Metrologie für KI in der Medizin

Hintergründe, Strategie und Umsetzungsempfehlungen

Metrologie ist ein anerkannter Vertrauensanker und wesentlicher Bestandteil der Qualitätsinfrastruktur. Dazu gehört die Charakterisierung der Messtechnik und Messmethoden, die Bewertung der Qualität von Messdaten und die Entwicklung neuer Messverfahren. Das gilt insbesondere auch für Methoden, Verfahren und Messgeräte in der Medizin. Wie in allen Bereichen wird auch die Medizin in zunehmendem Maße von der fortschreitenden Digitalisierung erfasst und stellt damit neue Herausforderungen an die PTB. Grundsätzlich ist die gesetzliche Beauftragung der PTB im Rahmen des EinhZG (§ 6 Abs. 3), des MessEG (§ 45) und des Medizinproduktegesetzes (§ 32 Abs. 2) sehr Technologie-offen formuliert. Insofern ist die PTB selbst auch kontinuierlich aufgefordert, ihre eigene Rolle im Sinne der gesetzlichen Beauftragung angesichts technologischer Entwicklungen zu bewerten und zu hinterfragen. Gleichzeitig ist bei neuen technologischen Entwicklungen immer davon auszugehen, dass ein Erwartungsdruck gegenüber der PTB besteht, ihrem gesetzlichen Auftrag in kompetenter Weise gerecht zu werden.

Eine Herausforderung, die schon seit langem besteht, ist die Verwendung von Software als (teilweise eigenständiges) Produkt in der Medizin. Die EU Medical Device Regulation (MDR) sieht bereits die Evaluation von reinen Softwareprodukten im Gesundheitssektor vor (Medical Device Software – MDSW). Ist diese Software essenzieller Bestandteil eines Medizinproduktes, so ist eine gemeinsame Evaluation von Hard- und Software notwendig. In jedem Fall sind klinische Studien für den Nachweis eines medizinischen Mehrwertes für die Zulassung relevant.

Im Zuge der Digitalisierung spielt die Künstliche Intelligenz (KI) zunehmend die Rolle eines Katalysators und beschleunigt die Entwicklung digitaler Produkte und Dienstleistungen enorm. Mit wachsendem Einsatzbereich von KI steigt die Notwendigkeit für klare Regeln und eine Berücksichtigung in der Qualitätsinfrastruktur. Das schließt ausdrücklich auch den Gesundheitssektor mit ein, in welchem die Komplexität der Zusammenhänge, der technische Fortschritt in der Messtechnik, der hohe Benefit für die Patienten und die enormen wirtschaftlichen Potenziale zu einem rasanten Anstieg an KI-Anwendungen führen wird.

In Anlehnung an die KI-Strategie der Bundesregierung wird hier unter *Künstlicher Intelligenz* ein Algorithmensystem zur Lösung konkreter Anwendungsprobleme auf Basis von Methoden aus der Mathematik und Informatik verstanden, wobei die entwickelten Systeme zur Selbstoptimierung fähig sind.

Die Enquete-Kommission der Bundesregierung sieht in der KI die nächste Stufe einer durch technologischen Fortschritt getriebenen Digitalisierung. Ein wesentliches Element ist dabei die Art, wie diese Algorithmen entwickelt werden. Ein klassischer Algorithmus setzt in der Regel ein vorher beschriebenes Verfahren in Software um. Das Verfahren basiert dabei auf mathematischen, statistischen oder anderen Annahmen, Theorien und Regeln. Im Gegensatz dazu wird der Algorithmus einer KI-Methode mit Hilfe von Daten trainiert. Diese Algorithmen zeichnen eine hohe Komplexität und einen sehr hochdimensionalen Parameterraum aus. Ein weiteres Merkmal ist die sehr hohe Anpassungsfähigkeit von KI-Methoden. Diese kann jedoch dazu führen, dass auch unerkannte Merkmale der Trainingsdaten ungewollt in den Algorithmus einfließen. Daher ist, im Gegensatz zu anderer Software, eine Prüfung des Algorithmus auf Basis des Quellcodes allein kaum durchführbar.

