alle weiteren beteiligten Stakeholder deutlich, bis hinauf in die Management-Ebene.

Mit der Anpassung der Erklärungen an die entsprechende Zielgruppe stellt sich zudem die Frage nach der konkreten Umsetzung bzw. Darstellung der Erklärung, die dem Adressaten den größten Nutzen bringen kann. Auf die Frage, auf welche Weise Erklärungen umgesetzt werden können oder sollten, antworteten die meisten Teilnehmenden, grafische Darstellungen seien gut geeignet (siehe Abbildung 8). Die Umfrage, bei der Mehrfachnennungen möglich waren, zeigt aber auch: Es gibt grundsätzlich viele gangbare Wege zur konkreten Ausgestaltung von Erklärungen, eine Universallösung ist nicht auszumachen. Vielmehr ist davon auszugehen, dass die Umsetzung von mehreren Faktoren – z. B. Zielgruppen oder zugrundeliegenden Datentypen – abhängig ist und individuell eine gut passende Lösung gefunden werden muss.

Einschätzung der Bedeutung von Entscheidungserklärungen für Zielgruppen*

Zielgruppe	Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit)			
	heute	Zukunft (5–10 Jahre)	Trend	
KI-Entwickler:innen	76 %	56 %	▼	-20 %
Domänenexpert:innen	59 %	59 %	▶	-1 %
Die Management-Ebene	38 %	57 %	▲	19 %
Endkund:innen, Endnutzer:innen	35 %	65 %	▲	31 %
Interne Prüfer:innen	41 %	57 %	▲	16 %
Externe Prüfer:innen	35 %	63 %	▲	28 %

Abbildung 7 – Umfrageergebnis: Erklärbarkeit heute besonders für KI- und Domänenexpert:innen wichtig, künftig wird eine vergleichbare Bedeutung für fast alle Zielgruppen prognostiziert.

* Aus der Grafik ist ablesbar, wieviel Prozent der Umfrageteilnehmer:innen die jeweilige Gruppe als bedeutende Zielgruppe für lokale Erklärbarkeit einschätzen.

Von den Teilnehmenden erwartete bzw. gewünschte Erklärungstypen (Mehrfachnennung war möglich)*

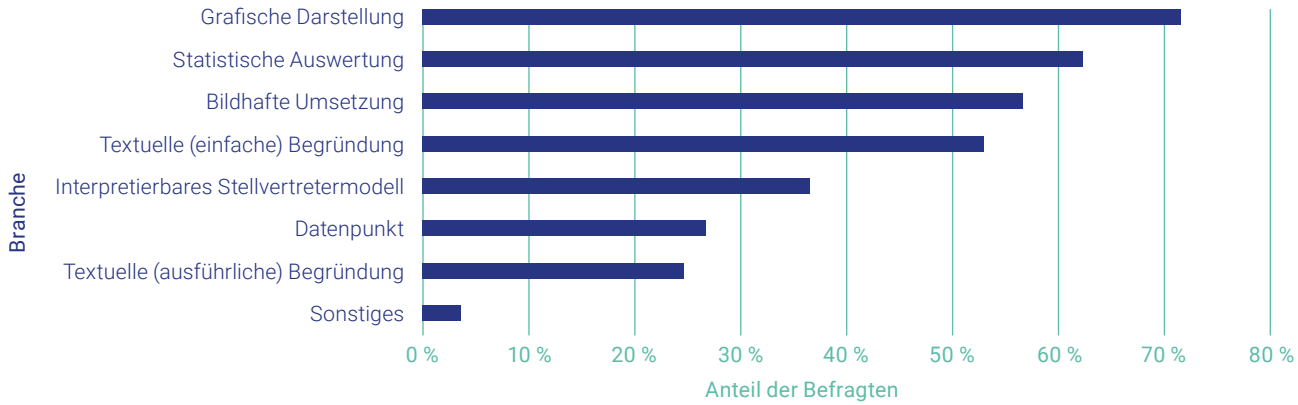


Abbildung 8: Umfrageergebnis: Grafische Darstellung von den meisten Teilnehmenden gewünscht.

* Aus der Grafik ist ablesbar, wie viel Prozent der Umfrageteilnehmenden die jeweilige Art der Umsetzung oder Verbesserung einer Erklärung erwarten. Unter „Sonstiges" wurden als weitere Möglichkeiten die Visualisierung durch Entscheidungsbäume bzw. der Wahrscheinlichkeiten der Ergebnisse sowie die Erläuterung des Einzelfalls bei Nachfrage benannt.

Erläuterung der Grafik:

- **Grafische Darstellung:** z. B. Darstellung des Einflusses einzelner Merkmale auf eine Entscheidung
- **Statistische Auswertung:** z. B. Darstellung ähnlicher Voraussetzungen und entsprechender Entscheidungen
- **Bildhafte Umsetzung:** etwa Highlighten ausschlaggebender Bereiche bei der Bilderkennung
- **Textuelle (einfache) Begründung:** z. B. Benennung des Hauptgrunds für Entscheidung
- **Interpretierbares Stellvertretermodell:** bspw. Generierung eines Entscheidungsbaums zur lokalen Approximation von komplexeren Modellen
- **Datenpunkt:** z. B. ein Datenpunkt, der eine gegensätzliche Entscheidung hervorgerufen hätte (Counterfactual Explanations), Muster-Datenpunkt für eine bestimmte Klasse (Prototyp)
- **Textuelle (ausführliche) Begründung:** etwa Erläuterung einzelner Schritte des Algorithmus

5 USE CASES FÜR ERKLÄRBARE KI

5 USE CASES FÜR ERKLÄRBARE KI

Die Umfrageergebnisse zur branchenspezifischen Bedeutung von Erklärungen für Einzelentscheidungen (lokale Erklärbarkeit) identifizieren die Gesundheitswirtschaft relativ eindeutig als eine Branche, in der sich KI ohne eine hinreichende Erklärbarkeit nicht etablieren wird (siehe Abbildung 6 in Kapitel 4). Auch für eine Reihe weiterer Branchen, wie die Finanz-, Produktions-, und Bauwirtschaft, die Prozessindustrie, die Energiewirtschaft sowie den Dienstleistungssektor, wird die hinreichende Erklärbarkeit mehrheitlich als unumgänglich angesehen, damit KI sich in diesen Branchen nachhaltig verfestigen kann. Während allerdings in diesen sechs Branchen die fehlende Akzeptanz für nicht erklärbare KI laut den Teilnehmenden der Umfrage vornehmlich in Vorbehalten von Kund:innen bzw. Anwender:innen gründet, werden in der Gesundheitswirtschaft bereits die regulatorischen Hürden als unüberwindbar für nicht erklärbare KI eingeschätzt.

Daher werden im Folgenden zunächst zwei Use Cases aus der Gesundheitsbranche vorgestellt. Im Anschluss folgen je ein Use Case aus dem Bereich der allgemeinen Produktionswirtschaft sowie der Prozessindustrie. Diese Anwendungsgebiete zeichnen sich durch einen hohen Automatisierungsgrad und besondere sicherheitstechnische Anforderungen aus. Sie sind daher anspruchsvolle Anwendungsfelder für KI-Anwendungen im Allgemeinen und gleichzeitig von hoher Bedeutung für den deutschen Wirtschaftsstandort. Die Informationen zu den Use Cases wurden im Wesentlichen aus den Interviews mit Expert:innen und Informationen aus der Literatur abgeleitet[6].

5.1 Use Cases Gesundheitswirtschaft

Im Medizinbereich werden KI-Algorithmen zur Lösung unterschiedlicher Problemstellungen eingesetzt. Bilddaten, beispielsweise Röntgenbefunde, digitalisierte Gewebeschnitte oder MRT-Scans, doch auch Textdaten (Arztbriefe und Befunde) oder sensorische Daten (EKG- und Blutdruckmesswerte) können mithilfe von KI-Algorithmen schneller und oft auch präziser analysiert werden. Die Auswertung der Analyseergebnisse kann dazu beitragen, frühzeitig Auffälligkeiten in den Daten zu entdecken und mit entsprechenden weiteren Untersuchungs- und Behandlungsschritten darauf zu reagieren.

Neben der wichtigen Aufgabe, Akzeptanz und Vertrauen seitens der Anwendenden und Betroffenen herzustellen, müssen gesundheitsspezifische regulatorische Vorgaben erfüllt sowie datenschutz- und datensicherheitsbezogene Vorkehrungen getroffen werden, um KI-Unterstützung bei kritischen Entscheidungen einbeziehen zu dürfen. Gerade die im Vergleich zu anderen Branchen strengen regulatorischen Hürden erschweren bzw. verzögern eine breite Verfügbarkeit von KI-Lösungen am Markt (BDVA Task Force 7 -Sub-group Healthcare 2020). Im Anschluss an die folgende Betrachtung von zwei Use Cases – KI-gestützte Bildanalyse und automatische Analyse von Arztbriefen – werden die regulatorischen Aspekte im Gesundheitsbereich gesondert diskutiert. Die beiden Use Cases wurden ausgewählt, da sie mit Bildverarbeitung und Natural Language Processing (NLP)[7] zwei typische praktische Anwendungsfälle von KI in der Gesundheitswirtschaft repräsentieren.

6 Dabei erfolgte die Formulierung auch hier ausschließlich durch das Team der Autor:innen, das somit die Verantwortung für vermeintliche übervereinfachte oder auch inkorrekte Darstellungen von Details trägt.

7 Bild- und Textdaten werden laut unserer Umfrage neben numerischen Daten am häufigsten verwendet (siehe Kapitel 4).

5.1.1 Use Case: KI-gestützte Bildanalyse histologischer Gewebeschnitte

Im Vergleich zur nicht invasiven medizinischen Bildgebung, z. B. Röntgen oder MRT, beruht die histologische Bildgebung auf entnommenen Gewebeproben, die mit Farbstoffen oder farbstoff-markierten Antikörpern angefärbt werden. Durch das Färben werden Strukturen und Zellen sichtbar gemacht. Histopathologische Gewebeschnitte von Proben, die einem Patienten oder einer Patientin mit Verdacht auf Krebs im Rahmen einer Biopsie entnommen wurden, werden von der Pathologin oder dem Pathologen nach der Anfärbung auf Auffälligkeiten untersucht. Derzeit werden diese Arbeiten zumeist noch manuell ausgeführt.

Im Rahmen entsprechender Forschungsansätze[8] wird erprobt, wie durch den Einsatz trainierter KI-Algorithmen Pathologinnen und Pathologen in ihrer Arbeit unterstützt werden können. Auf digitalen Scans der Gewebeschnitte werden vom KI-Modell Auffälligkeiten und Muster erkannt, die auf eine Erkrankung hinweisen. Der Mehrwert, der sich durch die Nutzung von KI in diesem Anwendungsfall ergibt, ist die Verbesserung der Qualität der Bildanalysen, was sich für Domänenexpert:innen vor allem im erhöhten Informationsgewinn niederschlägt (Nagpal et al. 2018). Das Ergebnis der KI-Analyse kann darüber hinaus zusätzliche Auffälligkeiten in den Daten aufdecken. Die Pathologin oder der Pathologe kann diese mit den eigenen Erfahrungen zusammenführen und so potenziell eine differenziertere Diagnose erstellen – beispielsweise indem die Gefahr, kritische Auffälligkeiten zu übersehen, verringert wird[9].

Bei den Eingangsdaten handelt es sich um Bilddaten, die in der Regel kein einheitliches Format haben und sich durch eine hohe Auflösung auszeichnen. Für die Analyse der Bilddaten werden vor allem Convolutional Neural Networks (CNNs) genutzt. Diese Black-Box-Modelle sind im Gegensatz zu White-Box-Modellen vor allem für die automatische Bildverarbeitung geeignet. Das liegt daran, dass komplexe Muster in Bildern implizit erkannt werden können, ohne dass die Entwicklerin oder der Entwickler Regeln vorgeben muss. CNNs kommen daher auch bei der KI-gestützten Bildanalyse histologischer Gewebeschnitte zum Einsatz. Dabei werden Lernverfahren vom Typ „Supervised Learning" angewendet[10].

In diesem Use Case soll die Ärztin oder der Arzt mittels automatisierter, KI-gestützter Anomalie-Erkennnung unterstützt werden. Das medizinische Fachpersonal entscheidet, inwiefern die Ergebnisse des KI-Systems einbezogen werden – und trägt somit auch die Verantwortung. Gleichzeitig werden vom System potenziell kritische Entscheidungen getroffen bzw. unterstützt, die das leibliche Wohl von Patient:innen maßgeblich beeinflussen und die bei Fehldiagnosen fatale Folgen haben können.

Zielgruppen und übergeordnete Ziele für die Nutzung erklärbarer KI

Die wichtigste Zielgruppe stellt das medizinische Personal dar, das das System einsetzen soll. Gleichzeitig müssen Zulassungsstellen – in diesem Fall die „Benannten Stellen", siehe 5.1.3 – adressiert werden. Schließlich spielt auch die KI-Entwicklerin oder der KI-Entwickler selbst eine wichtige Rolle, gerade hinsichtlich der (Weiter-)Entwicklung und Verbesserung des KI-Systems.[11]

Eine weitere Zielgruppe, für die Erklärbarkeit in Zukunft natürlich an Bedeutung gewinnen könnte, bilden die

8 Projekt EMPAIA des BMWi-Technologieprogramms Innovationswettbewerb KI (https://www.empaia.org/)

9 Eine zusätzliche Motivation für den Einsatz von KI stellt die Stratifizierung von Patient:innen dar, also die Einteilung dieser in verschiedene Behandlungsgruppen gemäß ihres individuellen Risikos. Darauf aufbauend lassen sich potenziell individuelle Therapieentscheidungen ableiten, was jedoch nicht mehr Fokus dieses Use Cases ist. Präzisionsmedizin und personalisierte Medizin können so perspektivisch weiterverfolgt und verbessert werden.

10 Prinzipiell kommen auch „Weakly Supervised Learning" und „Unsupervised Learning" sowie „Transfer Learning" in Betracht, um Modelle weiter zu verbessern, was in der Forschung auch untersucht wird.

11 Es handelt sich bei Anwendungsfällen, die mit diesem vergleichbar sind, häufig um Forschungsprojekte; weshalb entsprechende Forscherinnen und Forscher mit KI- oder Gesundheitsexpertise prinzipiell eine weitere Zielgruppe darstellen könnten. In diesem und den folgenden Use Cases sollen diese jedoch konsequent als Entwickler:in klassifiziert werden, falls sie entsprechende Systeme entwickeln. Aufgrund des Fokus auf praktische Umsetzungen werden Forschende, die allgemeine, medizinische Zusammenhänge untersuchen, nicht berücksichtigt. Erklärungen könnten dennoch solchen Forschenden dabei helfen, Assoziationen in den Daten (wie beispielsweise Korrelationen, die Hinweise auf kausale Zusammenhänge geben) und somit neue Biomarker zu entdecken, die für eine verbesserte Erkennung von Krankheiten allgemein, bzw. in diesem Fall zur Feindiagnostik von Krebstumoren, genutzt werden können.

Patient:innen selbst. Da das KI-System als Analysetool für Domänenexpert:innen (medizinisches Personal) verstanden wird, sollte es die Aufgabe von Zulassungsbehörden sein, angemessene, verpflichtende Anforderungen für Erklärbarkeit aus der Sicht von Patient:innen zu etablieren – und die Aufgabe der adressierten Domänenexpert:innen, persönliche Erklärungen für die Patient:innen zu geben. Da Letztere somit derzeit keine unmittelbare Zielgruppe darstellen, werden die Patient:innen hier nicht weiter explizit berücksichtigt.

Ein Ziel, das durch die Nutzung erklärbarer KI im Use Case angestrebt wird, ist die Identifikation von Kausalitätsbeziehungen, die gerade für Domänenexpert:innen wichtig ist – hier Pathologinnen und Pathologen. Entwickler:innen nutzen die Erklärungen, um die Konfidenz (Funktionalität, Robustheit und Stabilität) zu bestimmen und damit grundsätzliche Anfälligkeiten des Systems bezüglich diverser Störungen, wie z. B. statistischen Messfehlern, in Eingangs- oder Trainingsdaten zu identifizieren. Andererseits ist es auch das Ziel von Entwickler:innen, einen möglicherweise konkret existierenden Datenbias, d.h. einen sytematischen Fehler, in den Trainingsdaten zu finden (Fairness testen bzw. Datenbias aufdecken), da dies ebenfalls die Nichterkennung von Tumoren zur Folge haben kann. Außerdem müssen regulatorische Vorgaben bzw. Zulassungsanforderungen erfüllt werden, die derzeit jedoch größtenteils noch unklar sind.

Erklärbarkeitsanforderungen aus Perspektive der Zielgruppe(n)

Bei der Nutzung erklärbarer KI wird in dem hier betrachteten Use Case aus Sicht der Entwickler:innen insbesondere das Ziel verfolgt, die Konfidenz zu erhöhen. Dies erfordert nicht zwingend eine Modellerklärbarkeit. Es muss dieser Zielgruppe aber möglich sein, etwa einen möglichen Bias im Modell zu identifizieren und so das Risiko von Fehlentscheidungen zu minimieren.

Dafür ist eine alleinige Erklärung der Funktionsweise des Algorithmus nicht ausreichend – z. B., welche Schichten eines neuronalen Netzes für die Erkennung welcher Strukturen auf dem vermeintlichen Tumorbild verantwortlich sind. Vielmehr müssen Faktoren, die das Verhalten des KI-Systems bei der Entscheidungsfindung beeinflussen, untersucht werden: etwa ein möglicher Bias, der durch eine nicht ausgewogene Datengrundlage erzeugt wurde. Es muss vermieden werden, dass der Algorithmus während des Trainings Entscheidungskriterien lernt, die in der praktischen Anwendung nicht einsetzbar sind bzw. zu fehlerhaften Ergebnissen führen. Dies könnte beispielsweise die Klassifizierung histopathologischer Schnitte nach verwendetem Scanner oder Aufnahmedatum sein. Ein weiteres Beispiel ist die Nutzung bereits befundeter Daten. Patholog:innen haben womöglich bereits händisch Tumorbereiche markiert – und das KI-System lernt dann nur anhand dieser Markierungen, Tumore zu erkennen.

Ein Bias in den Trainingsdaten könnte auch dazu führen, dass die Nutzung der Anwendung für eine andere, beispielsweise deutlich jüngere, Patientengruppe zu schlechteren Ergebnissen führt. Ziel ist es, derartige Fehlerquellen bereits in der Entwicklung zu beseitigen. Jedoch sollte nach Möglichkeit auch das medizinische Personal die Möglichkeit haben, die Entscheidungen des Systems überprüfen zu können, um nachzuvollziehen, ob es sich um ein aus medizinischer Sicht belastbares Ergebnis handelt.

Domänenexpert:innen (das medizinische Fachpersonal) werden die Erklärungen im praktischen Einsatz nutzen. Auch wenn konkrete Kriterien für eine Zulassung in Bezug auf Erklärbarkeit noch nicht festgelegt sind, gehen Expert:innen davon aus, dass zumindest individuelle Entscheidungen zu einzelnen Patient:innen für Ärztinnen und Ärzte nachvollziehbar sein müssen. Eine fehlende granulare Erklärung der Funktionsweise des Algorithmus wäre in diesem Sinn folglich nicht zwingend ein Ausschlusskriterium. Vielmehr müssen Patholog:innen gezeigt oder erklärt bekommen, warum eine konkrete Entscheidung getroffen wurde, um auf dieser Grundlage mithilfe des eigenen Domänenwissens eine ent-

sprechende Einschätzung des Ergebnisses ableiten zu können. Hier können vor allem Darstellungen von Zwischenergebnissen hilfreich sein, z. B. die Segmentierung auffälliger Bildbereiche, um den Domänenexpert:innen die Beurteilung der Plausibilität eines vorliegenden Ergebnisses zu ermöglichen.

Obwohl es hinsichtlich der Erklärbarkeitsanforderungen der Zulassungsbehörden noch enorme Unklarheit gibt (siehe 5.1.3), sind gerade diese Anforderungen entscheidend für den späteren praktischen Einsatz von KI-Systemen. Derzeit wird für die Zulassung von KI im Medizinbereich eine Checkliste der Benannten Stellen[12] genutzt, die jedoch mit Blick auf die geforderte Erklärbarkeit unscharf bleibt (Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland 2020). Ein Punkt der Checkliste behandelt die Frage, ob sich der/die Entwickler:in bei der Implementierung auf erklärbare KI gestützt hat; bei einem weiteren wird thematisiert, inwiefern der/die Endnutzer:in dem Produkt Vertrauen entgegenbringt. Zum jetzigen Zeitpunkt können keine konkreten Zulassungsanforderungen bezüglich der Erklärbarkeit angegeben werden.

Erklärungsstrategien

Zur Erklärung von Entscheidungen neuronaler Netze existieren bereits Post hoc-Methoden, die insbesondere die Visualisierung der Ergebnisse fokussieren. Im Eingabebild werden einzelne Bereiche entsprechend ihres Einflusses auf das Ergebnis des KI-Algorithmus hervorgehoben. Auf dieser Grundlage kann der Domänenexperte bzw. die Domänenexpertin entscheiden, ob die Ergebnisse plausibel sind, die markierten Bildbereiche also so relevant sind, dass auch aus medizinischer Sicht diese als Grundlage für eine Klassifikation oder Segmentierung genutzt werden können.

Im konkreten Use Case der Bildanalyse von histopathologischen Gewebeschnitten werden vorwiegend neuronale Netze eingesetzt, um krankheitsrelevante Bildbereiche wie einen Tumor zu identifizieren. Der Fokus hier eingesetzter Erklärungswerkzeuge liegt auf der Erstellung einer Erklärung durch Visualisierung. Konkret können beispielsweise mit der Methode LRP Pixel im Eingangsbild identifiziert werden, die einen besonders hohen positiven oder negativen Einfluss auf das Klassifikationsergebnis haben. LIME erstellt lineare, sehr lokale Modelle, um so Einzelentscheidungen des neuronalen Netzes einfacher nachvollziehen zu können. Ein Vorteil von LIME ist die einfache Integration, LRP kann besonders schnell Erklärungen liefern. Beim Praxiseinsatz von LIME muss beachtet werden, dass die Methode für besonders hochdimensionale Eingabedaten eher weniger geeignet ist, sodass unter Umständen die Auflösung herunterskaliert werden muss. Beide Methoden sollten jedoch nicht ohne Auseinandersetzung mit deren Funktionsweise, sowie in diesem Fall nur mit Vorkenntnissen in Pathologie und neuronalen Netzen, angewendet werden.

Potenziell sind weitere Erklärungswerkzeuge, wie etwa Grad-CAM, Integrated Gradients oder DeepLIFT[13], die für die Anwendung auf neuronalen Netzen mit Bilddaten vorgesehen sind, für diesen Use Case anwendbar und alternative Umsetzungen von Visualisierungen für die Domänenexpert:innen möglich.

Ein weiterer Ansatz ist die Nutzung von Counterfactual Explanations. Diese können sowohl auf Bilddaten als auch anderen Datentypen angewendet werden. Die Grundidee ist dabei, mithilfe von Negativbeispielen die Nachvollziehbarkeit zu erhöhen. Konkret werden nötige Änderungen identifiziert, die bei einer Klassifikation die Einordnung in eine andere – möglichst nächste – Klasse bedingen würden. Beim Beispiel der Bilddaten werden zusätzliche Bilder synthetisch generiert, die den Eingabebildern möglichst ähnlich sehen, aber jeweils einer anderen Klasse zugeordnet werden (Goyal et al. 2019).

12 Benannte Stellen sind staatlich autorisierte Einrichtungen, die im Auftrag von Herstellern Konformitätsbewertungen durchführen, z. B. für die Zulassung von Medizinprodukten (Bundesinstitut für Arzneimittel und Medizinprodukte o. D.)

13 Pocevičiūtė et al. 2020 beschreiben weitere konkrete Möglichkeiten für den Einsatz von Erklärungswerkzeugen in der Pathologie, beispielsweise Excitation Backprop, PatternNet oder t-SNE (Pocevičiūtė et al. 2020).

USE CASE
KI-gestützte Bildanalyse histologischer Gewebeschnitte in der Kurzübersicht

TYP	Anomalieerkennung (Entscheidungsunterstützung)
KRITIKALITÄT	Sehr hoch (Medizinprodukt)
DATENTYPEN	Bilddaten (2D, 3D), digitalisierte Gewebeschnitte, hohe Auflösung
TYPISCHE KI-MODELLE	Neuronale Netze (CNNs, GANs), *(Erklärbarkeitsdefizite bzw. Black-Box-Modelle)*
(WICHTIGSTE) ZIELGRUPPEN und jeweilige übergeordnete Ziele für die Nutzung erklärbarer KI	**DOMÄNENEXPERT:INNEN** (medizinisches Personal): Kausalitätsbeziehungen finden
	ENTWICKLER:INNEN: Konfidenz (Robustheit, Stabilität) bestimmen, Fairness testen (Databias finden)
	ZULASSENDE BEHÖRDEN (Benannte Stellen): Zulassungsanforderungen prüfen
KONKRETE ANFORDERUNGEN an Erklärbarkeit	**DOMÄNENEXPERT:INNEN** (medizinisches Personal): Bewertung der Güte von Einzel- bzw. „lokalen" Entscheidungen
	ENTWICKLER:INNEN: Bewertung der Güte von Modellen und Bias in den Trainingsdaten aufdecken (über lokale Erklärungen)
	ZULASSENDE BEHÖRDEN: Reduzierung der Komplexität, „Nachvollziehbarkeit" (konkrete Anforderungen werden derzeit diskutiert)
GEEIGNETE ERKLÄRUNGSSTRATEGIEN	Entscheidungserklärungen (Post hoc), z. B. LRP, LIME

5.1.2 Use Case: KI-gestützte Textanalyse von Arztbriefen

In einem Arztbrief werden wichtige Daten der Krankengeschichte einer Patientin oder eines Patienten zusammengefasst. Das Dokument kann beispielsweise Informationen zu erfolgten Untersuchungen und entsprechende Befunde beinhalten. Arztbriefe werden unter anderem bei einer Überweisung an Fachärzt:innen oder bei der Entlassung aus dem Krankenhaus genutzt, um Informationen von Patient:innen an spezialisierte Fachärzt:innen oder die Hausärzt:innen weiterzugeben.