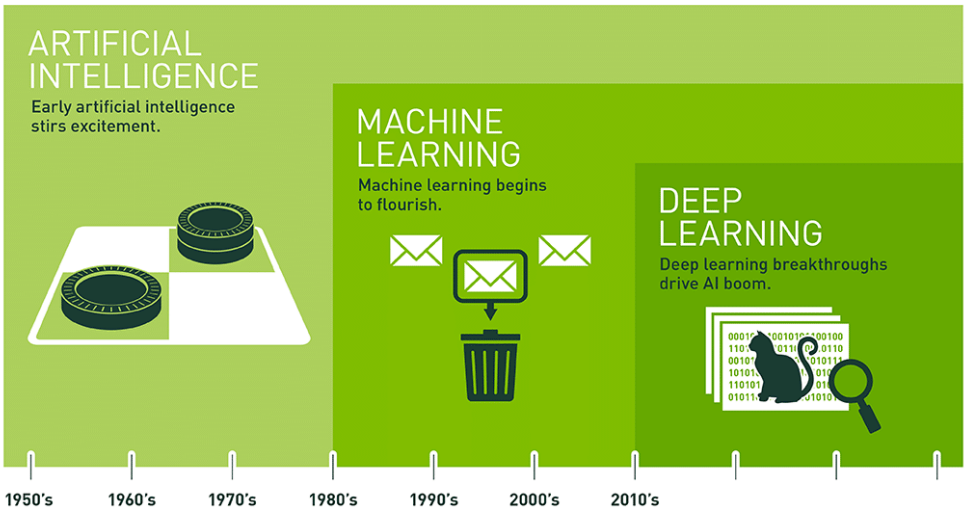

Abbildung 1 Quelle: https://blogs.nvidia.com/wp-content/uploads/2016/07/Deep_Learning_Icons_R5_PNG.jpg.png

Insbesondere für die Anwendung von KI-Methoden im Bereich der Medizin fehlt bisher ein allgemein anerkannter Vertrauensanker in der Qualitätsinfrastruktur. Dabei ist besonders dort Vertrauen in die KI-Methoden unabdingbar für ihre Akzeptanz. Allgemein verfolgt sowohl die Bundesregierung als auch die EU-Kommission das Leitbild einer auf den Menschen ausgerichteten KI. Das bedeutet, dass die KI dem Menschen und der Gesellschaft nutzen und dabei ein selbstbestimmtes Handeln stärken sollte. Dieses Leitbild wird auch als „KI mit europäischer Prägung“ bezeichnet in Abgrenzung zur wirtschaftlich ausgerichteten KI-Entwicklung, bspw. in den USA oder in China.

Viele Forschungseinrichtungen, Standardisierungsstellen und Unternehmen entwickeln Methoden, Leitfäden und Standards zu KI. Dieses Dokument adressiert die Fragestellung, wie in diesem Umfeld die Rolle der Metrologie und speziell der PTB aussehen sollte.

(1) Die PTB baut gezielt ihre Kompetenzen im Bereich KI aus und etabliert Kooperationen mit externen KI-Experten, um auch in Zukunft ihrem gesetzlichen Auftrag gerecht werden zu können.
(2) Aus der aktuellen gesetzlichen Beauftragung lässt sich die Notwendigkeit ableiten, dass sich die PTB intensiv mit dem Thema KI auseinandersetzen muss. Mittelfristig wird darauf hingearbeitet, die Beauftragung und Zuständigkeit der PTB im Themenfeld KI rechtlich schärfer zu verankern und bei der Neufassung eines Rechtsrahmens für KI die PTB von Beginn an zu integrieren.
(3) Arbeiten im Bereich KI konzentrieren sich zunächst auf konkrete Anwendungsfälle. Parallel werden grundlegende theoretische Arbeiten vorangetrieben, um langfristig eine allgemeine KI-Kompetenz an der PTB über spezifische Anwendungen hinaus zu etablieren.