Bei der Erstellung einer Differentialdiagnose durch die Ärztin oder den Arzt werden die im Arztbrief festgehaltenen Beschwerden und Symptome als Grundlage genutzt. Im Rahmen einer Differentialdiagnose werden Krankheitsbilder, die eine ähnliche Symptomatik aufweisen, in einem ersten Schritt identifiziert und anschließend voneinander abgegrenzt mit dem Ziel, nicht relevante Krankheitsbilder auszuschließen. So kann in vielen Fällen eine sicherere Diagnose gestellt werden[14].

Mithilfe von KI kann dieser Prozess unterstützt werden[15]. Konkret werden Verfahren des Natural Language Processing (NLP) genutzt, um den Inhalt eines Arztbriefes automatisiert zu erfassen und auszuwerten. Ziel des Einsatzes von KI-Verfahren ist dabei, Ärzt:innen zu unterstützen, indem ihnen weitere mögliche Krankheitsbilder vorgeschlagen werden, deren Symptome mit denen der Patient:innen übereinstimmen. Den Ärzt:innen wird also ein größeres Möglichkeitsspektrum präsentiert und das Ergebnis des KI-Systems kann wie die Zweitmeinung einer Kollegin oder eines Kollegen genutzt werden. Zeitersparnis ist ein weiterer Vorteil: Durch NLP kann die Ärztin/der Arzt automatisiert einen schnellen Überblick über die – oft unstrukturiert vorliegende – Krankheitsgeschichte der Patient:innen erhalten. Außerdem ist der Austausch zwischen zwei Medizinern bzgl. ähnlicher Patient:innen oft schwierig, da vor allem die Identifikation dieser Patient:innen hauptsächlich durch persönliche Gespräche erfolgt und diese sehr zeitaufwendig sind. Ein automatisches Matching auf Grundlage von Arztbriefen wäre für diese Problematik ebenfalls sehr hilfreich.

Die in diesem Use Case der medizinischen Textanalyse eingesetzten KI-Modelle sind neuronale Netze. Konkret werden Transformer Networks eingesetzt, die in der Literatur mehrfach als State of the Art bei NLP-Aufgaben bezeichnet werden (Otter et al. 2018; Wolf et al. 2019; Nambiar et al. 2020). Deep Learning, wozu Transformer Networks aufgrund ihrer vielen Schichten gezählt werden, bietet gegenüber anderen Verfahren den Vorteil, dass auch latente Features erkannt werden: Also auch nicht unmittelbar erkennbare Informationen, die vor allem bei der Erfassung und Verarbeitung von Sprache eine Rolle spielen, wie beispielsweise indirekte Bezüge oder logische Schlussfolgerungen. Ein indirekter Bezug wäre die Beschreibung der Patient:in durch Wörter wie „er“ oder „sie“ bzw. „ihm“ oder „ihr“. Die Netze werden auf großen medizinischen Datensätzen vortrainiert (Unsupervised Learning) und anschließend, dem Konzept des Transfer Learning folgend, für den konkreten Anwendungsfall angepasst: Vorhersage der Diagnose durch Supervised Learning. Erst nach abgeschlossenem Training soll das KI-System dann in der Praxis eingesetzt werden.

Wie beim vorangegangen Use Case der KI-gestützten Bildanalyse histologischer Gewebeschnitte handelt es sich hierbei in erster Linie um ein KI-System zur Entscheidungsunterstützung. Es wird eine Ähnlichkeitsanalyse durchgeführt, die der Arzt oder die Ärztin in die Erstellung der Differentialdiagnose einfließen lassen kann. Die Verantwortung trägt das medizinische Fachpersonal. Ebenfalls ist die Kritikalität sehr hoch, da die Ergebnisse des KI-Systems potenziell genutzt werden, um gesundheitskritische Entscheidungen zu treffen.

14 Wird ein/e Patient:in beispielsweise mit Schmerzen in der Brust eingeliefert, so könnte dieses Symptom unter Umständen auf ein akutes Koronarsyndrom oder eine Lungenembolie hinweisen. Im Rahmen einer Differentialdiagnose geht es darum, diese und weitere mögliche Krankheitsbilder zu identifizieren und dann auf dieser Grundlage, beispielsweise anhand von Vorerkrankungen oder Risikofaktoren des/der Patienten:in, einzelne Krankheitsbilder auszuschließen (Strong Medicine 2018).

15 Im Artikel zur Differentialdiagnose (https://www.bmwi.de/Redaktion/DE/Artikel/Digitale-Welt/GAIA-X-Use-Cases/differentialdiagnose.html) werden Praxisbeispiel und aktuelle Herausforderungen beschrieben.

Zielgruppen und übergeordnete Ziele für die Nutzung erklärbarer KI

Domänenexpert:innen, d. h. Ärztinnen und Ärzte, sollen vom KI-System unterstützt werden, Symptome und Beschwerden unterschiedlichen Krankheitsbildern zuzuordnen. Durch den Einsatz erklärbarer KI soll der Informationsgewinn für die Domänenexpert:innen erhöht und so eine Entscheidungsunterstützung überhaupt erst möglich gemacht werden. Unter Zuhilfenahme von einschlägigen Publikationen, die konkret aufgetretene Beschwerden in Bezug auf Krankheiten thematisieren, wird für diese Zielgruppe angestrebt, entsprechende Kausalitätsbeziehungen zu identifizieren bzw. entsprechende Korrelationen zu finden. Die Konfidenz – insbesondere für Entwickler:innen wichtig – soll erhöht werden, indem die Datengrundlage durch Vereinfachung plausibilisiert wird, sodass zwei Patient:innen mit den gleichen Symptomen auch das gleiche Krankheitsbild zugeordnet wird.

Erklärbarkeitsanforderungen aus Perspektive der Zielgruppe(n)

Für Domänenexpert:innen bzw. das medizinische Fachpersonal spielen insbesondere die zentralen Entscheidungskriterien des Systems eine wichtige Rolle. Ärzt:innen müssen für jede Patient:in individuell entscheiden können, inwiefern z. B. ein vorgeschlagenes Krankheitsbild auf Basis vorliegender Symptome aus medizinischen Gründen in Frage kommt. Dafür muss ein System, das bei dieser Entscheidung unterstützen soll, Einzelfallentscheidungen inhaltlich begründen können. Denn nur auf Basis der Begründung kann von einem medizinischen Experten oder einer Expertin effizient bewertet werden, ob ein Kriterium, das für die Klassifikation ausschlaggebend war, entweder plausibel oder medizinisch nicht sinnvoll ist. Folglich kann der Arzt oder die Ärztin auch nur auf diesem Weg entscheiden, ob die Ergebnisse des KI-Systems in die Differentialdiagnose einfließen sollten oder nicht.

Aus Sicht der medizinischen Expert:innen ist eine Bereitstellung entsprechender Erklärungen etwa durch das Anzeigen ähnlicher Fälle (Patient:innen mit ähnlichen Risikofaktoren und entsprechender Symptomzusammensetzung) oder Erkenntnisse aus der Literatur (Publikationen aus dem Medizinbereich) sinnvoll umsetzbar. Neben der Tatsache, dass das Heranziehen medizinischer Fachliteratur als Datenquelle per se einen enormen Mehrwert für das KI-System bietet, kann die Anzahl und die Anerkennung passender Quellen ebenfalls als ein Indikator für die Konfidenz von Entscheidungen genutzt werden. Auch wenn dies eher mittelbare Konsequenz ist, kann zudem der zeitaufwendige Austausch zwischen Ärzt:innen durch erklärbare KI vereinfacht werden, wenn diese Krankheitsverläufe automatisiert vergleichbar macht und die Identifikation von Patient:innen mit ähnlichen Verläufen beschleunigt[16].

Während der Implementierung eines derartigen KI-Systems geht es den Entwickler:innen vor allem darum, frühzeitig Fehler zu erkennen, die durch medizinisch nicht plausible „Features" entstehen. Dies beinhaltet vor allem die Betrachtung der Datengrundlage. Hierbei sollten etwa die Verteilung und Häufigkeit von Diagnosen untersucht werden, sodass spätere Entscheidungen des Systems nicht auf Grundlage eines Bias in der Datenbasis getroffen werden. Dafür ist eine systematische Prüfung des Algorithmus hinsichtlich einzelner Variablen und Korrelationen von Variablen notwendig. Dabei sollte etwa für eine Eingangsvariable wie das Alter überprüft werden, ob sich durch ihre systematische Veränderung bzw. Variation die Vorhersage des KI-Systems wie erwartet verändert oder ob diese durch Besonderheiten in der Trainingsgrundlage beeinflusst wird, die nicht der Realität entsprechen. Der Fokus liegt hier auf der Überprüfung von Einzelfallentscheidungen (lokale Erklärbarkeit), die unter Verwendung entsprechender Erklärungswerkzeuge auch für KI-Modelle mit Black-Box-Komponenten generiert werden können. Die Datengrundlage für das KI-Modell wird vor allem durch Plausibilisierung nachvollziehbarer.

Wie zuvor bereits beschrieben, sind konkrete Anforderungen an erklärbare KI im Gesundheitsbereich aus Sicht der Zulassungsbehörden noch in der Diskussion. Grundsätzlich muss gegenüber den zulassenden Stellen dargelegt werden, dass das verfolgte Ziel – z. B. Steigerung der Effizienz, Verbesserung der Differentialdiagnose – durch das System erreicht wird. Dieser Punkt ist auch für weitere Zielgruppen wie das Klinikmanagement essenziell, da sich daraus Indikatoren für die Wirtschaftlichkeit ableiten lassen. Wie im ersten Use Case ist eine mögliche zukünftige Zielgruppe die der Patientinnen und Patienten, die sich für eine Erklärung der ihnen gestellten Diagnose interessieren, die jedoch auch hier nicht explizit berücksichtigt werden[17].

16 Auf Basis einer entsprechenden Hervorhebung von Patient:innenfällen mit großen Übereinstimmungen könnte zeitnah entschieden werden, ob ein weitergehender Austausch zwischen den behandelnden Ärzt:innen sinnvoll erscheint.

17 Die behandelnde Ärztin/der behandelnde Arzt erstellt mithilfe der KI eine Diagnose, die anschließend der Patientin/dem Patienten vermittelt wird. Die Besprechung der Diagnose erfolgt also nur direkt mit der behandelnden Ärztin/dem behandelnden Arzt; die Patientin/der Patient muss das Ergebnis des KI-Systems nicht selbstständig nachvollziehen können.

Erklärungsstrategien

In dem betrachteten Anwendungsfall werden zwei konkrete Strategien zur Gewährleistung der Erklärbarkeit verfolgt. Beide fokussieren insbesondere auf die Zielgruppe der Domänenexpert:innen (medizinisches Fachpersonal). Die Bereitstellung von Entscheidungserklärungen wird dabei jedoch nicht „Post hoc" über den Umweg eines zusätzlichen Werkzeugs sichergestellt (wie im ersten medizinischen Use Case). Stattdessen kann das Modell selbst Entscheidungserklärungen bereitstellen[18]. Grundlage zur Bereitstellung von Erklärungen sind Black-Box-Modelle (neuronale Netze), die durch Prototypen oder externe Wissenssammlungen erweitert werden, womit das Modell selbst medizinisch nachvollziehbare Gründe für einzelne Entscheidungen geben kann und die Anforderung lokaler Erklärbarkeit erfüllt.

Beim ersten Ansatz arbeitet das neuronale Netz mit Prototypen. Bei einem Prototyp handelt es sich im einfachsten Fall um eine einzelne repräsentative Instanz aus der Datengrundlage. In diesem Anwendungsfall wäre ein Prototyp ein Krankheitsbild mit entsprechenden Symptomen, abgeleitet aus einem Arztbrief. Für eine stärkere Generalisierung werden oft mehrere Instanzen zu einem repräsentativen Prototyp zusammengefasst (mittels Supervised Learning). Beispielsweise klagt eine an Grippe erkrankte Patientin über Schnupfen und Kopfschmerzen, außerdem hat sie eine Körpertemperatur von 39° C. Diese Symptome werden im Arztbrief vermerkt. Ein anderer Patient, bei dem auch Grippe diagnostiziert wird, hat starke Halsschmerzen, Husten und ebenfalls Fieber. Bei der Erstellung des Prototyps zum Krankheitsbild „Grippe" würden nun die Symptome der beiden Patient:innen zusammengefasst und vermerkt werden.

Mithilfe von Prototypen kann also die Datengrundlage für die Nutzer:innen verständlicher dargestellt werden. Dieser Erklärungsansatz unterscheidet sich von anderen, wie Bewertung des Einflusses eines Parameters oder Approximation des KI-Modells insofern, als durch die verbesserte Nachvollziehbarkeit der Datengrundlage das Modell – und dessen Entscheidungen – interpretierbarer werden.

Soll ein „neuer" Arztbrief klassifiziert werden, so werden die beschriebenen Charakteristiken (Symptome wie beispielsweise Müdigkeit, Schnupfen, Halsschmerzen und erhöhte Temperatur) mit den in den einzelnen Prototypen benannten abgeglichen und der Prototyp mit den meisten Übereinstimmungen ausgewählt. Dafür können entsprechende Distanzfunktionen oder Klassifikationsverfahren wie K-Nearest-Neighbor verwendet werden. Nachdem der ähnlichste Prototyp (in diesem Beispiel „Grippe") identifiziert wurde, kann dieser mit dem Arztbrief verglichen und die entscheidenden Übereinstimmungen – Schnupfen, Halsschmerzen und erhöhte Temperatur –hervorgehoben werden, um die Diagnose des KI-Algorithmus nachvollziehbar zu machen.

Im zweiten Ansatz werden neuronale Netze mit externen Wissenssammlungen kombiniert. Diese „Knowledge-Bases" können Publikationen oder allgemein medizinische Werke sein, in denen Krankheiten beschrieben sind. Das neuronale Netz lernt auf dieser Datengrundlage den Zusammenhang von Symptomen und Krankheiten. Anschließend kann das Modell mit Arztbriefen weiter trainiert werden. Die Knowledge-Bases können als hochdimensionale Wissensgraphen verstanden werden, in denen die Repräsentationen thematisch ähnlicher Publikationen nah beieinander platziert werden. Soll nun ein „neuer" Arztbrief klassifiziert werden, so trifft das KI-Modell eine Entscheidung, die direkt auf die Aussagen der Publikationen aus der Wissensbasis zurückzuführen ist. Dadurch können Schlussfolgerungen leicht überprüft werden und die Nachvollziehbarkeit des Modells wird erhöht. Bei einer Entscheidung werden dem Arzt oder der Ärztin zu einzelnen Abschnitten des Arztbriefes relevante Publikationen angezeigt, die den betrachteten Sachverhalt beschreiben – beispielsweise konkrete Krankheitsbilder, die oft bei den beschriebenen Symptomen auftreten. So kann der Arzt bzw. die Ärztin gegebenenfalls über die Glaubwürdigkeit der Publikation auch Schlüsse auf die Glaubwürdigkeit der Entscheidung des Algorithmus ziehen.

18 In der Literatur sind widersprüchliche Angaben zur Bezeichnung derartiger Modelle zu finden. In einigen Quellen werden diese Modelle (trotz Black-Box-Anteil) als Ante-hoc klassifiziert (Sokol und Flach 2019; Holzinger 2018). In der Literatur ist jedoch umstritten, ob Ante-hoc-Erklärbarkeit eine Eigenschaft ist, die nur White-Box-Modelle für sich in Anspruch nehmen dürfen oder ob auch Black-Box-Modelle, die gewisse Erklärungen liefern, diese „ante hoc" zur Verfügung stellen. Im Folgenden wird Ante-hoc-Erklärbarkeit nur für White-Box-Modelle verwendet und explizit darauf hingewiesen, wenn auch Black-Box-Modelle gemeint sind, ohne dass dafür „Post hoc"-Erklärbarkeitswerkzeuge zum Einsatz kommen.

USE CASE
KI-gestützte Textanalyse von Arztbriefen in der Kurzübersicht

TYP	Ähnlichkeitsanalyse (Entscheidungsunterstützung)
KRITIKALITÄT	Sehr hoch (Medizinprodukt)
DATENTYPEN	Textdaten: Arztbriefe
TYPISCHE KI-MODELLE	Neuronale Netze (Transformer Networks) *(Erklärbarkeitsdefizite bzw. Black-Box-Modelle bei alleiniger Nutzung)*
(WICHTIGSTE) ZIELGRUPPEN und jeweilige übergeordnete Ziele für die Nutzung erklärbarer KI	**DOMÄNENEXPERT:INNEN** (medizinisches Personal): Informationsgewinn erhöhen, Kausalitätsbeziehungen finden
	ENTWICKLER:INNEN: Konfidenz (Robustheit, Stabilität) bestimmen
	ZULASSENDE BEHÖRDEN (Benannte Stellen): Zulassungsanforderungen prüfen
KONKRETE ANFORDERUNGEN an Erklärbarkeit	**DOMÄNENEXPERT:INNEN** (medizinisches Personal): Entscheidungsunterstützung ermöglichen durch inhaltliche Begründungen (lokale Erklärbarkeit)
	ENTWICKLER:INNEN: Tieferes Verständnis der Funktionsweise (durch lokale Erklärbarkeit) → Verbessern der Systeme
	ZULASSENDE BEHÖRDEN: Reduzierung der Komplexität, „Nachvollziehbarkeit“ (konkrete Anforderungen derzeit Diskussion)
GEEIGNETE ERKLÄRUNGSSTRATEGIEN	Entscheidungserklärungen durch Prototypen und externe Wissensbasen in Verbindung mit neuronalen Netzen

5.1.3 Regulatorik und Zertifizierung in der Gesundheitswirtschaft

Vor allem aufgrund der Kritikalität und Datensensibilität sind Zertifizierungsanforderungen im Medizinbereich besonders hoch. In den letzten Jahren wurden bereits einige Anpassungen hinsichtlich der Zulassung von Software im bzw. als Medizinprodukt umgesetzt. Jedoch ist die Formulierung und Umsetzung konkreter Anforderungen an KI-Systeme derzeit noch in Arbeit.

Ab 26. Mai 2021 tritt die neue Verordnung für Medizinprodukte (MDR) in Kraft und wird die bisher geltende Medical Device Directive (MDD) ablösen. Die MDD der Europäischen Union wurde bisher durch das Medizinproduktegesetz in deutsches Recht übertragen. Die MDR enthält im Vergleich zur MDD einige Änderungen auch hinsichtlich der Zulassung von Software, unter die KI-Anwendungen fallen. So wird Software fortan in höhere Risikoklassen eingeordnet, sodass auch höhere Anforderungen entstehen. Die MDR beschreibt vier verschiedene Risikoklassen, die Einteilung der Anwendungen richtet sich nach dem Verwendungszweck. Es werden konkrete Vorgaben beispielsweise zu Entwicklung, Validierung, Verifizierung der Funktionalität, Produktion und Überwachung der Algorithmen beschrieben, die wiederum durch die „Benannten Stellen" überprüft werden. Benannte Stellen sind staatlich autorisierte Einrichtungen, die im Auftrag von Herstellern Konformitätsbewertungen durchführen, beispielsweise zur Zulassung von Medizinprodukten (Bundesinstitut für Arzneimittel und Medizinprodukte o. D.; acatech 2020).

Hinsichtlich der Zertifizierung von KI-Systemen gibt es jedoch keine spezifischen Standards, sodass oft unklar bleibt, wie die konkreten Anforderungen umzusetzen sind. Zudem gibt es zahlreiche Normen und Richtlinien, die potenziell auch für KI-Systeme Anwendung finden. Für die Zertifizierung von KI-Systemen hat daher das Johner Institut – ein in der Branche bekanntes Unternehmen, das Beratung bei der Zulassung von Medizinprodukten anbietet – eine Checkliste als Orientierungshilfe entwickelt, die von den Benannten Stellen als Grundlage für eine eigene Checkliste genutzt wird. Diese beschränkt die Zulassung von Software im Medizinbereich ausdrücklich auf nicht selbstlernende KI-Systeme. Demnach können derzeit keine KI-Algorithmen zugelassen werden, die während der Anwendung weiterlernen, sich also verändern. Am Beispiel der automatisierten Tumorerkennung auf Bildern könnten neue Daten von Patient:innen, bei denen dem Arzt oder der Ärztin aufgefallen ist, dass diese vom KI-System nicht korrekt klassifiziert wurden, genutzt werden, um die Datengrundlage zu erweitern. So könnte der Algorithmus trainiert werden, spezielle Tumorvarianten, die er zuvor noch nicht erkannt hat, zuverlässiger korrekt zu klassifizieren. Des Weiteren wird eine erklärbare KI angestrebt, deren Entscheidungen nachvollziehbar sind.

Folgende Anforderungen mit Bezug zur Erklärbarkeit werden formuliert:

- Der Hersteller sollte sich damit beschäftigt haben, inwiefern erklärbare KI bei der Nachvollziehbarkeit des entwickelten Modells unterstützen kann – vor allem während des Entwicklungsprozesses.

- Zudem sollte untersucht werden, inwiefern einfachere und damit interpretierbarere Modelle eingesetzt werden könnten. Hinsichtlich der Interaktion mit Nutzenden soll geprüft werden, inwieweit diese dem System vertrauen oder Entscheidungen nachprüfen wollen.

- Weitere Anforderungen umfassen unter anderen das Risikomanagement (Bewertung der Risiken, die durch den Einsatz von KI oder deren nicht vorgesehene Nutzung entstehen, z. B. bzgl. Eingabewerten oder Gruppen von Patient:innen), die Güte und Performanz, verwendete Trainingsdatensätze – Umfang, Herkunft, möglicher Bias – die Reproduzierbarkeit der Ergebnisse und die Erstellung eines Post-Market-Surveillance-Plans zur Sicherstellung der Modellgüte auch nach der Zulassung.

Die Aspekte der Checkliste werden von der Benannten Stelle individuell geprüft, sodass es sich immer um Einzelfallentscheidungen handelt und einzelne Anforderungen auch unterschiedlich ausgelegt werden können. Aus Sicht der Systementwicklung ist eine Konkretisierung der Zulassungsanforderungen an ein KI-System bedeutend, um diese angemessen adressieren zu können. Vor allem konkrete Anforderungen an Erklärbarkeit sind unklar. Entsprechende Vorgaben könnten jedoch insbesondere Unternehmen während des Entwicklungsprozesses unterstützen, Zertifizierungen beschleunigen und die Sicherheit der Patient:innen erhöhen. Erklärbarkeit könnte zudem dazu beitragen, künftig selbstlernende KI-Systeme zu zertifizieren. Kontinuierlich weiterlernende KI birgt beispielsweise die Chance, Diagnosen und Prognosen durch stärkere Individualisierung zu verbessern. Gleichzeitig wird jedoch die Abschätzung von Risiken und die

Überprüfung erschwert (Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland 2020; Arbeitsgruppe Gesundheit, Medizintechnik, Pflege 2019).

In den USA wurden bereits Vorschläge für eine Zulassung selbstlernender KI-Systeme erarbeitet. Die U.S. Food and Drug Administration (FDA) beschreibt in einem ersten Entwurfspapier einen „Total Product Lifecycle Regulatory"-Ansatz, der einen zuvor festgelegten Plan für Änderungen – inklusive der Art der Änderung – nach Zulassung des Systems beinhaltet. Diese Änderungen müssen hinsichtlich möglicher Risiken für die Patient:innen überprüft werden können. Beispielsweise müssen neu hinzugefügte Daten ebenfalls qualitätsgesichert werden. Zusätzlich muss begründet werden, warum die neuen Daten erforderlich sind, und das konkrete Ziel des neuen Trainings des Algorithmus muss erklärt werden. Kriterien zur Performanzevaluation für den Vergleich des vorherigen Systems und des neu trainierten sollen aufgestellt und Transparenz bzgl. Änderungen für die Nutzenden hergestellt werden. Im von der FDA im Januar 2021 veröffentlichten Aktionsplan ist beschrieben, dass das zuvor erstellte Entwurfspapier in diesem Jahr entsprechend der Ergebnisse erfolgter Diskussionsrunden überarbeitet werden soll. Außerdem soll die Herstellung von Transparenz für die Nutzer:innen von KI-Systemen weiterverfolgt und stärker fokussiert werden (U.S. Food & Drug Administration 2020, 2021; Arbeitsgruppe Gesundheit, Medizintechnik, Pflege 2019).

5.2 Use Cases Produktionswirtschaft

Die Anforderungen der Produktionswirtschaft an Erklärbarkeit spiegeln wider, dass hier in der Regel deutlich mehr Vorwissen zu den untersuchten Prozessen zur Verfügung steht als in der Gesundheitswirtschaft. Statt des menschlichen Körpers stehen hier durch menschliches Know-how entworfene Produkte, Maschinen, Anlagen oder Prozesse und insbesondere Wirtschaftlichkeitsaspekte im Fokus von Software- und KI-Systementwicklungen.

In der Produktionswirtschaft reichen die Anwendungsfelder von Analyseaufgaben wie z. B. zur Maschinenüberwachung oder Qualitätskontrolle über Planungsunterstützung bis hin zu autonomen Systemen, etwa fahrerlosen Transportrobotern oder KI-gestützter Prozessführung. Auch erweist sich das gesamte Spektrum der Mensch-Maschine-Interaktion als ein weiter entwickelbares Anwendungsfeld.

In der Produktionswirtschaft unterscheidet man grundsätzlich zwischen der Fertigungsindustrie, wo Produkte als abzählbare Einheiten hergestellt werden, und der verfahrenstechnischen Industrie (Prozessindustrie). Durch die Unterschiede der Fertigungsprozesse ergeben sich Unterschiede im Gefahrenpotenzial. Während mögliche Störfalle in der Prozessindustrie, z. B. bei der Verarbeitung explosiver und giftiger Stoffe, weitreichende Folgen für Mensch und Gesellschaft in einer größeren Umgebung nach sich ziehen können, haben mögliche Unfälle in der Fertigungsindustrie eher lokalere Auswirkungen auf betroffene Personen und die Umgebung. Entsprechend sind die Zulassungsanforderungen für Systeme in der Prozessindustrie erheblich strenger, woraus sich gewisse Auswirkungen für KI und speziell die Erklärbarkeit von KI ergeben.

Auch hier werden zwei Use Cases vorgestellt. In Abschnitt 5.2.1 findet sich eine Beschreibung eines Use Cases zu KI-gestützter Maschinenzustandsüberwachung – eine der typischsten und auf viele Produktionsbereiche anwendbare KI-Anwendung auf Basis numerisch-tabellarischer Daten. In Abschnitt 5.2.2 wird ein Use Case zur KI-gestützten Prozessführung dargestellt, in dem KI-Komponenten in ein größeres Gesamtsystem eingebettet werden, das sicherheitskritische Prozesse in chemischen Anlagen zwar unter menschlicher Aufsicht steuert, aber zu einem wesentlichen Teil autonom ist. Der Use Case wurde ausgewählt, weil das KI-gestützte Gesamtsystem während des Betriebs weiterlernen muss („on-the-job-learning"). Dies ist bei Entscheidungsunterstützungssystemen, die auf vortrainierte – „pretrained" – Modelle zurückgreifen, zumeist nicht der Fall.

Während sowohl in der „diskreten" Fertigungsindustrie als auch in der Prozessindustrie das Schaffen von Akzeptanz und Vertrauen mittels Erklärbarkeit von großer Bedeutung ist, unterscheiden sich die Vorgaben, die für eine Zulassung erfüllt werden müssen, doch an einigen Stellen recht deutlich. In Abschnitt 5.2.3 werden regulatorische Aspekte der Produktionswirtschaft und entsprechende Unterscheidungen diskutiert.

5.2.1 Use Case KI-gestützte Maschinenzustandsüberwachung

Stillstände von einzelnen Apparaten, Maschinen oder Anlagen können schnell enorm teure Produktionsausfälle zur Folge haben. Dies gilt besonders dann, wenn sie in komplexe Produktionsprozesse eingebettet und dort für kritische Verarbeitungsschritte verantwortlich sind. Effektive Frühwarnsysteme, die Maschinenfehlverhalten oder Wartungsnotwendigkeiten anzeigen, können mögliche Ausfälle reduzieren oder ganz vermeiden sowie ein Wartungsmanagement wirtschaftlicher gestalten.

Denkbare Anwendungsfälle existieren in sämtlichen Teilbranchen der fertigenden Industrie. Während des Maschinen- oder Anlagenbetriebs fallen in aller Regel große Mengen an Daten an, die üblicherweise Zeitreihen numerischen Typs sind. Im geeignetsten Fall geben entsprechende Anlagen zusätzlich explizite Fehlercodes aus, wenn ein Maschinenausfall eingetreten ist oder unmittelbar bevorsteht. Unter Berücksichtigung des Zeitpunkts, wann ein Fehlercode gesendet wurde, können dann im Nachhinein die aufgezeichneten Datenpakete mit entsprechenden „Labels“ versehen werden, die den erfolgten oder sich ankündigenden Eintritt eines Fehlers anzeigen. Falls genügend Betriebs- und Ausfalldaten zur Verfügung stehen, ergibt sich ein klassisches Anwendungsszenario für Supervised-Learning-Verfahren, um Anomalien im Maschinenverhalten möglichst frühzeitig zu erkennen (Condition Monitoring).

Diese Informationen können Domänenexpert:innen enorm helfen, Instandhaltungsmaßnahmen entweder auf Basis eigener Expertise oder auf Basis entsprechender Modelle zu planen und den Zeitpunkt der Instandhaltungsmaßnahme unter Kenntnis des Maschinenzustandes genau festzulegen (Predictive Maintenance)[19].

Prinzipiell eignen sich verschiedene KI-Verfahren, wie z. B. Support Vector Machines, sehr gut für entsprechende Aufgabenstellungen von Condition Monitoring bzw. von Predictive Maintenance.

Im vorliegenden Fall sollen entsprechende Überwachungssysteme allerdings möglichst nachvollziehbare und lokalisierbare Angaben über potenzielle Anomalien des Maschinenverhaltens ausgeben. Dabei können statistische Informationen zu Messtechnik und Sensordaten als gegeben angenommen werden. Daher wird im Use Case einerseits ein Ansatz über Ensembles von Bayes-Netzen skizziert (Ansatz 1) und andererseits ein wissensbasierter Ansatz verfolgt (Ansatz 2).

Prinzipiell ist hier wie im Use Case zur KI-gestützten Bildanalyse histologischer Gewebeschnitte die Aufgabenstellung, Anomalien im Maschinenverhalten zu erkennen und Entscheidungsunterstützung zu liefern. Konkrete Klassifikationsergebnisse können häufig direkt für die Instandhaltungsplanung genutzt werden. Gleichzeitig ist die Kritikalität der Anwendung als hoch zu bewerten, da von Fehlklassifikationen bzw. nicht erkannten Anzeichen eines Maschinen- oder Anlagenschadens vor allem enormer wirtschaftlicher Schaden ausgehen kann.

Zielgruppen und übergeordnete Ziele für die Nutzung erklärbarer KI

Wichtigste Zielgruppe für Erklärungen sind bei diesem Anwendungsfall die Domänenexpert:innen (Instandhaltungs- bzw. „Maintenance Reliability“-Teams), bei denen die Verantwortung zur Festlegung von Wartungszyklen sowie die Verantwortung zum Initiieren von Wartungsvorgängen liegt. Das Kernziel für die Verwendung erklärbarer KI im Use Case ist es, Entscheidungen qualitativ unter Berücksichtigung des Prozessverständnisses der Domänenexpert:innen zu plausibilisieren (Kausalitätsbeziehungen „finden“). Von ähnlich wichtiger Bedeutung ist es dabei, die Hauptzielgruppe, in der Regel Ingenieur:innen, auch fortlaufend von der statistischen Aussagekraft von Einzelentscheidungen, z. B. der Robustheit gegenüber variierenden Messfehlern, zu überzeugen (Konfidenz bestimmen).

19 Für eine dedizierte, KI-gestützte Planung des optimalen Zeitpunkts für eine Instandhaltungsmaßnahme (Predictive Maintenance) ist es streng genommen erforderlich, ein Degradationsmodell für jeden einzelnen Schadenstyp zur Verfügung zu haben. Auf Basis entsprechender Modelle kann die Restlebenszeit einer Maschine oder Anlage dann mittels Regressionsverfahren geschätzt werden. Da geeignete Modelle jedoch häufig mangels entsprechender Schadensfälle nicht aufgestellt werden können, ist oft auch von Predictive Maintenance die Rede, wenn eigentlich nur eine Erkennung von Anomalien (Condition Monitoring) im Maschinen- oder Anlagenverhalten durchgeführt wird.

Zuletzt kann es auch ein wünschenswertes Ziel sein, Interaktionsmöglichkeiten für Domänenexpert:innen zu verbessern, damit diese auch selbst mittels entsprechender Eingaben die Erklärungen von KI-Systemen oder sogar die Systeme selbst verbessern können. Dies kann als eine grundsätzliche Motivation der Entwickler:innen von KI-Systemen verstanden werden, während die konkreten diesbezüglichen Anforderungen durch die Domänenexpert:innen definiert werden..

Die Entwickler:innen der KI-Systeme sind mittelbar ebenfalls Stakeholder, da sie – zumindest in der Einführungsphase entsprechender KI-Systeme – z. B. Schwellenwerte festzulegen haben, die bestimmen, wie ausgeprägt eine Anomalie sein muss, damit der Domänenexperte bzw. die Domänenexpertin eine Warnmeldung erhält. Diese Motivation liegt jedoch gleichfalls nah an dem übergeordneten Ziel, das die Domänenexpert:innen verfolgen (Konfidenz bestimmen), weshalb die Zielgruppe der KI-Entwickler:innen an dieser Stelle nur am Rande berücksichtigt wird.

Ein weiteres, hier nicht betrachtetes Ziel, das für Entwickler:innen relevant werden kann, ergibt sich, wenn Maschinen oder Anlagen ähnlicher Bauart analysiert werden sollen (Übertragbarkeit testen) und zu diesem Zweck in den gleichen Datenpool einspeisen. Da in dem vorliegenden Use Case Modelle nur individuell, auf einzelne Anlagen bezogen, trainiert werden, wird dieser Aspekt hier nicht berücksichtigt.

Erklärbarkeitsanforderungen aus der Perspektive der Zielgruppe(n)

Generell kann davon ausgegangen werden, dass Domänenexpert:innen zumindest nachhaltig von der Modellgüte überzeugt werden müssen. Aufgrund der entsprechenden Verantwortung für Entscheidungen, die gegebenenfalls unter Zeitdruck gefällt werden müssen, darf die Ermittlung von Erklärungen nicht zu lange dauern; im Idealfall sollte für die Zielgruppe sogar erkennbar sein, in welchen Betriebspunkten die Aussagekraft des Modells besonders gut oder besonders schlecht ist. Für KI-affine Anwender:innen kann es zudem eine wichtige, sogar essenzielle Anforderung sein, im Detail nachvollziehen zu können, wie ein Modell generiert wurde. Ferner sollten bereitgestellte Konfidenzwerte nach Möglichkeit auch statistische Fehlerverteilungen berücksichtigen, falls diese verfügbar sind (Erklärbarkeit von Einzelentscheidungen). Mit Blick auf die Zielgruppe und zur Berücksichtigung von eventuell gegebenen Fehlerverteilungen eignen sich vor allem statistische Konfidenzindikatoren wie Effekt- bzw. Signalstärken.

Zuletzt muss ein erklärbares KI-System eine intuitive und effiziente Interaktion ermöglichen, damit Nutzende dessen Verhalten zumindest stichprobenartig validieren können oder wünschenswerterweise sogar das zugrunde gelegte Modell nutzerseitig modifizieren und verbessern können. Letzteres kann in einer Minimalvariante geschehen, etwa durch anpassbare Schwellenwerte, oder auch durch einen direkt mitgedachten „Human-in-the-Loop"-Ansatz realisiert werden. Im letztgenannten Fall, der eine Fachexpertin oder einen Fachexperten als weitere „Datenquelle" einbezieht, ist es explizites Ziel, dass auf Basis der Mensch-Maschine-Interaktion das System erklärbarer wird und es gleichzeitig durch den Input von Expert:innen weiter dazulernt.

Eine zusätzliche, zumindest wünschenswerte Eigenschaft erklärbarer KI ist, dass Domänenexpert:innen bei Bedarf auch alternative „Erklärungskonzepte" erhalten können, wenn solche präferiert werden. Ein einfaches Beispiel wäre etwa, dass bei der Generierung von Erklärungen zu Bauteilen bei der Einzelerklärung über deren Farbe statt über deren vermeintlich komplizierte Typenbezeichnung argumentiert wird.

Erklärungsstrategien

Falls gemäß den Erklärbarkeitsanforderungen von Domänenexpert:innen explizit statistische Konfidenzwerte und die Berücksichtigung statistischer Zusatzinformationen gefordert sind, müssen entsprechend geeignete Machine-Learning-Modelle verwendet werden. Obwohl mittels geeigneter Validierungsstrategien (z. B. „Cross-Validation") die Vorhersagegüte jedes Modells auf Basis eines gegebenen Datensatzes bestimmt werden kann, können vermeintlich verfügbare statistische Zusatzinformationen, wie Eintrittswahrscheinlichkeiten von Ereignissen, nur von bestimmten Modelltypen umfassend berücksichtigt werden. Falls solche Zusatzinformationen vorliegen und explizit zur Erkennung von Anomalien und zur Bestimmung statistischer Konfidenzwerte herangezogen werden sollen, um die Nachvollziehbarkeit oder zumindest die Plausibilität zu erhöhen, eignen sich vor allem probabilistische Modelle, z. B. Bayes-Netze. Mit Bayes-Netzen können Wahrscheinlichkeiten von Ereignissen und deren Abhängigkeiten untereinander dargestellt werden.

Entscheidend für die Akzeptanz des KI-Gesamtprodukts ist es in diesem Use Case, die Domänenexpert:innen nachhaltig davon zu überzeugen, dass ein Modell sich „richtig" verhält. Aus diesem Grund sollte den Zielpersonen abseits des operativen Einsatzes die Möglichkeit einer qualitativen oder quantitativen Modellvalidierung

eröffnet werden, etwa über Simulation. Allerdings empfiehlt sich dafür, Domänenexpert:innen Szenarien selbstbestimmt wählen zu lassen, da eigens ausgewählte, „typische" Abläufe möglicherweise als zu selektiv empfunden werden könnten. Es ergeben sich unterschiedliche Wege, dies modellseitig konkret umzusetzen. Eine vergleichbar einfache Möglichkeit ist die geeignete Darstellung der Input-Output-Beziehungen des mittels Trainingsdaten generierten Ursprungsmodells.

Vielfach ist es jedoch erforderlich, den Domänenexpert:innen zunächst die grundsätzlichen, qualitativen Wirkmechanismen eines Modells über die Veranschaulichung von Zwischenergebnissen begreiflich zu machen. In diesen Fällen ist es mitunter sinnvoll, wenn der Zielgruppe ein White-Box-Surrogat-Modell bereitgestellt wird, das etwa auf Basis des Ursprungsmodells und/oder realer Betriebsdaten generiert werden kann. Hier können im einfachsten Fall z. B. Entscheidungsbäume herangezogen werden. Dieser Surrogat-Ansatz hat den Vorteil, dass Domänenexpert:innen Entscheidungsgrenzen und Zwischenergebnisse einfacher und schneller interpretieren können. Dies gilt natürlich insbesondere, wenn das Surrogat-Modell von mechanistischer Bauweise ist, d. h., ohnehin auf Basis physikalischer Gesetzmäßigkeiten erstellt wurde. Der Ansatz über adaptierbare mechanistische Surrogat-Modelle kann aufwendig sein. Möglicherweise existieren aber aus dem Designprozess der Maschine oder Anlage entsprechende Modelle, die zur Nutzung als Surrogat-Modell herangezogen werden können. Die Parameteradaption der Surrogat-Modelle an konkrete Maschinen kann dann mittels Ausgleichsrechnung bzw. Parameterschätzung erfolgen.

Der Aufwand zur Generierung eines mechanistischen Modells kann sich vor allem bei großer Ähnlichkeit von Maschinen und Anlagen und hinreichend einfach übertragbaren Prozessabläufen auszahlen. Ein Beispiel sind Turbokompressoranlagen, die in ihrer Baugröße enorm variieren können, aber ungeachtet dessen physikalisch betrachtet alle sehr ähnlich funktionieren und stets aus ähnlichen Teilmodulen bestehen. Ein weiterer Vorteil solcher mechanistischer Surrogat-Modelle ist zudem, dass Schadensfälle oder Betriebsdaten in ungewöhnlichen oder potenziell gefährlichen Betriebspunkten auch mittels Simulation ermittelt werden können, falls zunächst ein Mangel an entsprechenden Daten besteht. Dies kann einerseits das Prozessverständnis von Domänenexpert:innen erhöhen – hilfreich für Schulungen o. Ä. –; andererseits können so, falls nötig, auf künstliche Weise zusätzliche Daten für das Trainieren des KI-Modells (hier Bayes-Netz) generiert werden.

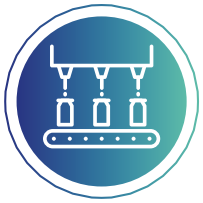
Eine Alternative, den Anforderungen der Modellerklärbarkeit von Domänenexpert:innen noch umfassender gerecht zu werden, ist, die Nutzer:innen direkt bei der Generierung der Modelle und der zugehörigen Erklärungsansätze zu beteiligen. Hier gibt es erste vielversprechende Ansätze entsprechender Machine-Learning-Verfahren auf der Basis von Wissensgraphen. Aktuell werden Ansätze erprobt[20], die erstmalig induktives logisches Programmieren und Methoden des Reinforcement Learning kombinieren, um auf diese Weise nachvollziehbare Machine-Learning-Verfahren zu erhalten, die darüber hinaus ein sogenanntes „Human in the Loop"-Konzept beinhalten. Mit entsprechenden Methoden soll es damit für die Branchenexpert:innen bald möglich werden, mittels natürlicher Sprache in Interaktion mit dem KI-System zu treten und somit – anhand vorliegender Beispiele für Maschinenanomalien und „normalen" Maschinenverhaltens – selbst zu definieren, welche Erklärungen für sie geeignet sind. Auf diese Weise können vollständig erklärbare Modelle generiert werden, anhand derer Nutzer:innen eines entsprechenden Condition-Monitoring-Systems mittels natürlichsprachlicher Erklärungen dargelegt werden kann, warum sich ein System in einem zulässigen oder einem unzulässigen Zustand befindet.

Die wichtigste Aufgabe der Domänenexpert:innen liegt bei diesem Human-in-the-loop-Konzept eigentlich darin, verständliche Bezeichnungen für einzelne Klassen wie Maschinenelemente (z. B. Motorraum, Fließband) oder Werkzeuge (z. B. Imbusschlüssel, Schraube) vorzunehmen. Gleichzeitig wird der Domänenexperte bzw. die Domänenexpertin bei diesem Ansatz auch nicht zwingend gebraucht für die Bereitstellung von Erklärungen (oder die Bereitstellung der eigentlichen Condition-Monitoring-Funktionalität). Wenn ausreichend viele Daten in entsprechend verarbeitungsfähiger Form vorliegen, kann die Verarbeitung auch automatisiert erfolgen. Allerdings schränkt eine automatisierte Vergabe von Klassenbezeichnungen unter Umständen die Möglichkeit ein, für Menschen klar verständliche, natürlichsprachliche Erklärungen zu generieren. Alternativ könnten diese Klassenbezeichnungen aus zusätzlichen externen Informationsquellen bezogen werden, falls diese zur Verfügung stehen.

20 Projekt RAKI des BMWi-Technologieprogramms Smarte Datenwirtschaft (https://raki-projekt.de/)

USE CASE
KI-gestützte Maschinenzustandsüberwachung in der Kurzübersicht

TYP	Erkennung von Anomalie(n) zum Zweck der Instandhaltungsplanung Entscheidungsunterstützung)
KRITIKALITÄT	Hoch (funktionale Sicherheit, Wirtschaftlichkeit)
DATENTYPEN	Numerische Daten (Sensordaten, Betriebsparameter), Textdaten (Fehlercodes, Maschinen-Logdaten)
TYPISCHE KI-MODELLE	**ANSATZ 1:** Bayes-Netze **ANSATZ 2:** Maschinelles Lernen auf Basis von Wissensgraphen
(WICHTIGSTE) ZIELGRUPPEN und jeweilige übergeordnete Ziele für die Nutzung erklärbarer KI	**DOMÄNENEXPERT:INNEN** (Instandhaltungsteams): Kausalitätsbeziehungen finden, Konfidenz (Robustheit, Stabilität) bestimmen, Interaktionsmöglichkeiten verbessern
	ENTWICKLER:INNEN: Konfidenz (Robustheit, Stabilität) bestimmen; Interaktionsmöglichkeiten verbessern
KONKRETE ANFORDERUNGEN an Erklärbarkeit	**DOMÄNENEXPERT:INNEN** und **ENTWICKLER:INNEN:** Bewertung (Plausibilität, statistische Bewertung) der Güte der Modelle (betrifft Teilaspekte globaler Erklärbarkeit); Bewertung der Einzelentscheidung (lokale Erklärbarkeit)
GEEIGNETE ERKLÄRUNGSSTRATEGIEN	**ANSATZ 1:** • Verwendung und Anpassung von Surrogat-Modellen (Modellplausibilität), • Extraktion statistischer Güteparameter (Bayes'sche Statistik) **ANSATZ 2:** Natürlichsprachliche Erklärungen (Wissensgraphen)



5.2.2 Use Case KI-gestützte Prozessführung in der Prozessindustrie

Bereits in den 1980er Jahren wurden etwa in der chemischen Industrie Verfahren eingesetzt, die durchaus als lernende Systeme bezeichnet werden können und deren Vorteile teilweise erst heute in anderen Branchen erkannt werden. Als Beispiele sind hier etwa statistische Filterverfahren oder modellprädiktive Regler zu nennen. Gleichzeitig werden die Einsatzmöglichkeiten datengetriebener KI-Verfahren in der Prozessindustrie erst seit einigen Jahren tiefgehender untersucht. Es wird von Branchenexpert:innen erwartet, dass hier noch nicht gehobene Potenziale zur Steigerung der Wirtschaftlichkeit von Prozessen und Anlagen existieren, beispielsweise durch die Auswertung von Bild- und Videodaten oder eine verbesserte Mensch-Technik-Interaktion.

Gleichzeitig wird die Erklärbarkeit als Schlüssel dafür gesehen, datengetriebene KI-Systeme in sicherheitskritischen Anwendungsfeldern der Prozessindustrie überhaupt einsetzen zu können, da sowohl zulassungstechnische Gesichtspunkte als auch Akzeptanzaspekte davon abhängen.

Um dies zu veranschaulichen, wird nachfolgend ein Use Case einer KI-gestützten Prozessführung beschrieben. Datengetriebene KI-Methoden sollen dabei im Rahmen einer „Zustandserkennung" und für die Ermittlung „optimaler Betriebsverläufe" verwendet werden. Die Zustandsbestimmung und die nachgelagerte Regelung, die mit der Bestimmung von optimalen Betriebsverläufen adressiert wird, sind zwei miteinander eng verbundene Teilanwendungsfälle, die gleichermaßen zu einer KI-gestützten Prozessführung beitragen. Dabei stellt die Zustandsbestimmung generell eine Voraussetzung für die Ermittlung optimaler Betriebsverläufe dar. Nur durch die Verkettung entsprechender algorithmischer Prozesse und Ausführung in „Echtzeit" kann das Setzen tatsächlicher Stellgrößen und damit die Prozessführung erfolgen:

- *Zustandserkennung über Bilddaten:*
  In den komplexen dynamischen Systemen der Prozessindustrie ist eine wesentliche Teilaufgabe der Prozessführung, ausreichend Kenntnis über die Zustände des Systems zu erlangen. Viele solcher Zustände sind oft nicht direkt messbar, sondern im besten Fall lediglich „beobachtbar" (also bei Verfügbarkeit eines ausreichend guten Modells des Systems aus den messbaren Größen schätzbar). Aufwendige Probenanalysen, etwa, um die genaue anteilige Zusammensetzung von Materialflüssen zu ermitteln, sind teuer und liefern Ergebnisse mit zeitlichem Verzug sowie mit einer viel niedrigeren Wiederholrate als Sensoren. Eine günstige und in der Prozessindustrie bislang wenig genutzte Möglichkeit, zusätzliches Datenmaterial bei hoher Abtastrate zu erhalten, sind Bilddaten, mit denen z. B. gewisse ungewünschte Effekte wie Luftblasenbildung erkannt werden können. Um aus Bild- und Videodaten Informationen über schwer messbare Zustandsgrößen zu erlangen, die anschließend für die Prozessführung eingesetzt werden, können, zumindest teilweise, KI-basierte Modelle und Verfahren eingesetzt werden.

- *Optimale Betriebsverläufe:*
  Das Finden optimaler Trajektorien zur Führung von Anlagen in der Prozessindustrie durch An- und Abfahrvorgänge ist häufig eine sehr anspruchsvolle Aufgabe. Hierbei müssen neben typischerweise komplexen Systemdynamiken meist eine Vielzahl von Unsicherheiten, z. B. Messfehler oder Unsicherheit bezüglich geschätzter Größen, sowie potenziell kritische Nebenbedingungen wie Temperatur- oder Drucklimits explizit berücksichtigt werden. Eine besondere Herausforderung für die robuste modellbasierte Regelung sind unvorhergesehene, „diskrete" Ereignisse, etwa Schaltvorgänge.

Prinzipiell kann hier von einem autonomen System ausgegangen werden, das Analyse-, Planungs- und

Regelungsaufgaben übernimmt. Die Kritikalität der Anwendung ist als sehr hoch zu bewerten, da potenzielle Störfälle nicht nur enormen wirtschaftlichen Schaden, sondern aufgrund der Verarbeitung gefährlicher Stoffe auch gesundheitliche Gefährdungen von Anwohnern und der Umwelt zur Folge haben können.

Zielgruppen und übergeordnete Ziele für die Nutzung erklärbarer KI

Es existieren drei Zielgruppen, die bei dem Entwurf einer erklärbaren KI berücksichtigt werden müssen. Das sind in erster Linie die zulassenden Behörden, ohne deren Zustimmung ein KI-gestütztes System für die Prozessführung nicht eingesetzt werden kann. Als zweites sind hier die Domänenexpert:innen (Betriebspersonal) zu nennen, die betreiberseitig entsprechend ausgebildet sein müssen, um ihren Teil für den sicheren Betrieb einer Anlage beizutragen. Die dritte Gruppe sind die für die Prozessführung verantwortlichen Entwickler:innen. Diese müssen für den Use Case ebenfalls über ausreichend Domänenexpertise verfügen, um Systeme entwickeln zu können, die gemäß dem Stand der Technik und des Wissens garantieren, dass die relevanten Schutzziele (Gesundheit, Umwelt, Wirtschaftlichkeit usw.) beim Betrieb der Anlage nicht verletzt werden.

Das wichtigste übergeordnete Ziel für die Verwendung erklärbarer KI im Use Case ist, die Anfälligkeit des KI-Systems für Störungen aller Art speziell in der Betriebsphase permanent überprüfen zu können (Konfidenz prüfen), um ggf. sogar Notfallmaßnahmen einleiten zu können. Dabei ist essenziell, dass Domänenexpert:innen, unter Berücksichtigung des individuellen Prozessverständnisses, auch in die Lage versetzt werden, mögliche Probleme hinreichend umfassend (Kausalitätsbeziehungen finden) und gleichzeitig hinreichend schnell (Informationsgewinn erhöhen/Vereinfachung) begreifen zu können, damit sie auf dieser Basis, falls erforderlich, schnell und zielgerichtet Anpassungen vornehmen können (Interaktionen verbessern).

Erklärbarkeitsanforderungen aus Perspektive der Zielgruppe(n)

In dem Moment, in dem die KI-basierten Verfahren eine sicherheitsrelevante Funktion darstellen oder beeinträchtigen, müssen diese als Teil der Genehmigung für zulassende Behörden im Hinblick auf die Wirkmechanismen erklärbar, im Hinblick auf die Risikominimierung nachvollziehbar und mit Blick auf ihre Wirksamkeit prüfbar sein. Auch wenn die Zulassungsanforderungen für Schutz- und Notfallkonzepte keine KI-spezifischen Vorgaben machen[21], benötigen zulassende Behörden folglich sowohl Entscheidungs- als auch detaillierte und umfassende Modellerklärungen (lokale und globale Erklärbarkeit). Modelle und individuelle Entscheidungen müssen für eine Zulassung also potenziell überprüfbar sein – und somit erklärbar.

Das Betreiberunternehmen einer Anlage ist verantwortlich für den sicheren Betrieb. Es muss diesen sicheren Betrieb durch Ausbildung des Betriebspersonals und technische Maßnahmen ermöglichen. Da den Domänenexpert:innen (Betriebspersonal) somit eine zentrale Rolle in diesem Sicherheitskonzept zukommt, müssen ihnen zumindest Entscheidungserklärungen (lokale Erklärbarkeit) bereitgestellt werden, um potenziell sicherheitskritische Ereignisse zu erkennen und somit vermeintliche Störfälle abzuwenden.