1. Hintergründe und Abgrenzung zu anderen Akteuren

Es gibt aktuell kaum einen Bereich, in dem das Thema KI nicht diskutiert wird. Etliche Publikationen zu neuen Methoden, Modellen und Anwendungsbeispielen werden in zunehmender Häufigkeit veröffentlicht. Zahlreiche Fraunhofer Institute, das DFKI und das DLR bauen ihre Aktivitäten im Bereich der KI-Forschung massiv aus. Insgesamt 100 KI-Professuren sollen in den kommenden Jahren eingerichtet werden. In der Normung und Standardisierung werden nicht nur bei ISO und IEC neue

Arbeitsgruppen eingerichtet. Auch das DIN ist mit der „Normungsroadmap KI“ [17] und der bereits teilweise veröffentlichten DIN SPEC 92001 sehr aktiv. Teilweise sind auch Gremien mit direktem Bezug zu Medizin und Gesundheit entstanden; bspw. die ITU/WHO FG-AI4H zu KI-Methoden im Gesundheitssektor. Die PTB muss in diesem Umfeld ihre Rolle und ihren Platz finden.

Standardisierung und Regulierung von KI

Die grundsätzlichen Anforderungen und Terminologien für die Bewertung von KI-Methoden sind bereits in Grundzügen erarbeitet. So definiert die DIN SPEC 92001-1 [1] drei wesentliche Anforderungen an die Qualität von KI:

- *Funktionalität und Performance* als Ausdruck der Fähigkeit der KI, die gestellte Aufgabe unter festgelegten Bedingungen zu erledigen
- *Robustheit* als Fähigkeit der KI mit fehlerhaften, verrauschten, unbekannten oder schädlichen Eingangsdaten umgehen zu können
- *Erklärbarkeit* als Ausdruck für die Möglichkeit, die Gründe für das Ergebnis einer KI-Methode verstehen und nachvollziehen zu können

Die DIN SPEC 92001-2 [7] konkretisiert den Begriff der Robustheit und unterscheidet zwischen *adversarial robustness* (AR) und *corruption robustness* (CR). Erstere bezeichnet die Robustheit ggü. schadhaften (adversarial) Änderungen an den Eingangsdaten, letzteres steht für Robustheit ggü. Rauschen oder Veränderung der statistischen Eigenschaften der Eingangsdaten. Für die Entwicklung von robusten KI-Methoden empfiehlt [7] einen Risikoanalyse-basierten Ansatz. Dazu werden auch Methoden zum gezielten Testen von KI-Methoden genannt (Fast Gradient Sign Method; Projected Gradient Descent). Allgemein wird ein Szenario-basiertes Testen empfohlen, welches den späteren Einsatzzweck und dessen Eigenschaften mit einbezieht. Als wichtig erachtet [7], dass die Risiko-Bewertung einer KI-Methode eigentlich kontinuierlich durchgeführt werden muss.

Auch die US-amerikanische *Food and Drug Administration* (FDA) geht in einem aktuellen Diskussionspapier [3] von der Notwendigkeit eines „Total Product Lifecycle Regulatory Approach“ (TPLC) für KI-Anwendungen aus. Im Zuge der Zertifizierung eines KI-basierten Medizinprodukts soll dabei insbesondere das Qualitätsmanagement des Unternehmens begutachtet werden hinsichtlich

- Qualitätssicherung in der Softwareentwicklung
- Tests und Performance-Monitoring der Produkte

Dabei stellt die FDA folgende grundsätzliche Prinzipien auf

- Etablierung anerkannter „good machine learning (ML) practices“
- Berücksichtigung des Produktlebenszyklus bei der Zulassung KI-basierter Medizinprodukte
- Erwartung, dass die Hersteller einen Risiko-basierten Ansatz für das Monitoring ihrer KI-basierten Medizinprodukte für den gesamten Produktlebenszyklus realisieren
- Transparente Aussagen für Kunden und Prüfer zu tatsächlicher Leistung und Verhalten von KI-basierten Medizinprodukten durch die Hersteller.

Dazu gehört für die FDA auch das Dokumentieren des geplanten Einsatzbereiches (Software as a Medical Device (SaMD) Pre-Specification – SPS) sowie der (Weiter-)Entwicklung in einem „Algorithm Change Protokoll“ (ACP): Data management, Re-training, Performance evaluation, Update procedures. SPS und ACP sind dann wesentliche Punkte bei der Zulassung neuer Produkte.