Die Aufgabe der Entwickler:innen der Prozessführung ist es, Systeme zu entwickeln, die entsprechend relevante Schutzziele beim Betrieb der Anlage nicht verletzen. Sobald KI-basierte Verfahren Sicherheitsaspekte berühren, sind lokale und globale Erklärbarkeit zwingend notwendig, um sie nachhaltig als notwendiges funktionales Element zu etablieren.

Für diesen Anwendungsfall ergibt sich allerdings noch eine weitere, bedeutsame Anforderung. Da die Echtzeitanforderung als eine allgemeine Anforderung eines automatisierungsbezogenen Use Case verstanden werden muss, überträgt sich diese selbstverständlich auch auf Erklärungen, die sowohl von den entsprechenden Stakeholdern als auch für die algorithmische Verarbeitung „in time“ zur Verfügung gestellt werden müssen.

Erklärungsstrategien

Wie konkrete Erklärungsstrategien ausgestaltet werden können, wird für diesen Anwendungsfall derzeit noch untersucht. Es zeigen sich allerdings bereits mehrere vielversprechende Ansätze.

21 Zulassungsanforderungen sind heute nicht KI-spezifisch. Das Bundesemissionsschutzgesetz, das Wasserhaushaltsgesetz oder die Störfallverordnung machen keinen Unterschied, ob ein selbstlernendes System zum Einsatz kommt oder nicht. Wenn jedoch ein KI-System direkt oder indirekt Schutzziele beeinflusst, ergeben sich Anforderungen auf lokaler und globaler Ebene.

Um den Basisanforderung der Domänenexpert:innen für Entscheidungserklärungen (lokale Erklärbarkeit) gerecht zu werden, werden bei der Zustandserkennung über Bilddaten unterschiedliche Ansätze als grundsätzlich geeignet erachtet. Dafür werden Post hoc-Erklärungswerkzeuge wie LIME, CAM und Guided Backpropagation bislang präferiert und entsprechend auf ihre Tauglichkeit für den Use Case analysiert. Mit diesen Tools können für Menschen mit entsprechendem Vorwissen interpretierbare Erklärungen auf der Basis von Bilddaten generiert und Biases vergleichbar gut erkannt werden. Diese Werkzeuge werden ebenfalls als hinreichend ausgereift angesehen, um auch die Anforderungen der Domänenexpert:innen für eine Überwachung der Zustandserkennung aus Bildern zu erfüllen.

Die Anforderung der sehr detaillierten Nachvollziehbarkeit der Entscheidungsprozesse, d. h., Modell- und Entscheidungserklärungen (globale und lokale Erklärbarkeit) bereitstellen zu müssen, betrifft bei diesem Use Case vor allem den zweiten Teil der Problemstellung: die Ermittlung optimaler Betriebsverläufe. Dabei stellt sich diese Anforderung, wie oben erläutert, aus Sicht der Entwickler:innen und der Zulassungsbehörden. Es deutet nach Meinung von Branchenexpert:innen alles darauf hin, dass die Entwicklung eines KI-Systems, das ausschließlich auf Black-Box-Modellen beruht, keine realistischen Chancen auf eine Zulassung habe. Die Zulassung jedes Prozessführungssystems erfordert die Berücksichtigung eines umfassenden, zulassungsfähigen Schutzkonzeptes, wofür die detaillierte Nachvollziehbarkeit der algorithmischen Systeme als unumgänglich erachtet wird (siehe hierzu auch den folgenden Abschnitt 5.2.3).[22]

Die Bereitstellung klassischer Post hoc-Erklärungen, die Analysewerkzeuge wie z. B. LIME bieten, wird hier folglich für die Ermittlung optimaler Betriebsverläufe von Branchenexpert:innen als nicht ausreichend angesehen.

Im Rahmen des Forschungsprojektes[23], das diesen Anwendungsfall konkret adressiert, wird eine Umsetzung über die Entwicklung passender, hybrider Methoden angestrebt. Vereinfacht betrachtet, könnten innerhalb einer entsprechenden hybriden KI die White-Box-Komponenten die Erfüllung der sicherheitstechnischen Anforderungen garantieren, während die Black-Box-Komponenten ausreichend geprüfte Informationen „zuliefern".

Dabei wäre es beispielsweise ein zunächst naheliegender Ansatz zu fordern, auf Basis von Black-Box-Modellen nur wirtschaftlich bedeutsame, nicht aber sicherheitsrelevante Stellgrößen zu ermitteln bzw. zu berechnen. Da allerdings in sicherheitskritischen Systemen alle mit dem Gesamtsystem interagierenden Stellglieder als potenziell kritisch betrachtet werden müssen, ist es grundsätzlich notwendig, alle Auswirkungen errechneter Stellgrößen permanent systemintern auf Sicherheitsrisiken zu prüfen.

Unabhängig von der Designphase kann dies teilweise sogar in Mensch-Maschine-Interaktionsprozessen mit Entwickler:innen bzw. Domänenexpert:innen geschehen. Ein Rückgriff auf menschliche Entscheider:innen ist dabei in vielen Fällen durchaus praktikabel. Nicht selten können potenziell kritische, jedoch wirtschaftlich vielversprechende Modellanpassungen – bezogen auf das Neutrainieren von KI-Modellen zur Zustandserkennung, aber auch in Bezug auf das Anpassen von Anlagenmodellen aufgrund ermittelter Zustände – erst einmal „zurückgehalten" und durch Menschen und/oder Simulation überprüft werden. Solche Modellanpassungen können den Betrieb eines Prozessleitsystems zwar aus wirtschaftlicher Sicht positiv beeinflussen, indem sie etwa den Produktionsdurchsatz steigern; hinsichtlich Risikoerwägungen können sie auch potenziell negativ beeinflussen, z. B. aufgrund eines unerkannten Datenbias, der die Zustandsbestimmung verfälscht und somit für die Prozessregelung ein unkalkulierbares Risiko darstellen kann. In bestimmten Szenarien ist es denkbar, entsprechende Modellanpassungen erst mit einer gewissen Verzögerung umzusetzen („deploy"). Zwischenzeitlich könnte auf nicht aktualisierte Modelle, vorbehaltlich der Wahrung der Schutzziele, und auf traditionelle Verfahren vertraut werden, z. B. Zustandsschätzung mit probabilistischen Filterverfahren (z. B. Kalman-Typ-Filterverfahren).

22 Bei haftungsrelevanten Unfällen oder Störfällen würde staatsanwaltschaftlich geprüft werden, ob der Stand der Technik zur Vermeidung von Schutzverletzungen berücksichtigt wurde, der sich in diesem Anwendungsfeld an traditioneller und interpretierbarer Mess- und Regeltechnik sowie White-Box-Modellen orientiert. Gegebenenfalls drohen bei der Feststellung eines Versäumnisses in Bezug auf das Schutzkonzept erhebliche wirtschaftliche Konsequenzen für die verantwortlichen Unternehmen. Im Falle eines Unfalls oder Störfalls, der sich trotz Wahrung eines zulassungskonformen Betriebs und trotz Einhaltung des zulassungsrelevanten Schutzkonzepts ereignet, stünden die zulassenden Behörden in der Verantwortung.

23 http://keen-plattform.de/

Angesichts des Autonomiegrads ist es indes nötig, dass Domänenexpert:innen vom System benachrichtigt werden, wenn eine entsprechende Entscheidung getroffen werden muss. Denkbar ist ebenfalls, dass auf Basis vergangener Entscheidungen ein Vorgehen vorgeschlagen wird.

In Bezug auf die Bestimmung optimaler Betriebsverläufe wird hier der Ansatz verfolgt, KI-basierte Methoden einzusetzen, um aus Anlagen- und Simulationsdaten geeignete „hybride" Modelle zu erzeugen. Dabei ist mit dem Konzept eines „hybriden" Modells gemeint, dass White-Box- mit Black-Box-Modellkomponenten kombiniert werden, um sowohl von der Deterministik mechanistischer Modelle, als auch von der Mustererkennung datengetriebener Ansätze zu profitieren. Aus den resultierenden Modellen können mittels moderner Regelungsverfahren – hier vor allem modellprädiktive Regelung[24] – optimale Steuertrajektorien abgeleitet werden. Diese hybriden Modelle müssen im Realbetrieb („online") angepasst und nachoptimiert und mit Blick auf Gültigkeit und Leistungsfähigkeit bei Störungen in der Anlage überwacht werden. Hier spielt die Erklärbarkeit – sowohl der Empfehlungen als auch der Anpassungen durch Re-Optimierung – eine große Rolle für das Vertrauen in das resultierende Empfehlungssystem.

Zuletzt wird von Branchenexpert:innen erwartet, dass ein solches hybrides KI-System auch die Anforderungen erfüllt, geeignete Erklärungen „in time" bzw. rechtzeitig vorlegen zu können. Dies liegt daran, dass auch die Kernfunktionalität des Gesamtsystems die Echtzeit-Anforderungen erfüllen muss und zugehörige Erklärungen zunächst automatisiert, z. B. mittels Simulation, aber bei Bedarf auch von Personen überprüfbar sein müssen. Eine systembasierte Unterstützung der zuständigen Person ist hier perspektivisch unumgänglich, obwohl die Zielgruppe der Domänenexpert:innen mit Prozessleitsystemen gut vertraut bzw. vertraut zu machen ist. Die Erklärungen müssen dabei an der Expertise der Zielpersonen und an den zeitlichen Vorgaben der Anwender:innen bzw. des Gesamtprozesses ausgerichtet werden. Durch eine vorläufige Verwendung der Black-Box-Komponenten für etwas weniger sicherheits- und zeitkritische Aufgaben wird angestrebt, mit den nicht unerheblichen Anforderungen an menschliche „Entscheider" und an eine effiziente Mensch-Maschine-Interaktion umgehen zu können. Infolgedessen können menschliche Entscheider:innen in den für sie möglichen Reaktionszeiten auf Ereignisse reagieren und wesentlich dazu beitragen, dass eine KI-gestützte Prozessführung sich sicherheitstechnisch angemessen verhält und möglicherweise sogar schrittweise dazulernt.

24 Modellbasierte Variante eines Reinforcement-Learning-Verfahrens, das eine Regelstrategie bzw. „Control Policy" generiert

USE CASE
KI-gestützte Prozessführung in der Prozessindustrie in der Kurzübersicht

TYP	*Es existieren zwei Teilaufgaben* (1) KI-gestützte Analyse (Zustandserkennung über Bilddaten), (2) KI-gestützte Feedback-Steuerung (Optimale Betriebsverläufe)
KRITIKALITÄT	Sehr hoch (Potenzial von Unfällen/Störfällen)
DATENTYPEN	Numerische Daten (Sensordaten, Betriebsparameter), Bilddaten
TYPISCHE KI-MODELLE	Für (1): Neuronale Netze (Erklärbarkeitsdefizite bzw. Black-Box) Für (2): Reinforcement Learning (modellprädiktive Regelung) auf Basis hybrider Modelle (zumindest weitgehend erklärbar)
(WICHTIGSTE) ZIELGRUPPEN und jeweilige übergeordnete Ziele für die Nutzung erklärbarer KI	**DOMÄNENEXPERT:INNEN** (Betriebspersonal) und **ENTWICKLER:INNEN** (Prozessführung): Konfidenz (Robustheit, Stabilität, Vulnerabilität) bestimmen, Kausalitätsbeziehungen finden, Informationsgewinn erhöhen (Vereinfachung), Interaktionsmöglichkeiten verbessern (insb. für Domänenexpert:innen)
	ZULASSENDE BEHÖRDEN: „Nachvollziehbarkeit" und Schutzkonzept prüfen
KONKRETE ANFORDERUNGEN an Erklärbarkeit	**DOMÄNENEXPERT:INNEN** (Betriebspersonal): Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit)
	ZULASSENDE BEHÖRDEN und **ENTWICKLER:INNEN** (Prozessführung): Einzelentscheidungserklärungen und Modellerklärungen (lokale und globale Erklärbarkeit)
GEEIGNETE ERKLÄRUNGSSTRATEGIEN	für (1): Post hoc-Erklärungen, z. B. LIME für (2): Einbindung der Black-Box-Modellkomponenten mittels hybrider Modellierung

5.2.3 Regulatorik und Zertifizierung in der Produktionswirtschaft

In der Produktionswirtschaft ist zu beobachten, dass zunehmend KI-Anwendungen in Robotik-Systeme oder Produktionsanlagen integriert werden. Dabei halten vor allem Computer-Vision- oder KI-gestützte Datenanalyse-Ansätze Einzug in die Fertigungsindustrie, um die zuvor eher regelbasiert arbeitenden algorithmischen Komponenten zu ergänzen.

Für die CE-Zertifizierung solcher Produktionsmaschinen und -anlagen stellt die Maschinenrichtlinie 2006/42/EG[25] vom 17. Mai 2006 die zentrale Norm der Europäischen Union (EU) dar. Sie regelt die Einhaltung der Prinzipien für die Sicherheit technischer Systeme und gibt somit den zulassungsseitigen Rahmen vor – und zwar zunächst unabhängig davon, ob KI-Elemente Einfluss auf das Maschinenverhalten nehmen oder nicht. Die Maschinenrichtlinie bietet somit einen Rahmen für detaillierte technische Regeln. So finden sich in ihrem Anhang I die allgemeinen Sicherheits- und Gesundheitsschutzvorgaben, die bei der Risikobeurteilung und -minderung zu beachten sind. Das zentrale Ziel ist die funktionsgerechte Auslegung der Maschine für den festzulegenden Anwendungsbereich und für die gesamte Lebensdauer, sodass Personen nicht gefährdet werden. Das umfasst etwa Vorgaben zu Handhabung, Steuerung und Instandhaltung der jeweiligen Anlage. Ein besonderes Augenmerk liegt auf Schutzmaßnahmen vor mechanischer Gefährdung. Zudem schreibt die Richtlinie vor, welche Informationsmaterialien zur Maschine und zu zugehörigen Schutzmaßnahmen in welcher Form vorliegen müssen.

Bisher zeigte sich die Maschinenrichtlinie robust gegenüber technologischen Veränderungen, da die grundlegenden Sicherheits- und Nutzungsprinzipien mechatronischer Systeme mit definierten, deterministischen Steuerungs- und Regelungskonzepten Bestand hatten. Eine systematische Evaluation der Richtlinie im Jahr 2018 seitens der europäischen Kommission mit Befragung zahlreicher Stakeholder aus dem Maschinen- und Anlagenbau ergab allerdings, dass mit dem verstärkten IoT[26] - und KI-Einsatz in Maschinenbauprodukten perspektivisch Anpassungsbedarf besteht, da dadurch ein Paradigmenwechsel zu vernetzten, eigenständig entscheidenden bis hin zu lernenden Maschinenbauprodukten erwartet wird (Europäische Kommission 2018). Im Falle der Verwendung von Black-Box-Modellen kann das Verhalten der Systeme nicht ausreichend sicher vorhersagbar sein. Durch die Einbettung in technische Systeme werden Entscheidungen in physische Handlungen überführt, selbst wenn nur eine Teilfunktionalität durch KI ermöglicht wird. So kann ein KI-Modul zur Bilderkennung dazu beitragen, einen Roboterarm zu steuern. Daraus können potenziell nicht kalkulierbare Sicherheitsrisiken resultieren, z. B. durch den Kontakt mit Menschen, durch die verarbeiteten Stoffe oder die Verknüpfung zur Infrastruktur.

Wenn sich KI-Verfahren im Produktionsumfeld hingegen während des Betriebs weiterentwickeln („Training-on-the-job"), ist es unter Umständen enorm schwierig bis unmöglich, Entscheidungen in ausreichendem Maße nachzuvollziehen. Nur mithilfe dedizierter Erklärungsstrategien oder werkzeuge ist eine Zertifizierung nach den Vorgaben der aktuellen Maschinenrichtlinie derzeit vorstellbar, wenn von einem KI-Produkt eine potenzielle Gefährdung für Personen ausgehen kann und keine alternativen Sicherheitsvorkehrungen getroffen werden. Es ist unklar, wie mit einem neu gelernten KI-System in sicherheitsrelevanten Anwendungen umgegangen werden kann.

Auch bereits vortrainierte KI-Black-Box-Systeme können ein Zertifizierungshindernis darstellen, insbesondere, wenn sie physische Handlungen oder andere sicherheitsrelevante Funktionen des Systems beeinflussen. Wenn diese nicht umfassend kontrollier- und nachvollziehbar sind, bleibt ein Risiko bestehen – oder der Mehrwert eines KI-Systems im Vergleich zur deterministischen Umsetzung ist gering. Im Kontakt mit Menschen werden von KI-gestützten Systemen also neue Sicherheits- und Gesundheitsrisiken erwartet, welche die bisherigen Zertifizierungsanforderungen der Richtlinie nicht abdecken.

Noch ist offen, ob eine Anpassung der Maschinenrichtlinie selbst erforderlich ist, sei es durch den Einbezug

25 Richtlinie 2006/42/EG des Europäischen Parlaments und des Rates vom 17. Mai 2006 über Maschinen, online verfügbar unter: https://eur-lex.europa.eu/eli/dir/2006/42/oj

26 IoT = Internet of Things (Internet der Dinge)

ethischer Regeln für autonome Systeme („Robotergesetze“) oder durch die Erweiterung der Sicherheits- und Gesundheitsregularien. Möglicherweise reichen auch technische Standards aus. Grundsätzlich gilt jedoch, dass Maschinen „in keinen unkontrollierten Zustand gelangen, der eine Gefahr für den Bediener oder für unbeteiligte Dritte darstellen würde“. Neben allgemeinen Sicherheitsvorrichtungen und -maßnahmen können ebenso erklärbare KI-Algorithmen bzw. entsprechende Warnsysteme wesentlich dazu beitragen, eine menschliche Aufsicht zu ermöglichen und die genannten Risiken maßgeblich reduzieren. Somit ist gut vorstellbar, dass künftige Richtlinien für die Zulassung lernender Systeme die Verwendung erklärbarer KI-Modelle und -Verfahren als einen Lösungsansatz aufgreifen und auch konkrete Anforderungen benennen werden, etwa bezogen auf die Vermittlung der Erklärungen an die Bediener:in, wenn von betreffenden Systemen eine potenzielle Gefahr ausgehen kann.

In der Prozessindustrie, in der häufig gefährliche Stoffe unter hohen Drücken verarbeitet werden, sind über die Maschinenrichtlinie hinaus zusätzliche Regulierungsvorgaben zu beachten (VERBAND DER CHEMISCHEN INDUSTRIE e.V. 2012). Die geltende Störfallverordnung[27] (SEVESO-III-Richtlinie) des Bundes-Immissionsschutzgesetzes schreibt folglich hohe Standards für die Zertifizierung der Prozesssicherheit vor. So müssen Produktionskonzepte grundsätzlich nachvollziehbar sein. Darüber hinaus muss kontrollierbar sein, dass die tatsächliche Implementierung dem geplanten Konzept entspricht. Für den Einsatz von KI in diesem Rahmen muss grundsätzlich sichergestellt sein, dass die Entscheidungen des Gesamtsystems ausreichend transparent, wiederholbar, detailliert nachvollziehbar und bei Bedarf korrigierbar sind.

Das Deutsche Institut für Normung (DIN) verfolgt seit 2020 eine dezidierte KI-Roadmap für die Normung (Wahlster und Winterhalter 2020). Ein Schwerpunktthema bildet die industrielle Automation. Neben der Softwarenormierung für industrielle Anwendungen werden auch die Bedarfe für lernende technische Systeme betrachtet. Zur ermittelten Herausforderung „Erklärbarkeit und Validierung“ hat der VDE bereits die technischen Regel E VDE-AR-E 2842-61-1:2020-07[28] veröffentlicht. In ihr sind die Terminologie und grundlegende Konzepte zu erklärbarer KI beschrieben. Aufbauend sollen in dem nationalen Umsetzungsprogramm „Trusted AI“ Qualitätskriterien und reproduzierbare, standardisierte Prüfverfahren für verlässliche KI-Systeme erarbeitet werden. Noch lässt sich nicht prognostizieren, wann sie für technische Systeme Anwendung finden werden (Wahlster und Winterhalter 2020).

Im Ergebnis gelten für KI-gestützte Systeme derzeit dieselben Sicherheitsbestimmungen wie für konventionell gesteuerte Produkte, auch wenn es erste Normierungsbestrebungen gibt. Damit bestehen für den Hersteller Haftungsregelungen, die der Produkthaftungsrichtlinie[29] entsprechen. Insbesondere gibt es keine anerkannten Prozesse für die Zertifizierung von KI-unterstützten Systemen. Das gilt speziell für Systeme, die ihr Verhalten während des Betriebs signifikant verändern, ohne dass diese Änderung einer menschlichen Aufsicht unterliegt, sodass die Regularien aktuell lernende Systeme dieser Art ausschließen. Auch vortrainierte KI-Systeme, die sich nicht weiterentwickeln, dürfen nur unter kontrollierbaren Rahmenbedingungen eingesetzt werden. Erklärbare KI-Algorithmen sind somit eine Grundlage dafür, Lern- und Entscheidungsprozesse von KI-Systemen nachvollziehbar und damit kontrollierbar umzusetzen und das Spektrum zertifizierter Anwendungen bei technischen Systemen signifikant zu erweitern.

27 Zwölfte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes, online verfügbar unter: http://www.gesetze-im-internet.de/bimschv_12_2000/index.html

28 VDE-Anwendungsregel E VDE-AR-E 2842-61-1:2020-07, Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen – Teil 61-1: Terminologie und Grundkonzepte, https://www.vde-verlag.de/normen/1800574/e-vde-ar-e-2842-61-1-anwendungsregel-2020-07.html

29 Gesetz über die Haftung für fehlerhafte Produkte/Produkthaftungsrichtlinie, online verfügbar unter: http://www.gesetze-im-internet.de/prodhaftg/index.html

5.3 Gesamtbetrachtung der Use Cases

Beim Vergleich der vier Use Cases fällt zunächst auf, dass zwei Motivationsgründe für die Nutzung erklärbarer KI in allen vier Fällen als wichtig erachtet werden, nämlich Kausalitätsbeziehungen zu identifizieren und Konfidenz zu bestimmen. Gleichzeitig gibt es indes sehr individuelle Motivationen für die Nutzung erklärbarer KI: Z. B. den eigentlichen Informationsgewinn durch erklärbare KI erhöhen oder die Interaktion zwischen Mensch und KI-System verbessern:

- Die Motivation, Kausalitätsbeziehungen zu identifizieren, steht etwa beim Use Case zur Bildanalyse histologischer Gewebeschnitte für die Domänenexpert:innen (hier Pathologinnen und Pathologen) eindeutig an erster Stelle. Das medizinische Personal will in diesem Fall mithilfe erklärbarer KI möglichst auf einen Blick erkennen, warum eine konkrete Klassifikationsentscheidung in Bezug auf z. B. einen erkannten Tumor getroffen wurde (lokale Erklärbarkeit). Die Schlussfolgerungen, die zu der Entscheidung führten, können dann im besten Fall von den Verantwortlichen dank ihres medizinischen Expert:innenwissens nachvollzogen werden und anschließend entweder in die medizinische Diagnose einfließen oder verworfen werden. Auch in Bezug auf Zulassungen gehen Fachleute davon aus, dass zumindest individuelle Entscheidungen zu einzelnen Patient:innen für Ärztinnen und Ärzte nachvollziehbar sein müssen[30]. Eine Lösung mit Post hoc-Erklärungswerkzeugen wie LRP oder LIME, dank denen Bildbereiche für die Domänenexpert:innen visuell hervorgehoben werden, wird hier als Erklärungsstrategie verfolgt. Diese Post hoc-Erklärungswerkzeuge werden ebenfalls von KI-Entwickler:innen eingesetzt, deren vorrangiges Ziel das Testen von Fairness (Aufdecken von Datenbias) oder das Bestimmen von Konfidenz ist.

- Das übergeordnete Ziel, Kausalitätsbeziehungen zu identifizieren, ist auch beim Use Case zur Maschinenzustandsüberwachung die Kernmotivation: Domänenexpert:innen (hier Ingenieur:innen) wollen der Entdeckung möglicher Anomalien durch ein KI-System zunächst nachgehen, bevor sie etwaige Instandhaltungsmaßnahmen einleiten. Eine absolute Mindestanforderung ist daher in diesem Fall, Erklärungen zu Einzelentscheidungen bereitzustellen (lokale Erklärbarkeit). Zulassungsseitig bestehen hier in der Regel keine wesentlichen Hürden. Allerdings müssen im Zweifel Entscheidungen von großer sicherheitstechnischer und wirtschaftlicher Tragweite und unter Zeitdruck durch Expert:innen für Instandhaltung gefällt werden, sodass die Erklärbarkeit von Modellwirkmechanismen (globale Erklärbarkeit) für eine Vorabprüfung der Modellverlässlichkeit meist ausschlaggebend dafür ist, ob ein KI-System überhaupt in den operativen Einsatz kommt. Als Erklärungsstrategie werden zwei unterschiedliche Wege verfolgt. Einerseits werden Bayes-Netze und ein Surrogat-Modell verwendet, um für Anwendende intrinsische Aussagen zu Eintrittswahrscheinlichkeiten von Ereignissen und ein flexibel simulierbares Anschauungsmodell bereitstellen zu können. Ein zweiter Ansatz gründet auf einer Machine-Learning-Methodik auf Basis von Wissensgraphen. Dabei werden natürlichsprachliche Erklärungen bereitgestellt, die der Nutzer selbst auf seine Anforderungen abstimmen kann. Die Verbesserung der Interaktion zwischen Mensch und KI-System wird gerade durch diesen Ansatz adressiert, was Anwender:innen zusätzlich motivieren kann, erklärbare KI zu nutzen.

- Das übergeordnete Ziel, den Informationsgewinn der Domänenexpert:innen zu erhöhen, ist die zentrale Motivation beim Use Case der medizinischen Textanalyse von Arztbriefen. Dabei sind die konkreten, fallbezogenen Anhaltspunkte, warum eine besonders große Nähe zwischen Krankheitsverläufen bei Patient:innen überhaupt gefunden wurde, für das medizinische Personal als Schlüsselinformation unerlässlich. Denn nur so kann von medizinischen Expert:innen effizient bewertet werden, ob ein Kriterium, das für die Klassifikation der KI ausschlaggebend war, entweder plausibel oder medizinisch nicht sinnvoll ist. Dafür muss ein System, das bei dieser Entscheidung unterstützen soll, Einzelfallentscheidungen inhaltlich begründen können (lokale Erklärbarkeit). Grundlage zur Bereitstellung von Erklärungen sind

30 Es existieren erste zugelassene Produkte für die KI-gestützte radiologische Bildanalyse am Markt, was einen Hinweis gibt, dass die Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit) zumindest in Einzelfällen bereits ausreichend für eine Zulassung war.

nominelle Black-Box-Modelle (neuronale Netze), die durch Prototypen bzw. externe Wissenssammlungen ergänzt werden; folglich kann das resultierende Modell selbst medizinisch nachvollziehbare Gründe für einzelne Entscheidungen geben, indem relevante Textstellen in Arztbriefen oder externen Publikationen für die Zielgruppe visuell hervorgehoben werden.

- Im Use Case der KI-gestützten Prozessführung gibt es unterschiedliche Motivationen für die Nutzung erklärbarer KI. Entscheidend ist vor allem das übergeordnete Ziel, entsprechende Konfidenzen zu bestimmen, vor allem bezüglich der Auswirkungen von Einzelentscheidungen für das komplexe Gesamtsystem. Unentdeckte Fehler in der visuellen Zustandserkennung bzw. Anfälligkeit für Störungen und Bias in den „hybriden" Modellen können im Zweifel unkalkulierbare Risiken für die robuste und stabile Steuerung sowie Regelung der chemischen Anlagen bedeuten. Daher offenbaren sich bei diesem Use Case im Vergleich auch die weitreichendsten Erklärbarkeitsanforderungen, sodass hier über die Erklärbarkeit von Einzelentscheidungen (lokale Erklärbarkeit) hinaus ebenfalls die detaillierte Erklärbarkeit von Modellwirkmechanismen (globale Erklärbarkeit) erforderlich ist. Hier wird der Ansatz verfolgt, aus mechanistischen Modellen und Simulationsdaten sowie Bild- und Sensordaten geeignete „hybride" Modelle zu erstellen, die White-Box- mit Black-Box-Komponenten zu selbsterklärenden Anlagenmodellen kombinieren. Moderne regelungstechnische Ansätze können dann die durch maschinelles Lernen verbesserten Anlagenmodelle für die „On-the-job"-Generierung von zeit- oder energieoptimalen Betriebsverläufen nutzen.

Die Bereitstellung von Modellerklärungen (globale Erklärbarkeit) ist von den betrachteten Use Cases nur für die KI-gestützte Prozessführung eine strikte Zulassungsanforderung. Es stehen hier zwar fachlich versierte Personen bereit, die ein entsprechendes KI-System beaufsichtigen, diese können aber unmöglich jede Einzelaktion des Gesamtsystems bzw. der Prozessführung überprüfen. Stattdessen müssen verantwortliche Personen aktiv darauf hingewiesen werden, wenn Entscheidungen von ihnen getroffen werden sollen. Insbesondere wenn sicherheitsrelevante Anwenderentscheidungen des Fachpersonals nicht getroffen werden, muss das System selbstständig alternative Maßnahmen entsprechend eines festzulegenden Schutzkonzeptes einleiten.

Anforderungen an die Form und den Umfang von Erklärungen sowie an die Zeitspanne, die eine Generierung von Erklärungen in Anspruch nehmen darf, sind höchst anwendungsspezifisch. Sie müssen an der Expertise der Zielpersonen und an den zeitlichen Vorgaben der Anwender:innen bzw. des Gesamtprozesses ausgerichtet werden. Dies erfordert Erklärungen, die gleichzeitig oder zumindest kurz nach der eigentlichen Entscheidung bzw. Empfehlung des KI-Systems vorliegen müssen und zusätzlich den – teils miteinander in Konflikt stehenden – Ansprüchen an Erklärungen genügen müssen: einfach, kurz und umfassend.

Anforderungen an die Form und den Umfang von Erklärungen sowie an die Zeitspanne, die eine Generierung von Erklärungen in Anspruch nehmen darf, sind höchst anwendungsspezifisch. Sie müssen an der Expertise der Zielpersonen und an den zeitlichen Vorgaben der Anwender:innen bzw. des Gesamtprozesses ausgerichtet werden.

Weitere Details und Literaturhinweise zu den verwendeten Ansätzen finden sich in Kapitel 3 und im Glossar in Anhang A. Ausgenommen hiervon sind der Machine-Learning-Ansatz auf Basis von Wissensgraphen und der Ansatz hybrider Modellierung, die sich beide noch zu sehr im Forschungsstadium befinden, um hier detaillierter dargelegt zu werden.

6 PRAKTISCHE ERSTE SCHRITTE: ORIENTIERUNGS-HILFE ZUR AUSWAHL VON ERKLÄRUNGS-STRATEGIEN

6 PRAKTISCHE ERSTE SCHRITTE: ORIENTIERUNGSHILFE ZUR AUSWAHL VON ERKLÄRUNGSSTRATEGIEN

Im Folgenden werden Empfehlungen zur Auswahl von Erklärungswerkzeugen vorgestellt, die aus den Experteninterviews, eigener Literaturrecherche, den Ergebnissen der Benchmark-Tests aus der vom KI-Fortschrittszentrum „Lernende Systeme und Kognitive Robotik" des Fraunhofer IPA durchgeführten Studie (Schaaf et al. 2021) sowie vom Bosch Center for Artificial Intelligence bereitgestellten Informationen abgeleitet wurden. Eine Übersicht über die Werkzeuge und die zugrundeliegenden Vor- und Nachteile findet sich in Kapitel 3.

Vor der Auswahl eines geeigneten Erklärungswerkzeugs müssen die Designkriterien für das Zielsystem betrachtet werden. Diese umfassen insbesondere die Zielgruppen der Erklärung, die Typen der zugrundeliegenden Daten und das ausgewählte KI-Modell, das die Entscheidungen trifft.

Im Hinblick auf die Auswahl des KI-Modells sollte (generell) stets geprüft werden, ob es möglich ist, ein weniger komplexes, damit verständlicheres und leichter nachvollziehbares Modell für die Lösung des Ausgangsproblems zu nutzen, das den Anforderungen genügt. Hinsichtlich der Erklärbarkeit handelt es sich im Idealfall um ein White-Box-Modell. Post hoc-Erklärungen für Entscheidungen von Black-Box-Modellen können als problematisch angesehen werden, da sie die Funktionsweise des Modells zu vereinfachen suchen, diese aber nicht in ihrer Vollständigkeit darstellen können. Daraus folgt, dass diese Erklärungen nicht immer vollständig akkurat sind und es sich vielmehr um Approximationen handelt, die auch zu Fehlern in der Interpretation führen können (Rudin 2019).

Im Hinblick auf die Auswahl des KI-Modells sollte (generell) stets geprüft werden, ob es möglich ist, ein weniger komplexes, damit verständlicheres und leichter nachvollziehbares Modell für die Lösung des Ausgangsproblems zu nutzen, das den Anforderungen genügt.

White-Box-Modelle und weitere KI-Modelle, die selbst sowohl Entscheidung als auch Erklärung liefern, bieten den Vorteil, dass kein zusätzliches Analysewerkzeug bzw. Surrogat-Modell für die Erklärung der Entscheidungen benötigt wird. Vielmehr ist das ursprüngliche Modell selbst nachvollziehbar(er) und kann überdies konkrete Entscheidungserklärungen liefern. Eine Auseinandersetzung mit der Funktionsweise des zusätzlichen Analysewerkzeugs, das für die Erklärung zuständig ist, ist nicht erforderlich. Außerdem liefern White-Box-Modelle zusätzlich die Möglichkeit, ein tieferes Verständnis des KI-Algorithmus selbst zu erreichen.

Die meisten der in Kapitel 3 besprochenen Post hoc-Erklärungswerkzeuge wurden von den Expert:innen für die Bereitstellung von Einzelentscheidungen für KI-Anwender:innen, z. B. Domänenexpert:innen, als nur bedingt geeignet eingeschätzt. Die Methoden haben oft den Nachteil, dass deren Funktionsweise für Nutzer:innen nicht intuitiv verständlich ist, sodass die korrekte Interpretation der Ergebnisse nicht automatisch gesichert ist. Allgemein wurde in den für die Studie geführten Interviews von mehreren Expert:innen geäußert, dass auch für Nutzende ohne KI-Expertise intuitive Erklärungsstrategien bereitgestellt werden müssen, die insbesondere inhaltliche Begründungen geben können. Ein gutes Beispiel dafür sind Counterfactual Explanations. Ebenso erfüllen Surrogat-Modelle, trotz der Diskrepanz zwischen Ausgangs- und Surrogat-Modell, gerade die Anforderungen einer möglichst intuitiven Nachvollziehbarkeit vergleichsweise gut, da z. B. an den dafür häufig genutzten Entscheidungsbäumen die Entscheidungskriterien unmittelbar ablesbar sind. Die Nutzung von Prototypen ist diesbezüglich ebenfalls empfehlenswert, um inhaltliche Argumente

für Entscheidungen zu liefern, da dieser Ansatz eine vergleichbar intuitive Form der Plausibilisierung von Entscheidungen darstellt. Entsprechende Strategien werden nicht nur eingesetzt, um Entscheidungen eines KI-Systems für Domänenexpert:innen nachvollziehbar zu gestalten – auch KI-Entwickler:innen können vom Einsatz profitieren und so mithilfe neuer Erkenntnisse zur Entscheidungsfindung des Modells die Qualität der Ergebnisse der eigenen Entwicklung verbessern.

Für KI-Entwickler:innen ist sowohl die Generierung von Entscheidungserklärungen als auch von Modellerklärungen wichtig. Einige Expert:innen sehen die besprochenen Post hoc-Methoden – mit Ausnahme der Counterfactual Explanations – als hauptsächlich geeignet, um KI-Entwickler:innen bei der Verbesserung der Algorithmen zu unterstützen. Beim Einsatz von Black-Box-KI sind derzeit die Methoden Integrated Gradients (für neuronale Netzwerke) bzw. SHAP (modellagnostisch) besonders gut geeignet. Ist eine schnelle Approximation ausreichend, so kann anstelle von Integrated Gradients auch auf DeepLIFT zurückgegriffen werden. Für die Anwendung auf vielschichtigen Modellen mit hoher Parameterzahl bzw. zur Verarbeitung hochdimensionaler Daten muss für SHAP überprüft werden, ob die Laufzeit weiterhin praxistauglich ist. Allgemein ist zu beachten, dass die Nutzbarkeit der Methoden stark vom jeweiligen Anwendungsfall abhängt und einzeln geprüft werden sollte, inwiefern der jeweilige Zweck gut erfüllt wird. Bewertungen einzelner Methoden sind zudem oft subjektiv. Bei Auswahl einer konkreten Methode sollte deren Funktionsweise gut bekannt sein, vor allem deren Nachteile. Als weitere Empfehlung wurde angemerkt, dass auch KI-Entwickler:innen sich nicht auf eine Methode verlassen sollten. Der Test mehrerer Methoden ist angeraten, um womöglich nicht direkt ersichtliche Probleme einer Methode für den spezifischen Anwendungsfall frühzeitig erkennen und umgehen zu können.

In der vom KI-Fortschrittszentrum „Lernende Systeme und Kognitive Robotik“ des Fraunhofer IPA durchgeführten Studie (Schaaf et al. 2021) wurden verschiedene Methoden unter anderem auf die Kriterien Laufzeit und Wiedergabetreue der Erklärungen untersucht[31]. Für die Anwendung auf Bilddaten wurden die Methoden Integrated Gradients und LIME wegen ihrer Wiedergabetreue zum Modell am besten bewertet. SHAP erreichte in diesem Punkt weniger gute Ergebnisse. Allerdings benötigen die beiden zuerst genannten Ansätze mehr Zeit für die Generierung der Erklärungen als andere Methoden, beispielsweise LRP. Dieses Ergebnis weist einen Unterschied zu den Einschätzungen der Expert:innen auf, die Integrated Gradients (hinsichtlich der Laufzeit) als gut geeignet für Bilddaten einschätzten.

Bei der Anwendung auf tabellarische Daten erreichten LIME und SHAP sehr ähnliche Ergebnisse. Counterfactual Explanations sind hervorzuheben, da bei dieser Methode die Wiedergabetreue zum Modell stets gegeben ist. Allerdings dauert die Erklärungsgenerierung relativ lang. Gut abgeschnitten hat außerdem das Surrogat-Modell (hier: generierter Entscheidungsbaum) – vor allem hinsichtlich einer kurzen Laufzeit (Schaaf et al. 2021).

Zur Unterstützung bei der Wahl eines geeigneten Erklärungswerkzeuges wurden die beschriebenen Empfehlungen als „Orientierungsbaum“ zusammengefasst (siehe Abbildung 9). Bei der Nutzung ist zu beachten, dass nur eine Auswahl von bereits etablierten Erklärungsstrategien und -werkzeugen betrachtet wurde und die Angaben auch auf Erfahrungen mit konkreten Anwendungsfällen beruhen. Der aufgezeigte „Orientierungsbaum“ soll die im Rahmen der Studie gewonnenen Erkenntnisse auf vereinfachte Weise darstellen und bei der Auswahl von Erklärungsstrategien und -werkzeugen eine grobe Orientierung bieten.

31 Die Wiedergabetreue gibt an, inwiefern die Erklärung das Verhalten des Modells widerspiegelt.

Der Orientierungsbaum

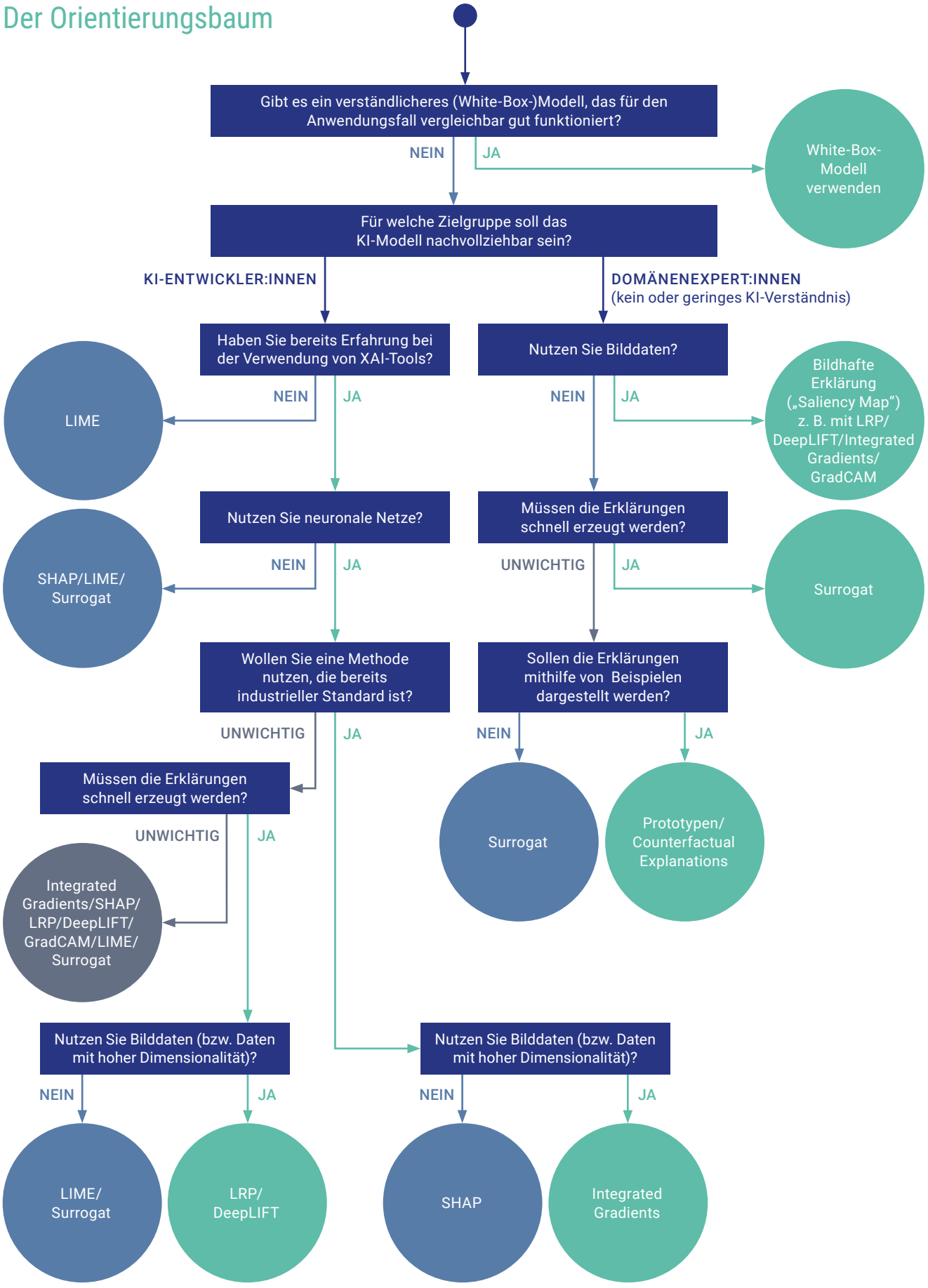


Abbildung 9: Der Orientierungsbaum unterstützt bei der Wahl geeigneter Erklärungswerkzeuge (XAI-Tools)

7 HERAUSFORDERUNGEN UND HANDLUNGSBEDARFE FÜR DIE ETABLIERUNG ERKLÄRBARER KI

7 HERAUSFORDERUNGEN UND HANDLUNGSBEDARFE FÜR DIE ETABLIERUNG ERKLÄRBARER KI

In den vorherigen Kapiteln wurden technische Möglichkeiten dargelegt, wie man eine Erklärbarkeit von KI-Systemen für konkrete Anwendungsfälle gewährleisten kann. Im Rahmen der für die Studie durchgeführten Interviews mit Expert:innen wurden ebenfalls die künftigen technischen und regulatorischen Herausforderungen und Handlungsbedarfe für die Realisierung erklärbarer KI-Systeme diskutiert.

Die Expertinnen und Experten wurden um ihre Einschätzungen zu vorausgewählten thematischen Aspekten gebeten – und zwar zur Relevanz und zum Schwierigkeitsgrad, die sie den vorgelegten Themen jeweils zuschreiben, sowie zu Zeiträumen, in denen Lösungen zur Überwindung der jeweiligen Herausforderung bereitstehen könnten.

Es folgt zunächst die Zusammenfassung der Diskussionsergebnisse zu den wesentlichen technischen Herausforderungen für die Realisierung erklärbarer KI-Systeme in Abschnitt 7.1. Im darauffolgenden Abschnitt 7.2 findet sich die Zusammenfassung der Diskussionsergebnisse zu den größten regulatorischen Herausforderungen.

7.1 Technische Herausforderungen und Handlungsbedarfe

Fünf technische Herausforderungen erwiesen sich als besonders relevant, welche jeweils von beinahe allen zum Themenfeld befragten Personen als sehr wichtig und zugleich lösbar erachtet wurden. Diese sind in Abbildung 10 dargestellt.

Wenn man zunächst die Themen betrachtet, die laut Expert:innen vergleichsweise kurzfristig umgesetzt werden können und sollten, dann findet sich dort das Formulieren von **Best Practices zur Auswahl geeigneter Erklärungsstrategien**, wozu auch diese Studie einen Beitrag leisten soll. In der Diskussion wurde deutlich, dass sich in der Wissenschaft in Teilgebieten bereits Best Practices herausbilden, insbesondere im Bereich Supervised Learning. Es kann auf eine wachsende wissenschaftliche Literatur, entsprechende Software-Prototypen und vereinzelte Success Stories zurückgegriffen werden. Diese wissenschaftlichen Best Practices sind jedoch für Unternehmen, die keine eigenen KI-Forschungsabteilungen unterhalten, häufig unbrauchbar, da sie zumeist auf sehr eng gefasste Anwendungsszenarien bezogen sind.

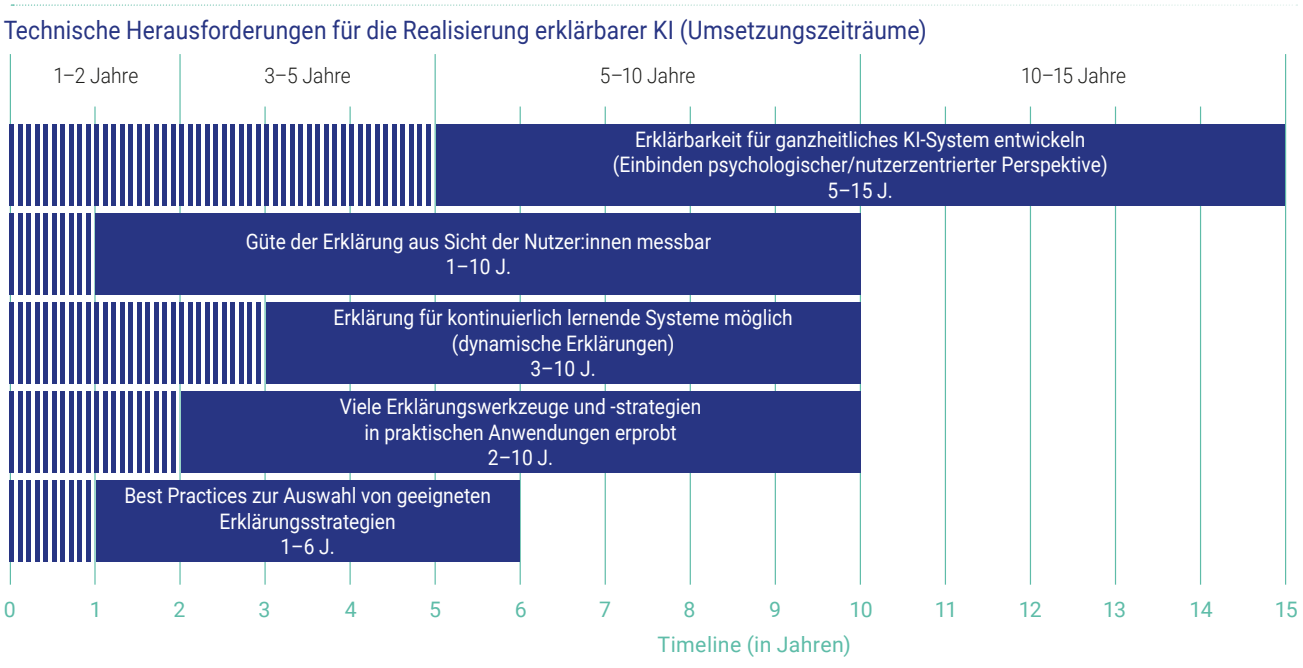


Abbildung 10: Ergebnis aus Experteninterviews – Größte Herausforderungen für die Realisierung erklärbarer KI-Systeme aus technischer Perspektive und mögliche Zeiträume für die Umsetzung.

Typische Rahmenbedingungen aus Unternehmenssicht werden dabei in der Regel nicht berücksichtigt: z. B. weniger KI-affine Nutzer:innen, Zeitdruck oder unrealistische, störungsbehaftete Datengrundlagen. Derartige Faktoren und auch die Passfähigkeit der Erklärungsstrategien in bestehende europäische oder internationale Regulatorik- und Ethik-Systeme spielen für Unternehmen, insbesondere KMU, eine wichtige Rolle.

Für die Herausbildung anwendungsbezogener Best Practices setzen die Expert:innen je nach Anwendungsgebiet, Problemkomplexität und Testmöglichkeit ein bis fünf Jahre an. Die Spanne erklärt sich aus den sehr unterschiedlichen Anforderungen, die sich auf reine Systeme zur Entscheidungsunterstützung oder auch auf hochautonomisierte Prozesse beziehen können, sowie aus dem fehlenden, einheitlichen Verständnis davon, was ausreichende Erklärung bedeutet.

Von einer großen Mehrheit der Expert:innen wird zudem der Eindruck geteilt, viele Methoden für erklärbare KI seien **praktisch noch nicht ausreichend erprobt** und folglich in vielerlei Hinsicht noch Forschungsgegenstand. Eine Ausnahme bildet hier der Bereich der natürlich-sprachlichen Erklärungen im Umfeld des „Natural Language Processing", wo von US-amerikanischen Großunternehmen wie Google und Facebook bereits umfangreiche professionelle und kommerziell erfolgreiche Umsetzungen existieren. Abgesehen davon, wird aber vielfach noch eine große Lücke zwischen wissenschaftlicher Theorie und industrieller Umsetzung gesehen, vor allem für KMU: Für Praktiker:innen aus der Industrie sind publizierte Methoden häufig nicht ausreichend gut portierbar implementiert und in wissenschaftlichen Artikeln behandelte Testprobleme häufig wenig realistisch. Trotz dieser Hindernisse ist das Unternehmensinteresse an erklärbarer KI laut den Expert:innen vorhanden. Ungeklärte regulatorische Vorgaben werden von mehreren Fachleuten ebenfalls als wesentliche Hürde für die praktische Erprobung von erklärbarer KI sowie von KI im Allgemeinen empfunden. Folglich wird von den Expert:innen für eine umfangreiche Erprobung erklärbarer KI aufgrund unterschiedlicher Anwendungsgebiete und Anforderungen eine Spanne von zwei bis zehn Jahren veranschlagt.

Ein weiterer Bereich, der aufgrund der Unterschiedlichkeit der KI-Anwendungen und der zugehörigen Anforderungen an Erklärungen ebenfalls sehr unterschiedlich von den Expert:innen eingeschätzt wurde, ist die vermeintliche **Schwierigkeit bei der Bereitstellung von Erklärungen für kontinuierlich lernende Systeme**. Laut mehreren Expert:innen ist eine technische Lösung bereits heute für solche Anwendungen möglich, für die eine einfache Erklärungsstrategie mit Entscheidungserläuterungen grundsätzlich ausreicht. Gleiches gilt, falls die Phasen, in denen Systeme sequenziell neu trainiert werden, ausreichend Zeit lassen, um auch „offline" die Erklärungsstrategien anzupassen. Dynamische Erklärungen, die sich tatsächlich individuell an System und Anwender:innen anpassen, werden hingegen von den Expert:innen als deutlich schwerer umsetzbar eingeschätzt. Lösungen halten viele Fachleute nur über den Ausbau der Mensch-Maschine-Interaktion für möglich. Für Systeme dieser Kategorie werden von den Expert:innen erste Lösungen in zehn Jahren erwartet, sodass insgesamt entsprechende Lösungen in drei bis zehn Jahren erwartet werden.

Die **Messbarkeit der Erklärungsgüte aus Sicht der Nutzer:innen** hält die Mehrheit der Expert:innen ebenfalls für eine wichtige Herausforderung, für einige stellt sie sogar ein Schlüsselthema dar. Gerade um Akzeptanz und Vergleichbarkeit herzustellen, wird dieser Aspekt als besonders relevant angesehen. Bei der Entwicklung entsprechender Ansätze sollten Methoden aus weiteren Disziplinen, wie beispielsweise Verhaltenswissenschaften und Psychologie, mitberücksichtigt werden. Jedoch wirft die konkrete Umsetzung noch Fragen auf. Eine automatisierte, algorithmische Lösung wird als sehr schwierig bzw. große Herausforderung angesehen. Wahrscheinlicher ist aus Sicht der Expert:innen, dass Lösungsansätze nur anwendungsspezifisch entwickelt werden oder dass eine Umsetzung über Studien mit Nutzer:innen erforderlich sein wird. Je nach angestrebter konkreter Umsetzung – Studien oder algorithmisch – wird ein Lösungszeitraum von einem bis zu zehn Jahren erwartet.

Eine Herausforderung, die das Einbinden einer nutzer:innenzentrierten Perspektive und die zuvor thematisierte Messbarkeit der Erklärungsgüte beinhaltet, ist es, eine **Erklärbarkeit für ganzheitliche KI-Systeme zu entwickeln**. Von einigen Expert:innen wird hier ebenfalls die Berücksichtigung von Ansätzen aus der Psychologie und den Kognitionswissenschaften betont. Die vordergründige Problematik ist insbesondere, dass die Nutzerin oder der Nutzer ein falsches Sicherheitsgefühl

vermittelt bekommen könnte, wenn das KI-System nicht ausreichend nachvollziehbar ist. Der Fokus einer entsprechenden Lösung sollte immer auf der Herstellung gerechtfertigten Vertrauens oder Misstrauens beim Anwender und der Anwenderin liegen. Verlässt sich der Mensch blind auf die Entscheidung des Systems, droht der Verlust von Problemlösungskompetenzen und von technischem Know-how. Während eine Mehrheit der Expert:innen das Thema als wichtig und herausfordernd einordnet, hielten es einige für weniger bedeutend. Von den Expert:innen wird als Lösungszeitraum eine Spanne von 5–15 Jahren erwartet.

7.2 Regulatorische Herausforderungen und Handlungsbedarfe

Unterschiedliche Gremien auf nationaler und europäischer Ebene beschäftigen sich bereits rege mit den besonderen Anforderungen, die sich aus den Eigenschaften von KI-Systemen für deren Zulassung ergeben. Regulatorische Anforderungen werden generell anwendungsbezogen und unabhängig von konkreten Modellen und Verfahren formuliert. Dennoch ist offensichtlich, dass sich insbesondere die Forderung nach der Erklärbarkeit von KI-Systemen maßgeblich aus der besonderen Eigenschaft von Black-Box-KI-Systemen – ohne ein vorgefertigtes Regelwerk Entscheidungen treffen zu können – ableitet.

Heute existieren in strikt regulierten Anwendungsfeldern wie Gesundheit, Prozessindustrie, kritische Infrastrukturen etc. vonseiten des Gesetzgebers zumeist keine klaren Vorgaben, an denen sich zuständige Zulassungsstellen und Entwickler:innen in Bezug auf Erklärbarkeit orientieren könnten. Und falls Vorgaben existieren, sind

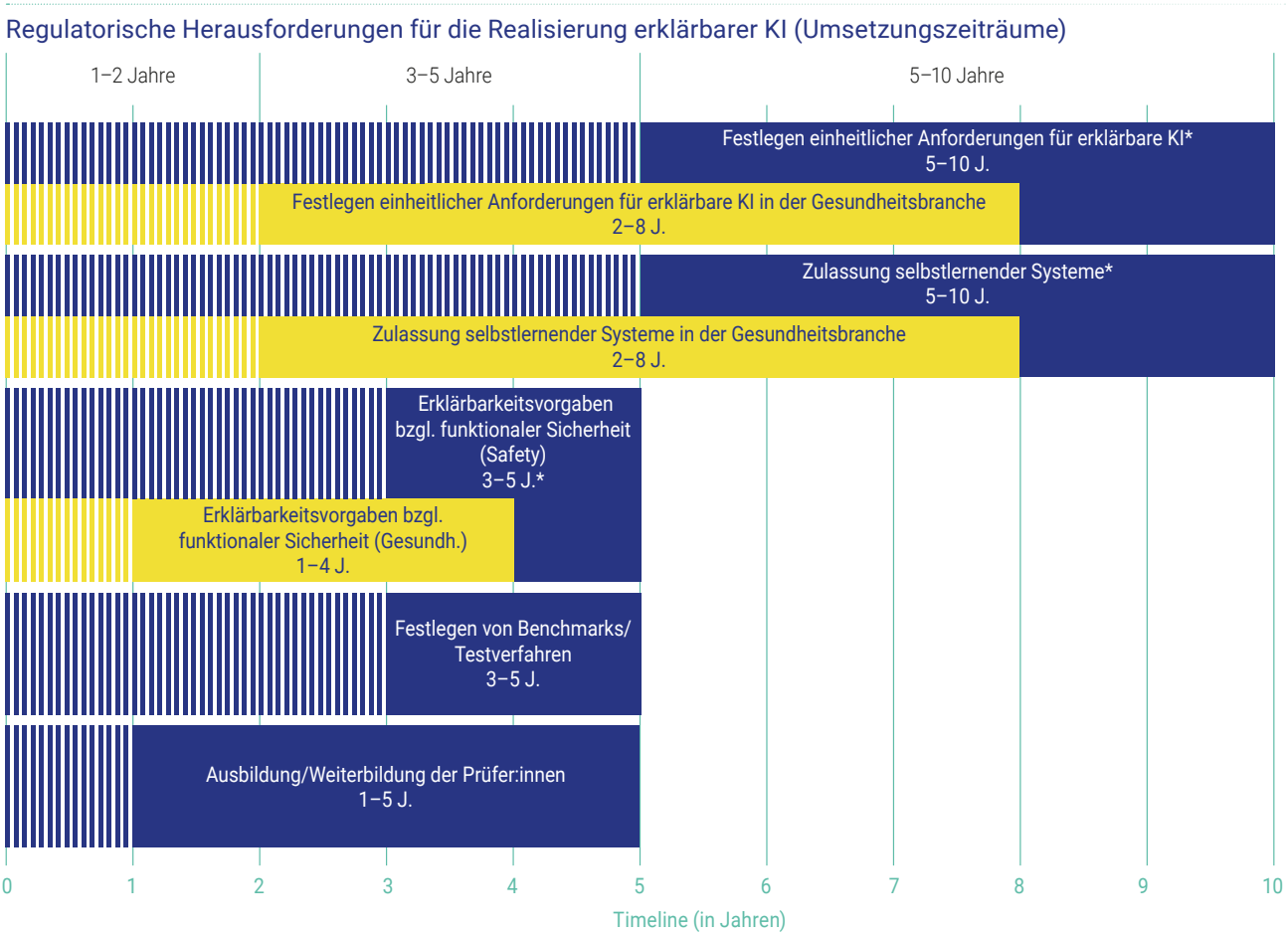


Abbildung 11: Ergebnis aus Experteninterviews – Größte Herausforderungen für die Realisierung erklärbarer KI-Systeme aus regulatorischer/rechtlicher Perspektive und mögliche Zeiträume für die Umsetzung (Einschätzung der allgemeinen Entwicklung in Blau und der teilweise schneller erwarteten Entwicklung in der Gesundheitsbranche in Gelb).

* Die Aussagen der Expert:innen aus der Gesundheitsbranche sind hier ausgenommen.

diese technisch nicht selten so herausfordernd, dass die Nutzung bestimmter KI-Modelle bzw. -Verfahren implizit ausgeschlossen wird, ohne dass dies anwendungsseitig zwangsläufig gerechtfertigt ist. Da sich dies hemmend auf den effektiven Einsatz von Verfahren der KI auswirkt, wurden in den Gesprächen mit den Expert:innen regulatorische Herausforderungen hinsichtlich ihrer Bedeutung und Umsetzbarkeit diskutiert. Das in Abbildung 11 dargestellte Meinungsbild repräsentiert eine Zusammenfassung der Diskussionsergebnisse mit allen Expert:innen.

Fünf Aspekte erwiesen sich in der Diskussion mit den Expert:innen als besonders relevant: **Aus- und Weiterbildung der Prüfer:innen**, Festlegung von Benchmarks und Testverfahren, funktionale Sicherheit, Festlegen einheitlicher Anforderungen für erklärbare KI sowie die Zulassung selbstlernender Systeme. Bei den Zeiträumen, die die Expert:innen für mögliche Lösungen veranschlagen, zeigen sich durchaus unterschiedliche Auffassungen. Dabei fällt auf, dass die Mehrzahl der Personen aus dem Gesundheitsbereich kürzere Zyklen bei der Umsetzung einzelner Aspekte für machbar halten. Wir werden im Folgenden auf die allgemeinen Einschätzungen und beobachtete Besonderheiten eingehen.

Wenn man zunächst die Themen betrachtet, die laut Expert:innen vergleichsweise kurzfristig umgesetzt werden können und sollten, findet sich dort die Aus- und Weiterbildung der Prüfer:innen. Obwohl der Mangel an entsprechender Expertise derzeit als wesentliches Hemmnis und potenzieller Flaschenhals in Bezug auf Zulassungsverfahren wahrgenommen wird und je nach Anwendung sehr unterschiedliche Anforderungen beurteilt werden müssen (u. a. Funktionalität, Sicherheitskonzepte, angemessene Einbindung von Domänenexpert:innen in Entscheidungsprozesse), wurde von den meisten Fachleuten eine Umsetzung innerhalb von drei Jahren für möglich gehalten. Von Personen, die auch die vorgelagerte Entwicklung geeigneter Weiterbildungsprogramme als aufwendig erachteten, wurden drei bis fünf Jahre angesetzt.

Das **Festlegen von Benchmarks/Testverfahren** erachtete eine Mehrzahl der Expert:innen als eine mögliche Basis dafür, wie System- und Erklärverhalten von Prüfinstitutionen analysiert und validiert werden könnte. Unterschiedliche Expert:innen wiesen jedoch darauf hin, dass Benchmarks nur bei bestimmten Anwendungen eingesetzt werden können – z. B. autonomes Fahren, Robotikanwendungen – und ein regelmäßiges Anpassen entsprechender Datensätze bei einer derartigen Prüfmethodik zwingend erforderlich wäre. Als Vorteil wurde jedoch die hohe Vergleichbarkeit und die Zeitersparnis hervorgehoben, die sich bei einer Anwendbarkeit ergeben können. Mehrere Expert:innen, die sich zu einer realistischen Umsetzungsdauer äußerten, nannten einen Zeitraum von drei bis fünf Jahren, um für geeignete Zielanwendungen entsprechende Prüfverfahren zu etablieren.

Der Aspekt der **funktionalen Sicherheit** gilt ebenfalls einer Mehrzahl der Interviewpartner:innen als eine elementare Herausforderung für die Realisierung erklärbarer KI-Systeme aus regulatorischer Perspektive. Dies unterstreicht die Bedeutung, die Erklärungen von „Safety"-Aspekten haben – und zwar unabhängig davon, ob der Mensch derjenige ist, der Analyseergebnisse oder vorgeschlagene Entscheidungen umsetzt, oder entsprechende Aktoren. In dem Zusammenhang wurde von einem Experten die Sicht geäußert, dass jedes KI-System, dessen Entscheidungen nicht von einer oder einem informierten Nutzer:in hinreichend häufig überprüft wird, prinzipiell als autonom betrachtet werden muss. Mit Blick auf die gegenwärtige regulative Umsetzung zeigen sich gewisse Unterschiede zwischen den betrachteten Anwendungsbranchen in Abhängigkeit vom Automatisierungsgrad: Bereits heute müssen etwa in der Prozessindustrie, unabhängig von der zugrunde gelegten Algorithmik, detaillierte Schutzkonzepte gemäß Störfallverordnung vorgelegt werden. Schutzkonzepte, die in aller Regel auch explizit eine angemessene Systemüberwachung durch qualifizierte Mitarbeitende einschließen, müssen nicht nur bei der initialen Zulassung für die zulassenden Behörden nachvollziehbar sein, sondern ebenso in regelmäßigen Abständen an den Stand der Technik angepasst und rezertifiziert werden. Während die Pflicht zur entsprechenden Darlegung und Umsetzung von Schutzkonzepten in der Prozessindustrie und teilweise auch in der Produktion (s. Maschinenrichtlinie) somit dem Betreiberunternehmen auferlegt wird, gibt es in anderen Bereichen vielfach noch keine eingespielte Zulassungspraxis. Für die Gesundheitswirtschaft, in der kritikalitätsbedingt praktisch keine autonomen Systeme, sondern nur Entscheidungsunterstützungssysteme eingesetzt werden, erwarten mehrere Expert:innen in ein bis vier Jahren eine Klärung der Anforderungen für erklärbare KI-Systeme in Bezug auf die funktionale Sicherheit.

Bei den zentralen Aufgaben des **Festlegens einheitlicher Anforderungen für erklärbare KI und der Zu-**

lassung selbstlernender Systeme, denen sich der Gesetzgeber stellen muss, prognostizieren mehrere Expert:innen, die in der Gesundheitsbranche tätig sind, eine schnellere Entwicklung konkreter Anforderungen für Erklärbarkeit als in den Anwendungsbereichen Produktion und Prozessindustrie. Übereinstimmend empfehlen beim **Festlegen einheitlicher Anforderungen für erklärbare K**I mehrerer Expert:innen dringend eine europäische Lösung. Auch betonen einzelne Interviewpartner:innen die große politische Dimension dieser Herausforderung, da verschiedene Länder und Gremien in der EU sich dafür auf eine Linie einigen müssen. Das Fehlen einheitlicher Anforderungen in den betroffenen Branchen wird von einem Großteil der Expert:innen als sehr großes Hemmnis für die Entwicklung erklärbarer KI wie auch für KI allgemein in Deutschland wahrgenommen. Dieser Umstand bremst in den betroffenen Branchen Zulassungen, die Investitionsbereitschaft sowie die Entwicklung erklärbarer KI im Allgemeinen. Die Dimension dieser Herausforderung, die zwingend auch unter Berücksichtigung von Branchenspezifika adressiert werden muss, spiegelt sich auch in dem langen prognostizierten Zeitraum für eine mögliche Umsetzung von fünf bis zehn Jahren wider.

Zwingend muss laut den Expert:innen vor allem eine Lösung für die **Zulassung selbstlernender Systeme** gefunden werden. Dabei stellt sich die Kernfrage, wie die entsprechenden Kriterien, die abhängig von den festgelegten Anforderungen sind, angemessen geprüft werden können, wenn ein System sich verändert. Dies beinhaltet u. a. auch die Überprüfung, ob ein System im Lernprozess gegebenenfalls ethisch nicht vertretbare Entscheidungsstrategien entwickelt, was bei einer initialen Zulassung unter Umständen nicht erkennbar ist. Die Aufgabe, die sich dabei den zulassenden Behörden stellt, wurde von der Mehrheit der hierzu befragten Personen als herausfordernd, aber lösbar eingestuft. Es gab jedoch eine Einzelstimme, die dies für derzeit nicht umsetzbar hielt. Mehrere Befragte sahen keine oder kaum eine Herausforderung in dieser Aufgabe, zumindest für klar definierte Anwendungsbereiche. Eine vorgeschlagene Herangehensweise umfasst etwa extern durchgeführte Rezertifizierungen, die nach einer Initialisierungszulassung regelmäßig beim Neutrainieren von Modellen angestoßen werden könnten, unterstützt z. B. durch ein Ticket-System. Bei Modellen, die sich allerdings zwangsläufig permanent verändern – etwa, weil auf die Veränderung von Umgebungsparametern reagiert werden muss –, lassen sich Intervalle für extern angestoßene Zertifizierungen nur schwerlich oder gar nicht sinnvoll festlegen. Weil eine quasikontinuierliche Prüfung durch externe Stellen aus unterschiedlichsten Gründen in solchen Fällen nicht in Frage kommt, wurde die Selbstzertifizierung durch Betreiberunternehmen als sinnvolle Variante vorgeschlagen.

8 FAZIT

8 FAZIT

Es existieren Konzepte, um Ausprägungen von „Black-Box"-Modellen und transparenten Modellen („White-Box") über die drei Transparenzstufen der Simulierbarkeit, der Unterteilbarkeit und der algorithmischen Transparenz zu unterscheiden. Dabei zeichnen sich White-Box-Modelle durch die algorithmische Transparenz und die Nachvollziehbarkeit der Eingangsgrößen aus. Dies unterscheidet sie entscheidend von Black-Box-Modellen, die bei Modellen niedrigster Dimension, die in der Anwendung noch von praktischem Nutzen sind, keine der drei zuvor genannten Eigenschaften erfüllen – insbesondere nicht die niedrigste Transparenzstufe der algorithmischen Transparenz. Dabei ist laut Literatur das entscheidende Kriterium für die Eigenschaft der algorithmischen Transparenz, ob ein Modell bzw. die Modellgenerierung hinreichend zugänglich für mathematische Analysen ist. Es wurde eine Zuordnung von gängigen Verfahren in die White-Box- und Black-Box-Kategorie sowie die zuvor genannten Transparenzkonzepte aus der Literatur bereitgestellt (siehe Kapitel 2).

→ Es wurde im Rahmen der Studie nicht erhoben, wie bekannt das Konzept der Modelltransparenz in den jeweiligen wissenschaftlichen Communities ist. Die einschlägigen Literaturbeiträge, die sich auch auf erklärbare KI beziehen, wurden erst in den letzten fünf Jahren veröffentlicht. Teilweise unklar formulierte und widersprüchliche Begrifflichkeiten sowie heute noch erscheinende Artikel, die die Kategorie von Black-Box-Modellen generell ablehnen, oder andererseits solche, die den Einsatz von Black-Box-Modellen, wie z. B. neuronale Netze, für kritische Anwendungen ablehnen, deuten jedoch an: Entsprechende Diskurse werden mitunter noch sehr unterschiedlich in den wissenschaftlichen Communities geführt. Eine Vereinheitlichung der Taxonomie steht von wissenschaftlicher Seite noch aus.

Die Betrachtung der etablierten Erklärungsstrategien und -werkzeuge zeigt, dass einzelne Methoden darauf ausgelegt sind, nur für einen bestimmten KI-Modelltypus Erklärungen zu generieren; andere können nur bei der Verwendung bestimmter Datenarten eingesetzt werden (siehe Kapitel 3). Es wurden konkrete Vor- und Nachteile bei der Verwendung der einzelnen Ansätze herausgestellt, wobei sich vor allem eines zeigt: Falls nur Entscheidungserklärungen (lokale Erklärbarkeit) gefordert sind, bieten etablierte Post hoc-Analysewerkzeuge Möglichkeiten, um Black-Box-Modelle, z. B. neuronale Netze, besser nachvollziehen zu können. Erklärungswerkzeuge wie Integrated Gradients und SHAP, die vor allem zur Erklärung von Einzelentscheidungen verwendet werden, erlangten bereits industrielle Reife, sind aber in ihrer Bedienung nicht sehr intuitiv und daher im Allgemeinen als Werkzeuge für KI-Entwickler:innen zu begreifen. Für KI-Nutzende werden oft intuitivere Ansätze wie Saliency Maps und Counterfactual Explanations bevorzugt.

Nach Einschätzung der meisten Entwickler:innen und Anwender:innen werden in fünf bis zehn Jahren die Modellvarianten der neuronalen Netze den wichtigsten Modelltyp im Bereich der Künstlichen Intelligenz darstellen (siehe Kapitel 4). Dabei kommen durchschnittlich rund zwei Drittel der Personen, die Varianten neuronaler Netze auch aktiv nutzen, zu dem Schluss, dass diese bereits heute teilweise erklärbar sind – zumindest in Bezug auf Einzelentscheidungen und bei Nutzung geeigneter Erklärungsstrategien. Dass umgekehrt ein Drittel der Personen, die neuronale Netze nutzen oder entwickeln, diese für gar nicht erklärbar halten, deutet auf eine gewisse Unbekanntheit existierender Analysewerkzeuge hin.

Gleichzeitig teilt vor allem eine große Mehrheit der Branchenkenner:innen der Gesundheitswirtschaft die Einschätzung, dass KI für einen professionellen Einsatz in der Domäne zwingend erklärbar sein muss. Aber auch in diversen anderen Branchen (Finanz-, Produktions-, Bauwirtschaft, Prozessindustrie, Energiewirtschaft, Dienstleistungssektor) erachtet eine Mehrheit der Personen mit entsprechender Branchenkenntnis eine gewisse Erklärbarkeit als unverzichtbar. Dies liegt in diesen Anwendungsfeldern jedoch meist weniger an strikten Zulassungsvorgaben, sondern vielmehr an den potenziellen Kund:innen und Anwender:innen, die nicht erklärbare KI-Systeme für „typische" Branchenanwendungen heute schlicht nicht akzeptieren würden. Dass KI-Erklärbarkeit zukünftig auch für weitere Akteur:innen wie interne Prüfer:innen, Management und Endkund:innen laut Umfrage zunehmend wichtig wird (siehe Kapitel 4), unterstreicht den perspektivischen Bedarf für erklärbare KI, auch abseits von regulatorischen Anforderungen.

→ Die Beobachtungen zur Nichtkenntnis von Methoden deuten an, dass der wissenschaftlich-methodische Diskurs zwischen den Disziplinen der Informatik (insbesondere Data Science) und der Mathematik (insbesondere Statistik und numerische Mathematik) in Grundlagenforschung und Ausbildung sowie die Herausbildung von Best Practices vorangetrieben werden sollten.

Der Vergleich anhand von vier Use Cases (siehe Kapitel 5) zeigt: Zwei Motivationsgründe für die Nutzung

erklärbarer KI sind allen gemein, nämlich Kausalitätsbeziehungen zu identifizieren und Konfidenz zu bestimmen. Dabei steht die erste, eher typische, Motivation bei den beiden Anwendungsfällen zur Anomalieerkennung – Bildanalyse histologischer Gewebeschnitte und Maschinenzustandsüberwachung – klar im Vordergrund. Ein Unterschied besteht neben der unterschiedlichen Datengrundlage darin, dass im medizinischen Anwendungsfall der Standard für Erklärbarkeitsanforderungen von Seite der Zulassungsbehörden definiert wird, während das ärztliche Personal für den operativen Betrieb faktisch nur Entscheidungserklärungen (lokale Erklärbarkeit) benötigt. Bei der Maschinenzustandsüberwachung bestehen hingegen keinerlei regulatorische Anforderungen, allerdings erwarten Anwendende hier über Entscheidungserklärungen hinaus häufig auch Modellerklärungen (globale Erklärbarkeit). Entsprechend unterschiedlich sind die Lösungswege: Die Anforderung lokaler Erklärbarkeit bei der Bildanalyse wird mithilfe von Post hoc-Erklärungen eines Black-Box-Modells adressiert, die Anforderung lokaler und globaler Erklärbarkeit mit selbsterklärenden White-Box-Ansätzen (Maschinenzustandsüberwachung). Erklärungen werden im zweiten Fall aber auch aktiv bereitgestellt: einerseits in Form statistischer Eintrittswahrscheinlichkeiten (Bayes-Netze) und eines zusätzlichen Surrogatmodells, anderseits durch natürlichsprachliche Erklärungen, die ein Anwender oder eine Anwenderin selbst verbessern kann.

Das individuelle Ziel, den Informationsgewinn der Domänenexpert:innen zu erhöhen, ist bei dem Use Case der medizinischen Textanalyse von Arztbriefen die zentrale Motivation. Nur so kann von medizinischen Expert:innen überhaupt bewertet werden, ob ein Kriterium, das für die Klassifikation durch die KI ausschlaggebend war, entweder plausibel oder medizinisch nicht sinnvoll ist. Grundlage zur Bereitstellung von Entscheidungserklärungen sind nominelle Black-Box-Modelle (neuronale Netze), die durch Prototypen bzw. externe Wissenssammlungen ergänzt werden, sodass das resultierende Modell selbst medizinisch nachvollziehbare Gründe für einzelne Entscheidungen geben kann. Im Use Case Prozessführung ist „Konfidenzen bestimmen“ eines von mehreren, aber letztlich das entscheidende übergeordnete Ziel, auch wenn es intuitiv nicht zwangsläufig mit Erklärbarkeit in Verbindung gebracht wird. Unentdeckte Fehler in der visuellen Zustandserkennung bzw. Anfälligkeit für Störungen und Bias in den „hybriden“ Modellen können unkalkulierbare Risiken für die robuste und stabile Steuerung und Regelung der chemischen Anlagen nach sich ziehen. Daher entstehen bei diesem Use Case im Vergleich ebenso die weitreichendsten Erklärbarkeitsanforderungen (Erklärbarkeit von Einzelentscheidungen und Modellwirkmechanismen). Hier wird der Ansatz verfolgt, aus mechanistischen Modellen und Simulationsdaten sowie Bild- und Sensordaten geeignete „hybride“ Modelle zu erstellen, die White-Box- mit Black-Box-Komponenten zu erklärbaren Anlagenmodellen kombinieren.

→ Der Mehrwert von anwendungsbezogenen Fallstudien ist deutlich erkennbar. Die Übertragbarkeit technologischer Ansätze auf andere Anwendungsfelder ist vergleichsweise leicht möglich, wenn Problemstellungen sich strukturell ähneln, z. B. beim Datentyp, den Zielen etc. Dabei ist unbedingt empfehlenswert, verstärkt Anwendungen zu adressieren, die auch die Erklärung von Modellwirkmechanismen (Generierung globaler Erklärbarkeit durch „hybride“ Systeme), die Interaktion zwischen Mensch und KI-System zur Verbesserung von Erklärungen (wie z. B. im Use Case Maschinenzustandsüberwachung) oder Erklärungen für (teil-)autonome Systeme fokussieren (wie z. B. im Use Case zur KI-gestützten Prozessführung). Diese bedeutsamen Anwendungsfelder wurden bisher nur vereinzelt in der anwendungsorientierten Forschung adressiert.

Die Empfehlungen der Expertinnen und Experten sowie die Erkenntnisse aus der Literatur wurden in eine praktische Orientierungshilfe überführt (siehe Kapitel 6). Diese soll hinsichtlich erster praktischer Schritte bei der Auswahl von Erklärungsstrategien unterstützen. Dabei ist eine zentrale Erkenntnis, dass zumindest auf absehbare Zeit Erklärungswerkzeuge fehlen werden, die detaillierte und quantitativ nutzbare Modellerklärungen für Black-Box-Modelle, wie z. B. neuronale Netze, bereitstellen können. Grundsätzlich gilt deshalb, dass White-Box-Modelle bei entsprechenden Erklärbarkeitsanforderungen stets zu bevorzugen sind, wenn sie im Vergleich zu Black-Box-Modellen ähnlich gut funktionieren, oder zumindest hinreichend gut in Bezug auf die Anwendung. Falls Modellerklärungen und die Nutzung von Black-Box-Modellen erforderlich sind, ist perspektivisch auch die Nutzung „hybrider“ Ansätze vielversprechend, die White-Box- mit Black-Box-Komponenten kombinieren und eigenständig Erklärungen bereitstellen[32]. Falls Erklärungen von Einzelentscheidungen ausreichen, bietet der „Orientierungsbaum“ eine Entscheidungshilfe, bezüglich der in Kapitel 3 diskutierten Erklärungsstrategien.

32 Einzelne Verfahren werden derzeit in Forschungsprojekten entwickelt oder weiterentwickelt und wurden teilweise in der Studie vorgestellt (Projekte KEEN und Service-Meister des BMWi-Technologieprogramms KI-Innovationswettbewerb sowie das Projekt RAKI des BMWi-Technologieprogramms Smarte Datenwirtschaft).

→ Der Orientierungsbaum berücksichtigt zwar die meistzitierten Ansätze (im Falle ihrer praktischen Anwendbarkeit), stellt damit jedoch lediglich eine Momentaufnahme für den Stand der Technik dar. Im Sinne der Vervollständigung von Best Practices ist es generell empfehlenswert, diese eingeschlagene Richtung weiterzuverfolgen und dabei insbesondere quantitative Vergleiche und die Prüfung der Übertragbarkeit der Ansätze noch stärker zu berücksichtigen.

In Bezug auf die technischen Herausforderungen für erklärbare KI zeigt sich: Derzeit liegen noch keine vervollständigen Best Practices vor, die für Unternehmen, insbesondere KMU, nutzbar sind und genügend Anwendungsfelder abdecken (siehe Kapitel 7). Damit eng verbunden ist das Defizit, dass in der wissenschaftlichen Forschung häufig praxisferne Beispielanwendungen untersucht werden und Erprobungen an realen Problemstellungen fehlen – wenn man von einzelnen High-Tech-Unternehmen absieht.

→ Zwar werden in einzelnen Success Stories bzw. Forschungsprojekten (siehe Use Cases) Anwenderanforderungen bezüglich der Nachvollziehbarkeit von KI bereits adressiert; die Umfrageergebnisse und die Aussagen der Expert:innen legen jedoch nahe, dass der Marktbedarf für erklärbare KI weiter anwachsen wird. Es ist daher empfehlenswert, die Entwicklung von branchenspezifischen Lösungen mit gezielten Aktivitäten der angewandten Forschung zu unterstützen.

→ Angesichts der von den meisten Expert:innen beobachteten Defizite und der Reichhaltigkeit methodischer Ansätze ist insbesondere zu empfehlen, die Effizienz und Anwenderfreundlichkeit der Lösungen von praktischen Problemstellungen aus der Praxis verstärkt ins Zentrum von Forschungsaktivitäten zu stellen.

→ Um die laut Meinung der Expert:innen teilweise vernachlässigte Anwender:innenperspektive zu stärken sowie die bestehende Kluft zwischen Wissenschaft und Industrie zu adressieren, sollten stets quantitative Vergleiche zwischen Alternativansätzen und dem Stand der Technik angestellt werden, wenn diese möglich sind.

→ Die Neu- und Weiterentwicklung geeigneter „hybrider" Ansätze, die daten- und wissensgetriebene Ansätze – bzw. White- und Black-Box-Modellierungsansätze – kombinieren, kann durch die Etablierung interdisziplinärer anwendungsorientierter Forschungsverbünde aus Fachleuten der Informatik, Mathematik (Statistik, Numerik) sowie der verschiedenen Anwendungsdisziplinen gefördert werden.

Die Berücksichtigung verhaltens- bzw. kognitionswissenschaftlicher Aspekte von erklärbarer KI (Messbarkeit der Erklärung, Erklärbarkeit ganzheitlicher KI-Systeme, automatisierte Erklärungsanpassungen an Nutzende und selbstlernende Systeme) wird von einer Mehrheit der interviewten Expert:innen als wichtige Forschungsrichtung und aktuell große Herausforderung gesehen.

→ Die Aussagen der Expert:innen zu den nutzerzentrierten Themenfeldern erklärbarer KI sowie die Einschätzung, Lösungen seien eher mittel- bis langfristig zu erwarten, verdeutlichen: Hier gibt es diverse offene Fragen, die zunächst vonseiten der Grundlagenforschung beantwortet werden müssen.

Generell werden die technischen Herausforderungen als überwindbar eingeschätzt. In vielen potenziellen Zielbranchen von erklärbarer KI sind Systeme jedoch zulassungspflichtig.

Es fehlen in den meisten betroffenen Branchen, z. B. der Gesundheitswirtschaft, allerdings klare regulatorische Vorgaben oder Zulassungsrichtlinien, an denen Unternehmen sich orientieren oder technische Entwicklungen ausrichten können.

In Bezug auf die größten regulatorischen Herausforderungen (siehe Kapitel 7) zeigt sich folgendes: Das Fehlen einheitlicher Anforderungen für Erklärbarkeit von KI ist derzeit das größte Hemmnis für die Entwicklung erklärbarer KI mit Folgen für die generelle KI-Nutzung. Hierdurch werden in den betroffenen Branchen Zulassungen verhindert und folglich die Investitionsbereitschaft sowie die Entwicklung erklärbarer KI gebremst. Indirekt betrifft dieser Missstand aber ebenso die Innovationsentwicklungen in anderen unregulierten Bereichen, da Innovationen bzw. „Technologieschübe" bezüglich erklärbarer KI ausbleiben. Nach Meinung der Expert:innen steht auch fest, dass Mechanismen für die Zulassung und (Re-) Zertifizierung selbstlernender Systeme im gleichen Zuge gefunden werden müssen. Bedenklich ist dabei, dass viele Expert:innen für die Festlegung von Anforderungen für Erklärbarkeit, die nach Möglichkeit in eine europäisch einheitliche Regelung münden soll, eine Umsetzung erst in fünf bis zehn Jahren erwarten. Von Kenner:innen der Gesundheitsbranche werden teilweise schnellere Umsetzungen erwartet, insbesondere für die Ausgestaltung erklärbarer KI in Bezug auf funktionale Sicherheit bzw. körperliche Unversehrtheit von Patient:innen. Gleichzeitig veranschlagen die Fachleute für das Festlegen von Testverfahren und Benchmarks nur drei bis fünf Jahre.

Die Aus- und Weiterbildung von Prüfer:innen ist perspektivisch eine enorm wichtige Aufgabe, da diese Personen zukünftig viele Aufgaben von großer gesellschaftlicher Tragweite wahrnehmen müssen: Es müssen vortrainierte Systeme, Systeme, die in größeren zeitlichen Abständen auf Basis aktualisierter Trainingsdaten neutrainiert werden, und kontinuierlich weiterlernende Systeme initial zugelassen und dann regelmäßig rezertifiziert werden.

Dieser Umstand könnte sich zu einem Flaschenhals für die Zulassung von KI-Produkten entwickeln. Unter Berücksichtigung der Tatsache, dass Ausbildungsprogramme entworfen werden müssen, setzen die Expert:innen ein bis fünf Jahre zur Umsetzung dieser Maßnahme an.

→ Bislang fehlen vor allem die Festlegung von Anwendungs- und Risikoklassen, aus denen das grundsätzliche Erfordernis einer Erklärung ableitbar ist, sowie die Festlegung nachvollziehbarer, angemessener und möglichst einheitlicher sowie quantitativer Anforderungen an Erklärbarkeit, zumindest auf Anwendungs- oder Risikoklassenebene. Zwar wird auf Basis der in Kürze von der EU zu veröffentlichenden Vorlage für die Regulierung von KI* möglicherweise eine Risikoklasseneinteilung vorliegen; trotzdem ist nicht zu erwarten, dass dadurch branchen- und anwendungsspezifische Erklärbarkeitsanforderungen bzw. quantitative Zulassungs- und Zertifizierungsrichtlinien für KI-Produkte vorgelegt werden. Es ist daher empfehlenswert – angesichts der laut Expert:innen langwierigen Umsetzungsprozesse – , zeitnah spezifische Zulassungs- und Zertifizierungsrichtlinien für KI-Produkte zu entwickeln, zumindest für die in Deutschland gesellschaftlich und wirtschaftlich bedeutsamsten Anwendungsgebiete von erklärbarer KI. Dabei sollten Wissenschafts-, Unternehmens- und Standardisierungsvertreter:innen sowie Prüfeinrichtungen an einem solchen Entwicklungsprojekt beteiligt werden, um einerseits einen möglichst breiten gesellschaftlichen Konsens zu erzielen und andererseits die praktische Umsetzbarkeit in Prüfung und Zertifizierung bei der Entwicklung der Richtlinien sicherzustellen. ●

* Bei Redaktionsschluss der Studie wurde hier gemäß Ankündigung der EU-Kommission von einer Veröffentlichung im April 2021 ausgegangen.

A ÜBERBLICK KI-VERFAHREN UND -MODELLE

A ÜBERBLICK KI-VERFAHREN UND -MODELLE

Durch KI-Verfahren lassen sich Lösungen für vielfältige Klassifikations-, Regressions- und Clustering-Probleme[33] finden oder zumindest approximieren. Häufig werden die Begriffe KI und maschinelles Lernen synonym verwendet. Beim maschinellen Lernen geht es darum, dass ein Algorithmus auf der Grundlage von Trainingsdaten „lernt" ein Problem zu lösen[34]. Dabei werden während des eigentlichen Lernprozesses die Freiheitsgrade einer vorgegebenen Modellstruktur an die jeweiligen Daten bzw. die spezifische Problemstellung angepasst. Die „Beurteilung" neuer, unbekannter Daten gemäß der Aufgabenstellung erfolgt dann über eine entsprechende Auswertung des angepassten „KI-Modells".

Die eigentliche Anpassung der Modelle kann dabei – abhängig von der Anzahl der spezifischen Modellfreiheitsgrade – die Manipulation einiger weniger oder von Millionen von Parametern erfordern. Während die Anpassung eines linearen Regressionsmodells in einem einfachen Fall nur die Bestimmung eines einzigen Modellparameters erfordert, um abhängige Variablen mit unabhängigen Variablen in Beziehung zu setzen, müssen für die Anpassung vielschichtiger Modelle, die eine starke Verflechtung aufweisen, wie beispielsweise neuronale Netze, zumeist hunderte von Parametern angepasst werden. Handelt es sich sogar um sogenannte tiefe (neuronale) Netze – umfasst das Netz also viele Parameter und Schichten (wobei es keine genaue Definition für „viel" gibt) –, so erreicht man schnell die angeführten Millionen von Parametern, die angepasst werden müssen. Man spricht in diesem Fall auch vom „Deep Learning".

Im Folgenden werden ausgewählte KI-Modelle und -Verfahren kurz beschrieben.

Ein **Autoencoder** ist ein neuronales Netz, das zur komprimierten Kodierung von Daten genutzt wird. Der Autoencoder besteht aus zwei Komponenten: einem Encoder, der die Eingabedaten komprimiert, und einem Decoder, der aus den komprimierten Daten die ursprünglichen Daten rekonstruiert. Encoder und Decoder können auch separat verwendet werden. Typische Anwendungen sind z. B. Anomalieerkennung oder Dimensionsreduktion (Badr 2019).

Bayes-Netze sind probabilistische bzw. graphische Modelle, die in Form gerichteter azyklischer Graphen vorliegen, deren Knoten Zufallsvariablen und deren Kanten bedingte Wahrscheinlichkeiten beschreiben. Sie sind besonders gut geeignet, um Wahrscheinlichkeiten für die mögliche Ursache von eingetretenen Ereignissen zu quantifizieren. Bayes-Netze können z. B. genutzt werden, um Entscheidungsprobleme unter Unsicherheiten zu lösen.

Clustering-Modelle werden genutzt, um eine Datenmenge automatisch in Untergruppen ähnlicher Datenpunkte aufzuteilen. Der häufig verwendete K-Means-Clustering-Algorithmus ist ein schneller iterativer Algorithmus, der nach anfänglich zufällig ausgewählten Clusterzentren diese immer weiter anpasst, sodass der „Clustering Error" minimiert wird.

In der Bildverarbeitung häufig eingesetzte neuronale Netze sind **Convolutional Neural Networks (CNNs)**. Aufgrund der lokalen Vernetzungsarchitektur zwischen Schichten (ähnlich den Schichten im visuellen Kortex) kann ihre Auswertung als Faltung (engl. „convolution") erfolgen. CNNs eignen sich besonders für Anwendungsgebiete, wo Nachbarschaft zwischen Features eine Rolle spielt, z. B. Pixel in einem Bild oder Worte in einem Satz. Ähnlich wie in den Sehregionen des Säugetiergehirns werden hier von Schicht zu Schicht die rezeptiven Felder größer und die Komplexität der Features, auf die die Einheiten reagieren, nimmt zu.

Dimensionsreduktion wird eingesetzt, um die Anzahl der Daten bzw. Features zu reduzieren, z. B. um die rechnerische Komplexität einer Datenverarbeitung zu verringern oder um die wichtigsten Charakteristika einer Datenmenge zu extrahieren. Principal Components Analysis (PCA) ist ein Beispiel für eine Methode, bei der Features unüberwacht transformiert werden. Auf der anderen Seite kann die Berechnung des Informations-

33 Das Ziel einer Klassifikation ist es, Eingabewerte einer (diskreten) Klasse oder Gruppe zuzuordnen. Ein Beispiel hierfür ist die Bildklassifikation: Bilder, auf denen Tiere abgebildet sind, müssen in eine der beiden Klassen „Hund" oder „Katze" eingeordnet werden. Über Regression wird die Beziehung einer abhängigen Variable, z. B. Körpergröße des Menschen, von unabhängigen Variablen wie der Schuhgröße des Menschen abgebildet. Beim Clustering werden Gemeinsamkeiten und Unterschiede in den Daten analysiert und diese entsprechend gruppiert. Beispielsweise werden im Marketingbereich Gruppen aus ähnlichen Produkten gebildet (ohne dass diese Gruppen vorher bekannt sein müssen), um dem Kunden bzw. der Kundin bei einer Onlinesuche passende Angebote präsentieren zu können.

34 Lernverfahren werden oft unterteilt in überwacht (supervised), unüberwacht (unsupervised), teilüberwacht (semi-supervised) sowie in Reinforcement-Verfahren. Beim überwachten Lernen erhält der Algorithmus Paare aus Eingabe- und Ausgabewerten, die vorgeben, welches Ergebnis bei einer bestimmten Eingabe erwartet wird. Auf dieser Grundlage lernt der Algorithmus, auch neue Eingaben korrekt zuzuordnen. Beim unüberwachten Lernen werden nur Eingabewerte zur Verfügung gestellt. Der Algorithmus muss eigenständig Strukturen in den Daten erkennen. Teilüberwachtes Lernen kombiniert beide zuvor beschriebenen Verfahren. Beim Reinforcement-Learning lernt der Algorithmus selbstständig eine Strategie nur auf Grundlage von positivem oder negativem Feedback (Alloghani et al. 2020; Oladipupo 2010).

gehalts herangezogen werden, um überwacht spezielle Features auszuwählen. Weitere Beispiele sind t-SNE (t-distributed stochastic neighbor embedding) und LDA (Linear Discriminant Analysis) (Cunningham 2008; Balakrishnama und Ganapathiraju 1998). Auch Auto-Encoding Neural Networks können genutzt werden, um eine Dimensionsreduktion zu erwirken (siehe oben).

Bei **Ensemble-Modellen** wird nicht eine einzelne Funktion bzw. ein einzelnes Modell zur Klassifizierung oder Regression auf Grundlage von Daten gelernt, sondern mehrere verschiedene reundante. Die Ergebnisse dieser werden anschließend zusammengeführt, beispielsweise über Bildung des gewichteten Mittelwerts oder Mehrheitsentscheidung. Das Ziel, das bei der Verwendung von Ensemble-Modellen verfolgt wird, ist die Ableitung eines „starken" Klassifizierers (strong learner) aus mehreren „schwachen" Klassifizierern (base classifiers/weak learners). Die einzelnen Klassifizierer müssen sowohl möglichst akkurat als auch divers sein. Random Forest ist ein sehr bekanntes Ensemble-Modell, das auf einzelnen Entscheidungsbäumen beruht. Es können aber auch unterschiedliche Modelle (z. B. Entscheidungsbäume und neuronale Netze) in einem Ensemble kombiniert werden (Goos et al. 2000).

Die Idee hinter **Entscheidungsbäumen** ist, eine komplexe Entscheidung in mehrere einfache Entscheidungen zu unterteilen. Entscheidungsbäume sind hierarchische Strukturen, die sowohl für Klassifikation als auch für Regression eingesetzt werden können. Beispielhafte Algorithmen für die Erstellung von Entscheidungsbäumen sind ID3 oder C4.5. Entscheidungsbäume sind in der Regel auch für Laien gut nachvollziehbar, da an jedem Punkt innerhalb des Baumes erkennbar ist, welche Entscheidung gerade getroffen wurde (Mitchell 2010; Arrieta et al. 2019).

Expertensysteme (oder wissensbasierte Systeme) sind auf dem Wissen von Expert:innen aufgebaute Wissensdatenbanken (meist als Wenn-Dann-Logik repräsentiert), die unter der Verwendung von Inferenzmechanismen aus der Wissensbasis Schlussfolgerungen und Handlungsempfehlungen ableiten oder den Wahrheitsgehalt von Aussagen überprüfen können. Zum Aufbau eines wissensbasierten Systems ist es nötig, über detaillierte Kenntnisse des Anwendungsgebiets zu verfügen und diese entsprechend einer Problemlösestrategie zu formalisieren. Expertensysteme sind in der Regel gut nachvollziehbar (Karst 1992; Puppe 1988; Spreckelsen und Spitzer 2009; Wagner 2000; Lucas und Van der Gaag, Linda C. 1991).

Generative Adversarial Networks (GANs) ist ein Verfahren zum Training von neuronalen Netzen, das insbesondere dann eingesetzt wird, wenn die Menge der Trainingsdaten begrenzt ist. Es werden zwei verschiedene Netze trainiert: Eines (der sogenannte Generator) erzeugt Daten, die so realistisch wie möglich aussehen, das andere (der Diskriminator) versucht, echte Daten von den synthetisch erzeugten zu unterscheiden. Beide Netze werden konkurrierend trainiert mit dem Ziel, möglichst realistische synthetische Daten zu erzeugen, d. h. Daten, die der Diskriminator nicht mehr von den Trainingsdaten unterscheiden kann. Diese können zur Erweiterung der Trainingsgrundlage oder zur Vervollständigung unvollständiger Datensätze in der Anwendung genutzt werden (Creswell et al. 2017).

Long Short Term Memory Networks (LSTMs) gehören zu den Rekurrenten Neuronalen Netzen (RNNs). LSTMs sind besonders gut geeignet, um Sequenzen von Daten zu verarbeiten, z. B. Sprache. Der Vorteil gegenüber RNNs liegt darin, dass „gelernte" Informationen länger behalten werden können – und somit auch Kontextinformationen.

Mathematische Optimierung bezeichnet eine Methodik zur Minimierung oder Maximierung einer mathematischen Zielfunktion, wobei Variablen Nebenbedingungen unterliegen können. Falls Zielfunktionen und Nebenbedingungen lineare Funktionen bezüglich der Entscheidungsvariablen sind, spricht man von linearer Optimierung, andernfalls von nichtlinearer Optimierung. Die analytische Lösung von Optimierungsproblemen ist nur selten möglich, sodass in der Regel numerische Lösungsverfahren eingesetzt werden, um Parameter zu finden, die den jeweiligen Optimalitätskriterien entsprechen. Der Einsatz hocheffizienter Lösungsmethoden ist faktisch unumgänglich für die Lösung nichtlinearer Optimierungsprobleme – insbesondere, wenn über Nebenbedingungen komplexe Modelle eingebunden sind.

Mithilfe von **Metaheuristiken** können Suchräume mit unterschiedlichen Strategien erkundet werden. Heuristiken werden genutzt, um möglichst gute, angenäherte Lösungen für Optimierungsprobleme zu finden, die zu komplex sind, um sie exakt zu lösen. Dabei wird eine Metaheuristik oft als übergeordnete Strategie angesehen, mit deren Hilfe „untergeordnete" Heuristiken angeleitet werden, um passende Lösungen zu finden. Ein Beispielalgorithmus für eine Metaheuristik ist Simulated Annealing (Bianchi et al. 2008; Voß 2001).

Neuronale Netze (engl. Neural Networks) bestehen aus mehreren Schichten (Layer), die wiederum aus einzelnen Einheiten (Units) mit einer (meist nichtlinearen) Transferfunktion bestehen, die die Summe der Inputs in einen Output transferiert, der über gewichtete Verknüpfungen an den nächsten Layer weitergegeben wird. Ein Netz besteht aus einem Input Layer, mindestens einem Hidden Layer und einem Output Layer. Die Anzahl der Layer wird auch als Tiefe des Netzes bezeichnet. Die Gewichte der Verbindungen und die Parameter der Transferfunktion werden im Laufe des Trainings des Netzes angepasst. Die Komplexität neuronaler Netze – Anzahl der Units und Layer sowie Gewichte der einzelnen Verbindungen und somit Abhängigkeiten zwischen den Units – führen dazu, dass diese Modelle kaum nachvollziehbar sind.

Durch **Regressionsmodelle** wird der Zusammenhang zwischen einer oder mehreren abhängigen und einer oder mehreren unabhängigen Variablen dargestellt. Bei der linearen Regression wird die Annahme getroffen, dass die abhängige Variable kontinuierlich ist und in einem linearen Verhältnis zu den Input-Variablen steht. Bei der logistischen Regression wird die abhängige Variable als binär betrachtet. Logistische Regression ist weit verbreitet und findet beispielsweise auch innerhalb neuronaler Netze Anwendung (Cucchiara 2012; Karlaftis und Vlahogianni 2011). Die beiden vorgestellten Beispiele – lineare und logistische Regressionsmodelle – werden den statistischen bzw. probabilistischen Modellen zugeordnet. Regressionsmodelle sind in der Regel einfach interpretierbar.

Die Idee des **Reinforcement Learning** besteht darin, dass ein „Agent" selbstständig mit seiner Umgebung interagiert, um ein Ziel zu erreichen. Der autonome Agent muss sich in seiner Umgebung zurechtfinden und Interaktionen durchführen, für die er durch seinen „Trainer" eine Belohnung, Bestrafung oder eine neutrale Rückmeldung erhält. Der Agent muss eine Strategie entwickeln, die die Anzahl der Belohnungen maximiert. Er verfügt über eine Menge von Interaktionsmöglichkeiten und kann seine Umgebung normalerweise nicht komplett, sondern in Ausschnitten und eventuell verrauscht wahrnehmen. Auf Grundlage der wahrgenommenen Umgebung wird eine Aktion ausgewählt und durchgeführt, die die Umgebung verändert. Diese Änderung wird wiederum vom Agenten wahrgenommen. Beispiele für Algorithmen des Reinforcement Learning sind Q-Learning oder SARSA (Kaelbling et al. 1996; Mitchell 2010; Harmon und Harmon 1997; Sutton und Barto 2010).

Rekurrente Neuronale Netze (Recurrent Neural Networks, RNNs) stellen eine Art von neuronalen Netzen dar, bei denen Einheiten mit sich selbst, mit Einheiten der gleichen Schicht oder Einheiten aus vorhergehenden Schichten verbunden sein können. So entstehen Kreise in der Konnektivitätsstruktur, was diese neuronalen Netze in dynamische Systeme verwandelt. RNNs können komplexe dynamische Zusammenhänge abbilden und haben ein „Gedächtnis"; sie sind allerdings bei gleicher Anzahl von Einheiten rechenaufwändiger zu trainieren. Anwendung finden RNNs beispielsweise in der Spracherkennung.

Statistische und probabilistische Modelle dienen seit jeher dazu, Messdaten auszuwerten und aus gemessenen und bekannten Wahrscheinlichkeitsverteilungen Schätzungen und Vorhersagen über die modellierten Phänomene abzuleiten. Die Methoden der Statistik ähneln zum Teil denen des Machine Learning und können für ähnliche Ziele eingesetzt werden – z. B. zur Vorhersage oder Klassifizierung. Derartige Ansätze erfordern jedoch, dass die Wahl des Modells auf klaren und nachvollziehbaren Annahmen bzgl. der zugrundeliegenden Daten und Prozesse beruht. Für Datenmengen, deren Struktur sehr komplex oder unbekannt ist, eignen sich daher annahmenfreie Machine Learning-Modelle besser. Ein bekanntes Beispiel sind Regressionsmodelle.

Support Vector Machines (SVMs) wurden ursprünglich entwickelt, um binäre Klassifikationsprobleme zu lösen, können aber auch für weitere Problemstellungen wie Regressionen genutzt werden. Mithilfe von SVMs kann ein nichtlineares Problem mittels Transformation in ein lineares überführt werden. SVMs sind aufgrund guter Performance und effizienten Trainingsverfahren (im Gegensatz zu NNs) häufig die erste Wahl, wenn ein überschaubares ML-Problem gelöst werden soll. SVMs können sehr komplex werden, sodass Nachvollziehbarkeit nicht immer gegeben ist.

Transformer Networks sind eine Art von neuronalen Netzen, die speziell in der Sprachverarbeitung angewendet werden, beispielsweise für Übersetzungsaufgaben, die Generierung von Texten oder Zusammenfassungen. Transformer Networks bestehen aus zwei Komponenten: dem Encoder und dem Decoder. Der Encoder erstellt eine Repräsentation einer Eingabe, anschließend generiert der Decoder Wort für Wort eine entsprechende Ausgabe (Uszkoreit 2017).

Die Basis für **Wissensgraphen und Semantic-Web-Technologien** sind Modelle zur Repräsentation von Wissen, die vom Computer gelesen und „verstanden" werden können. Rein datengetriebene ML-Verfahren haben das Problem, dass sie nur Muster in den Daten erkennen, aber diese Muster nicht in den weiterführenden realweltlichen Kontext stellen können. Wissensgraphen können diesen Kontext herstellen. Diese Modelle werden als Grundlage für die Weiterverwendung der „verlinkten" Informationen in intelligenten Systemen genutzt. Ziel des Semantic Web ist es, detailliertere Informationen schneller zu identifizieren und semantische Interoperabilität im Internet zu steigern. Dies geschieht mithilfe von Ontologien, die wiederum Konzepte spezifizieren, die innerhalb einer Anwendungsdomäne essenziell bzw. wichtig sind. Zur Repräsentation dieser Ontologien werden unterschiedliche Sprachen entwickelt: Beispiele sind das Resource Description Framework (RDF), die Web Ontology Language (OWL) und das Rule Interchange Format (RIF) (Hitzler 2008).

LITERATUR-VERZEICHNIS

LITERATURVERZEICHNIS

acatech (2020): Machine Learning in der Medizintechnik. Analyse und Handlungsempfehlungen. München: acatech (acatech Position). Online verfügbar unter https://www.acatech.de/publikation/machine-learning-in-der-medizintechnik/, zuletzt geprüft am 14.04.2021.

Adadi, Amina; Berrada, Mohammed (2018): Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access (Volume: 6). Online verfügbar unter https://ieeexplore.ieee.org/document/8466590, zuletzt geprüft am 14.04.2021.

Alloghani, Mohamed; Al-Jumeily, Dhiya; Mustafina, Jamila; Hussain, Abir; Aljaaf, Ahmed J. (2020): A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: Michael W. Berry, Azlinah Mohamed und Bee Wah Yap (Hg.): Supervised and Unsupervised Learning for Data Science, Bd. 9. Cham: Springer International Publishing (Unsupervised and Semi-Supervised Learning), S. 3–21. Online verfügbar unter https://link.springer.com/chapter/10.1007%2F978-3-030-22475-2_1, zuletzt geprüft am 14.04.2021.

Arbeitsgruppe Gesundheit, Medizintechnik, Pflege (2019): Lernende Systeme im Gesundheitswesen. Grundlagen, Anwendungsszenarien und Gestaltungsoptionen. Hg. v. Plattform Lernende Systeme, Deutsche Akademie der WIssenschaften. München. Online verfügbar unter https://www.plattform-lernende-systeme.de/publikationen-details/lernende-systeme-im-gesundheitswesen.html, zuletzt geprüft am 20.01.2021.

Arrieta, Alejandro Barredo; Díaz-Rodríguez, Natalia; Ser, Javier Del; Bennetot, Adrien; Tabik, Siham; Barbado, Alberto et al. (2019): Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Online verfügbar unter http://arxiv.org/pdf/1910.10045v2, zuletzt geprüft am 14.04.2021.

Bach, Sebastian; Binder, Alexander; Montavon, Grégoire; Klauschen, Frederick; Müller, Klaus-Robert; Samek, Wojciech: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. In: PloS One (7). Online verfügbar unter https://pubmed.ncbi.nlm.nih.gov/26161953/, zuletzt geprüft am 14.04.2021.

Badr, Will (2019): Auto-Encoder: What Is It? And What Is It Used For? (Part 1). Online verfügbar unter https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726, zuletzt geprüft am 12.03.2021.

Baehrens, David; Schroeter, Timon; Harmeling, Stefan; Kawanabe, Motoaki; Hansen, Katja; Mueller, Klaus-Robert (2009): How to Explain Individual Classification Decisions. Online verfügbar unter http://arxiv.org/pdf/0912.1128v1, zuletzt geprüft am 14.04.2021.

Balakrishnama, S.; Ganapathiraju, Aravind (1998): Linear Discriminant Analysis - A Brief Tutorial. Online verfügbar unter https://www.researchgate.net/publication/240093048_Linear_Discriminant_Analysis-A_Brief_Tutorial, zuletzt geprüft am 14.04.2021.

Barbalau, Antonio; Cosma, Adrian; Ionescu, Radu Tudor; Popescu, Marius (2020): A Generic and Model-Agnostic Exemplar Synthetization Framework for Explainable AI. Online verfügbar unter http://arxiv.org/pdf/2006.03896v3, zuletzt geprüft am 14.04.2021.

BDVA Task Force 7 -Sub-group Healthcare (2020): AI IN HEALTHCARE WHITEPAPER. Hg. v. Big Data Value Association. Online verfügbar unter https://www.bdva.eu/sites/default/files/AI%20in%20Healthcare%20Whitepaper_November%202020_0.pdf, zuletzt geprüft am 20.01.2021.

Bhatt, Umang; Xiang, Alice; Sharma, Shubham; Weller, Adrian; Taly, Ankur; Jia, Yunhan et al. (2019): Explainable Machine Learning in Deployment. Online verfügbar unter http://arxiv.org/pdf/1909.06342v4, zuletzt geprüft am 14.04.2021.

Bianchi, Leonora; Dorigo, Marco; Gambardella, Luca Maria; Gutjahr, Walter J. (2008): A survey on metaheuristics for stochastic combinatorial optimization. In: Natural Computing volume 8, pages 239–287. Online verfügbar unter https://link.springer.com/article/10.1007%2Fs11047-008-9098-4, zuletzt geprüft am 14.04.2021.

Bundesinstitut für Arzneimittel und Medizinprodukte (o. D.): Benannte Stellen. Online verfügbar unter https://www.dimdi.de/dynamic/de/medizinprodukte/institutionen/benannte-stellen/, zuletzt geprüft am 10.02.2021.

Chattopadhyay, Aditya; Sarkar, Anirban; Howlader, Prantik; Balasubramanian, Vineeth N. (2017): Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. Online verfügbar unter http://arxiv.org/pdf/1710.11063v3, zuletzt geprüft am 14.04.2021.

Cortez, Paulo; Embrechts, Mark J. (2011): Opening black box Data Mining models using Sensitivity Analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). Online verfügbar unter https://ieeexplore.ieee.org/document/5949423, zuletzt geprüft am 14.04.2021.

Creswell, Antonia; White, Tom; Dumoulin, Vincent; Arulkumaran, Kai; Sengupta, Biswa; Bharath, Anil A. (2017): Generative Adversarial Networks: An Overview. In: IEEE Signal Process. Mag. (IEEE Signal Processing Magazine) (1). Online verfügbar unter http://arxiv.org/pdf/1710.07035v1, zuletzt geprüft am 14.04.2021.

Cucchiara, Andrew (2012): Applied Logistic Regression. Online verfügbar unter https://www.researchgate.net/publication/261659875_Applied_Logistic_Regression, zuletzt geprüft am 14.04.2021.

Cunningham, Pádraig (2008): Dimension Reduction. In: Matthieu Cord und Pádraig Cunningham (Hg.): Machine Learning Techniques for Multimedia, Bd. 12. Berlin, Heidelberg: Springer Berlin Heidelberg (Cognitive Technologies), S. 91–112. Online verfügbar unter https://link.springer.com/chapter/10.1007%2F978-3-540-75171-7_4, zuletzt geprüft am 14.04.2021.

Danilevsky, Marina; Qian, Kun; Aharonov, Ranit; Katsis, Yannis; Kawas, Ban; Sen, Prithviraj (2020): A Survey of the State of Explainable AI for Natural Language Processing. Online verfügbar unter http://arxiv.org/pdf/2010.00711v1, zuletzt geprüft am 14.04.2021.

Datta, Anupam; Sen, Shayak; Zick, Yair (2016): Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems (IEEE Symposium on Security and Privacy 2016). Online verfügbar unter https://ieeexplore.ieee.org/document/7546525, zuletzt geprüft am 14.04.2021.

Doshi-Velez, Finale; Kim, Been (2017): Towards A Rigorous Science of Interpretable Machine Learning. Online verfügbar unter http://arxiv.org/pdf/1702.08608v2, zuletzt geprüft am 14.04.2021.

eco - Verband der Internetwirtschaft e.V. (Hg.) (2019): Künstliche Intelligenz. Potenzial und nachhaltige Veränderung der Wirtschaft in Deutschland. Online verfügbar unter https://www.eco.de/kuenstliche-intelligenz-potenzial-und-nachhaltige-veraenderung-der-wirtschaft-in-deutschland/#download, zuletzt geprüft am 18.03.2021.

Erhan, Dumitru; Bengio, Y.; Courville, Aaron; Vincent, Pascal (2009): Visualizing Higher-Layer Features of a Deep Network. Université de Montréal. Online verfügbar unter https://www.researchgate.net/profile/Aaron_Courville/publication/265022827_Visualizing_Higher-Layer_Features_of_a_Deep_Network/links/53ff82b00cf24c81027da530.pdf, zuletzt geprüft am 14.04.2021.

Europäische Kommission (Hg.) (2018): Commission Staff Working Document - Evaluation of the Machinery Directive. Brüssel. Online verfügbar unter https://ec.europa.eu/transparency/regdoc/rep/10102/2018/EN/SWD-2018-160-F1-EN-MAIN-PART-1.PDF, zuletzt geprüft am 01.03.2021.

European Commission (Hg.) (2020): White Paper: On Artificial Intelligence. A European approach to excellence and trust. Online verfügbar unter https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, zuletzt geprüft am 24.03.2021.

Gilpin, Leilani H.; Bau, David; Yuan, Ben Z.; Bajwa, Ayesha; Specter, Michael; Kagal, Lalana (2018): Explaining Explanations: An Overview of Interpretability of Machine Learning. Online verfügbar unter http://arxiv.org/pdf/1806.00069v3, zuletzt geprüft am 14.04.2021.

Gondal, Waleed M.; Köhler, Jan M.; Grzeszick, René; Fink, Gernot A.; Hirsch, Michael (2017): Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. Online verfügbar unter http://arxiv.org/pdf/1706.09634v1, zuletzt geprüft am 14.04.2021.

Google (Hg.) (2020): AI Explanations Whitepaper. Online verfügbar unter https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf, zuletzt geprüft am 01.02.2021.

Goos, Gerhard; Hartmanis, Juris; Leeuwen, Jan (2000): Multiple Classifier Systems. First International Workshop, MCS 2000 Cagliari, Italy, June 21-23, 2000 Proceedings. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science, 1857). Online verfügbar unter http://dx.doi.org/10.1007/3-540-45014-9, zuletzt geprüft am 14.04.2021.

Goyal, Yash; Wu, Ziyan; Ernst, Jan; Batra, Dhruv; Parikh, Devi; Lee, Stefan (2019): Counterfactual Visual Explanations. Online verfügbar unter http://arxiv.org/pdf/1904.07451v2, zuletzt geprüft am 14.04.2021.

Harmon, Mance E.; Harmon, Stephanie S. (1997): Reinforcement Learning: A Tutorial. Online verfügbar unter https://www.semanticscholar.org/paper/Reinforcement-Learning%3A-A-Tutorial.-Harmon-Harmon/e7ac195959e2ce078902b000fc16ef2096fcec10, zuletzt geprüft am 14.04.2021.

High-Level Expert Group on AI (Hg.) (2019): Ethics guidelines for trustworthy AI. Online verfügbar unter https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, zuletzt geprüft am 25.02.2021.

Hitzler, Pascal (2008): Semantic Web. Grundlagen. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (eXamen.press). Online verfügbar unter http://dx.doi.org/10.1007/978-3-540-33994-6, zuletzt geprüft am 14.04.2021.

Holzinger, Andreas (2018): Explainable AI (ex-AI). Hg. v. Informatik Spektrum. Online verfügbar unter https://www.springerprofessional.de/explainable-ai-ex-ai/15586620, zuletzt geprüft am 19.02.2021.

Holzinger, Andreas; Biemann, Chris; Pattichis, Constantinos S.; Kell, Douglas B. (2017): What do we need to build explainable AI systems for the medical domain? Online verfügbar unter http://arxiv.org/pdf/1712.09923v1, zuletzt geprüft am 14.04.2021.

Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland (Hg.) (2020): Fragenkatalog „Künstliche Intelligenz bei Medizinprodukten“. Online verfügbar unter http://www.ig-nb.de/dok_view?oid=795601, zuletzt geprüft am 20.01.2021.

Kaelbling, L. P.; Littman, M. L.; Moore, A. W. (1996): Reinforcement Learning: A Survey. In: jair (Journal of Artificial Intelligence Research). Online verfügbar unter https://www.jair.org/index.php/jair/article/view/10166, zuletzt geprüft am 14.04.2021.

Karlaftis, M. G.; Vlahogianni, E. I. (2011): Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. In: Transportation Research Part C: Emerging Technologies (3). Online verfügbar unter https://www.sciencedirect.com/science/article/abs/pii/S0968090X10001610?via%3Dihub, zuletzt geprüft am 14.04.2021.

Karst, Michael (1992): Methodische Entwicklung von Expertensystemen. Wiesbaden, s.l.: Deutscher Universitätsverlag (DUV Wirtschaftswissenschaft). Online verfügbar unter http://dx.doi.org/10.1007/978-3-663-14584-4, zuletzt geprüft am 14.04.2021.

Kawaguchi, Kenji (2016): Deep Learning without Poor Local Minima. Online verfügbar unter http://arxiv.org/pdf/1605.07110v3, zuletzt geprüft am 14.04.2021.

Li, Oscar; Liu, Hao; Chen, Chaofan; Rudin, Cynthia (2017): Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. Online verfügbar unter http://arxiv.org/pdf/1710.04806v2, zuletzt geprüft am 14.04.2021.

Lipton, Zachary C. (2016): The Mythos of Model Interpretability. Online verfügbar unter http://arxiv.org/pdf/1606.03490v3, zuletzt geprüft am 14.04.2021.

Lucas, Peter J.; Van der Gaag, Linda C. (1991): Principles of expert systems. Wokingham: Addison-Wesley. Online verfügbar unter https://www.researchgate.net/publication/224818110_Principles_of_Expert_Systems, zuletzt geprüft am 14.04.2021.

Lundberg, Scott; Lee, Su-In (2017): A Unified Approach to Interpreting Model Predictions. Online verfügbar unter http://arxiv.org/pdf/1705.07874v2, zuletzt geprüft am 14.04.2021.

Mangalathu, Sujith; Hwang, Seong-Hoon; Jeon, Jong-Su (2020): Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Engineering Structures. Online verfügbar unter https://www.sciencedirect.com/science/article/abs/pii/S0141029620307513?via%3Dihub, zuletzt geprüft am 14.04.2021.

Mazzanti, Samuele (2020): SHAP values explained exactly how you wished someone explained to you. Online verfügbar unter https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30, zuletzt geprüft am 16.02.2021.

Mitchell, Tom M. (2010): Machine learning. International ed., [Reprint.]. New York, NY: McGraw-Hill (McGraw-Hill series in computer science).

Molnar, Christoph (2019): Interpretable machine learning. A guide for making black box models explainable. 1st edition. Online verfügbar unter https://christophm.github.io/interpretable-ml-book/, zuletzt geprüft am 14.04.2021.

Montavon, Grégoire; Binder, Alexander; Lapuschkin, Sebastian; Samek, Wojciech; Müller, Klaus-Robert (2019): Layer-Wise Relevance Propagation: An Overview. In: Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen und Klaus-Robert Müller (Hg.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer International Publishing, S. 193–209. Online verfügbar unter https://doi.org/10.1007/978-3-030-28954-6_10, zuletzt geprüft am 14.04.2021.

Nagpal, Kunal; Foote, Davis; Liu, Yun; Po-Hsuan; Chen; Wulczyn, Ellery et al. (2018): Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer. In: npj Digit. Med (1). Online verfügbar unter http://arxiv.org/pdf/1811.06497v1, zuletzt geprüft am 14.04.2021.

Nambiar, Ananthan; Liu, Simon; Hopkins, Mark; Heflin, Maeve; Maslov, Sergei; Ritz, Anna (2020): Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks (13). Online verfügbar unter https://www.biorxiv.org/content/10.1101/2020.06.15.153643v1, zuletzt geprüft am 14.04.2021.

Nguyen, Daria (2020): Explain Your ML Model Predictions With Local Interpretable Model-Agnostic Explanations (LIME). Hg. v. Publicis Sapient Engineering. Online verfügbar unter https://medium.com/xebia-france/explain-your-ml-model-predictions-with-local-interpretable-model-agnostic-explanations-lime-82343c5689db, zuletzt geprüft am 15.02.2021.

Oladipupo, Taiwo (2010): Types of Machine Learning Algorithms. In: Yagang Zhang (Hg.): New Advances in Machine Learning: InTech. Online verfügbar unter https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms, zuletzt geprüft am 14.04.2021.

Otter, Daniel W.; Medina, Julian R.; Kalita, Jugal K. (2018): A Survey of the Usages of Deep Learning in Natural Language Processing. Online verfügbar unter http://arxiv.org/pdf/1807.10854v3, zuletzt geprüft am 14.04.2021.

Pocevičiūtė, Milda; Eilertsen, Gabriel; Lundström, Claes (2020): Survey of XAI in digital pathology. Online verfügbar unter http://arxiv.org/pdf/2008.06353v1, zuletzt geprüft am 14.04.2021.

Puppe, Frank (1988): Einführung in Expertensysteme. Berlin, Heidelberg: Springer (Studienreihe Informatik). Online verfügbar unter http://dx.doi.org/10.1007/978-3-662-00706-8, zuletzt geprüft am 14.04.2021.

Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos (2016): „Why Should I Trust You?": Explaining the Predictions of Any Classifier. Online verfügbar unter http://arxiv.org/pdf/1602.04938v3, zuletzt geprüft am 14.04.2021.

Rudin, Cynthia (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. In: Nat Mach Intell. Online verfügbar unter http://arxiv.org/pdf/1811.10154v3, zuletzt geprüft am 18.02.2021.

Salehi, Mohammadreza (2020): A Review of Different Interpretation Methods in Deep Learning (Part 2: Pixel-wise Decomposition, DeepLIFT, LIME). Online verfügbar unter https://medium.com/@mrsalehi/a-review-of-different-interpretation-methods-in-deep-learning-part-2-input-gradient-layerwise-e077609b6377, zuletzt geprüft am 28.02.2021.

Samek, Wojciech; Montavon, Grégoire; Vedaldi, Andrea (2019): Explainable AI. Interpreting, explaining and visualizing deep learning (Lecture notes in computer series Lecture notes in artificial intelligence). Online verfügbar unter https://link.springer.com/book/10.1007/978-3-030-28954-6, zuletzt geprüft am 14.04.2021.

Schaaf, Nina; Wiedenroth, Saskia Johanna; Wagner, Philipp (2021): Erklärbare KI in der Praxis: Anwendungsorientierte Evaluation von xAI-Verfahren. Hg. v. Marco Huber und Werner Kraus. Online verfügbar unter https://www.ki-fortschrittszentrum.de/de/studien/erklaerbare-ki-in-der-praxis.html, zuletzt geprüft am 14.04.2021.

Selvaraju, Ramprasaath R.; Cogswell, Michael; Das, Abhishek; Vedantam, Ramakrishna; Parikh, Devi; Batra, Dhruv (2019): Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Int J Comput Vis (International Journal of Computer Vision) (2). Online verfügbar unter http://arxiv.org/pdf/1610.02391v4, zuletzt geprüft am 14.04.2021.

Shiebler, Dan (2017): Understanding Neural Networks with Layerwise Relevance Propagation and Deep Taylor Series. Online verfügbar unter http://danshiebler.com/2017-04-16-deep-taylor-lrp/, zuletzt geprüft am 28.02.2021.

Shrikumar, Avanti; Greenside, Peyton; Kundaje, Anshul (2017): Learning Important Features Through Propagating Activation Differences. In: PMLR 70:3145-3153. Online verfügbar unter http://arxiv.org/pdf/1704.02685v2.

Shrikumar, Avanti; Greenside, Peyton; Shcherbina, Anna; Kundaje, Anshul (2016): Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. Online verfügbar unter http://arxiv.org/pdf/1605.01713v3, zuletzt geprüft am 14.04.2021.

Sokol, Kacper; Flach, Peter (2019): Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. Conference on Fairness, Accountability, and Transparency (FAT* ,20), January 27-30, 2020, Barcelona, Spain. Online verfügbar unter http://arxiv.org/pdf/1912.05100v1, zuletzt geprüft am 14.04.2021.

Spreckelsen, Cord; Spitzer, Klaus (2009): Wissensbasen und Expertensysteme in der Medizin. KI-Ansätze zwischen klinischer Entscheidungsunterstützung und medizinischem Wissensmanagement. 1. Aufl. Wiesbaden: Vieweg+Teubner Verlag / GWV Fachverlage GmbH Wiesbaden (Medizinische Informatik). Online verfügbar unter http://dx.doi.org/10.1007/978-3-8348-9294-2, zuletzt geprüft am 14.04.2021.

Springenberg, Jost Tobias; Dosovitskiy, Alexey; Brox, Thomas; Riedmiller, Martin (2014): Striving for Simplicity: The All Convolutional Net. Online verfügbar unter http://arxiv.org/pdf/1412.6806v3, zuletzt geprüft am 14.04.2021.

Stepin, Ilia; Alonso, Jose M.; Catala, Alejandro; Pereira-Farina, Martin (2021): A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. In: IEEE Access. Online verfügbar unter https://ieeexplore.ieee.org/document/9321372, zuletzt geprüft am 14.04.2021.

Strong Medicine (2018): An Approach to Chest Pain, 29.01.2018. Online verfügbar unter https://www.youtube.com/watch?v=-i67erljNYI, zuletzt geprüft am 19.02.2021.

Sundararajan, Mukund; Taly, Ankur; Yan, Qiqi (2017): Axiomatic Attribution for Deep Networks. Online verfügbar unter http://arxiv.org/pdf/1703.01365v2, zuletzt geprüft am 14.04.2021.

Sutton, Richard S.; Barto, Andrew G. (2010): Reinforcement learning. An introduction. [Nachdr.]. Cambridge, Mass.: MIT Press (A Bradford book). Online verfügbar unter https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf, zuletzt geprüft am 14.04.2021.

Tjoa, Erico; Guan, Cuntai (2020): A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. In: IEEE Trans. Neural Netw. Learning Syst. (IEEE Transactions on Neural Networks and Learning Systems). Online verfügbar unter http://arxiv.org/pdf/1907.07374v5, zuletzt geprüft am 14.04.2021.

Touretzky, David S. (Hg.) (1996): Advances in neural information processing systems 8. Proceedings of the 1995 conference ; [papers presented at the Ninth Annual Conference on Neural Information Processing Systems (NIPS), held in Denver, Colorado from Nov. 27 to Nov. 30, 1995. Conference on Neural Information Processing Systems; Annual Conference on Neural Information Processing Systems; NIPS. Cambridge, Mass.: MIT Press. Online verfügbar unter https://mitpress.mit.edu/books/advances-neural-information-processing-systems-8, zuletzt geprüft am 14.04.2021.

U.S. Food & Drug Administration (Hg.) (2020): Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). Discussion Paper and Request for Feedback. Online verfügbar unter https://www.fda.gov/media/122535/download, zuletzt geprüft am 20.01.2021.

U.S. Food & Drug Administration (2021): Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Online verfügbar unter https://www.fda.gov/media/145022/download, zuletzt geprüft am 10.02.2021.

Uszkoreit, Jakob (2017): Transformer: A Novel Neural Network Architecture for Language Understanding. Online verfügbar unter https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html, zuletzt geprüft am 12.03.2021.

van Aken, Betty; Papaioannou, Jens-Michalis; Mayrdorfer, Manuel; Budde, Klemens; Gers, Felix A.; Löser, Alexander (2021): Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. Online verfügbar unter http://arxiv.org/pdf/2102.04110v1, zuletzt geprüft am 14.04.2021.

VERBAND DER CHEMISCHEN INDUSTRIE e.V. (Hg.) (2012): Leitfaden zur Anwendung der Maschinenrichtlinie in verfahrenstechnischen Anlagen. Online verfügbar unter https://www.vci.de/langfassungen-pdf/leitfaden-zur-anwendung-der-maschinenrichtlinie-in-verfahrenstechnischen-anlagen.pdf, zuletzt geprüft am 01.03.2021.

Voß, Stefan (2001): Meta-heuristics: The State of the Art. In: G. Goos, J. Hartmanis, J. van Leeuwen und Alexander Nareyek (Hg.): Local Search for Planning and Scheduling, Bd. 2148. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Computer Science), S. 1–23. Online verfügbar unter https://link.springer.com/chapter/10.1007%2F3-540-45612-0_1, zuletzt geprüft am 14.04.2021.

Wachter, Sandra; Mittelstadt, Brent; Russell, Chris (2017): Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. In: Harvard Journal of Law & Technology. Online verfügbar unter http://arxiv.org/pdf/1711.00399v3, zuletzt geprüft am 14.04.2021.

Wagner, Marc (2000): Bayes-Netze. Eine Einführung. Online verfügbar unter https://itp.uni-frankfurt.de/~mwagner/talks/Bayes.pdf, zuletzt geprüft am 14.04.2021.

Wahlster, Wolfgang; Winterhalter, Christoph (Hg.) (2020): Deutsche Normungsroadmap Künstliche Intelligenz. DIN/DKE. Online verfügbar unter https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf, zuletzt geprüft am 01.03.2021.

Wolf, Thomas; Debut, Lysandre; Sanh, Victor; Chaumond, Julien; Delangue, Clement; Moi, Anthony et al. (2019): HuggingFace's Transformers: State-of-the-art Natural Language Processing. Online verfügbar unter http://arxiv.org/pdf/1910.03771v5, zuletzt geprüft am 14.04.2021.

Ye, Andre (2020): Every ML Engineer Needs to Know Neural Network Interpretability. Online verfügbar unter https://medium.com/analytics-vidhya/every-ml-engineer-needs-to-know-neural-network-interpretability-afea2ac0824e, zuletzt geprüft am 28.02.2021.

Zeiler, Matthew D.; Fergus, Rob (2013): Visualizing and Understanding Convolutional Networks. Online verfügbar unter http://arxiv.org/pdf/1311.2901v3, zuletzt geprüft am 14.04.2021.

Zeiler, Matthew D.; Taylor, Graham W.; Fergus, Rob (2011): Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 International Conference on Computer 06.11.2011 - 13.11.2011. Online verfügbar unter https://ieeexplore.ieee.org/document/6126474, zuletzt geprüft am 14.04.2021.

Zhou, Bolei; Khosla, Aditya; Lapedriza, Agata; Oliva, Aude; Torralba, Antonio (2015): Learning Deep Features for Discriminative Localization. Online verfügbar unter http://arxiv.org/pdf/1512.04150v1, zuletzt geprüft am 14.04.2021.