Ein wichtiger Teil des Qualitätssicherungs- und -Managementsysteme ist laut Aussage der FDA [3] die Wahl der Trainings- und Testdaten sowie die Auswahl von Anwenderdaten für das Re-Training. Beim Datenmanagement sieht die FDA daher

- Protokolle zur Datenerhebung
- Qualitätssicherungssysteme für die Daten
- Bestimmung eines Referenzstandards
- Auditierung und Sicherung von Test- und Trainingsdaten

durch die Hersteller vor. Auch der Fragenkatalog der deutschen „Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland“ (IG-NB) für die Zulassung von KI bei Medizinprodukten widmet viele Fragen der Auswahl und Beurteilung der verwendeten Daten [19]. Gleichzeitig fehlen gerade zu den wichtigen Fragen für die Bewertung der KI und der zugrundeliegenden Daten in [19] entsprechende Normen und Standards als Grundlage für die Prüfung.

Zertifizierung von KI

Im Whitepaper des Fraunhofer IAIS [4] wird eine durch akkreditierte Prüfer operativ durchführbare Zertifizierung für KI-Anwendungen diskutiert. Demzufolge soll ein Zertifikat für KI

- Einen gewissen Qualitätsstandard bescheinigen,
- Dabei helfen, KI-Anwendungen überprüfbar rechtskonform zu gestalten und
- KI-Anwendungen vergleichbar machen.

Dabei stellt das IAIS fest, dass für die Bewertung der Verlässlichkeit Domänenwissen und mathematisch-statistische Expertise notwendig sind [4].

EUROLAB macht eine deutliche Unterscheidung zwischen indirekter (indirect conformity assessment - ICA) und direkter (direct conformity assessment - DCA) Anwendung von KI-Methoden [5]. Bei ICA dient die KI-Methode als Unterstützung für die Entscheidungsfindung. Hier wird das Beispiel einer KI-Methode für die Auswertung einer Röntgen-Messung verwendet: Die KI-Methode ermittelt aus den Messdaten eine qualitative Aussage, z. B. über den Gesundheitszustand des Patienten. In diesem Fall wäre bei einer Akkreditierung keine Bewertung der KI-Methode an sich notwendig, sondern es würde die Kompetenz des Personals der zu akkreditierenden Stelle festgestellt werden müssen mit der Methode umzugehen. Das entspräche dem bereits heute praktiziertem Vorgehen für beliebige andere nichtlineare numerische Verfahren. Wenn die KI-Methode jedoch das Ergebnis als Teil der Messung präsentiert (z. B. als Overlay) und damit den Anschein eines „echten Ergebnisses“ erweckt, ist die KI-Methode selbst als „Autorität“ zu verstehen und in der Akkreditierung mit zu beachten. EUROLAB empfiehlt derzeit, DCA nur in sehr unkritischen Bereichen einzusetzen (z. B. Bewertung von Musikqualität), bis die Methoden ausgereifter sind. Im Positionspapier von EUROLAB [5] wird auch für KI-Methoden eine Art Kalibrierung wie für gängige Messmittel gefordert. Diese sollte die Verlässlichkeit der KI-Methode bewertbar machen, indem die Fähigkeit der Methode erfasst wird, das Ergebnis eines „Standards“ zu reproduzieren.

Im aktuellen Whitepaper „Zertifizierung von KI-Systemen“ [18] der Plattform Lernende Systeme heißt es:

> *„Bevor eine gelungene Zertifizierung von KI-Systemen etabliert werden kann, sind daher noch offene Fragen zu klären. Diese betreffen den Gegenstand der Zertifizierung, die Prüfkriterien, den Zeitpunkt und die Notwendigkeit der Zertifizierung, den Detailgrad der Zertifizierung sowie den Umgang mit weiterlernenden Systemen.“*

Außerdem verweist [18] auch auf die AI *High Level Expert Group* der EU-Kommission (COM), welche folgende Kriterien bei der Regulierung von KI empfiehlt

> *„Vorrang menschlichen Handelns und menschlicher Aufsicht, technische Robustheit und Sicherheit, Privatsphäre und Datenqualitätsmanagement, Transparenz, Vielfalt,*

Nichtdiskriminierung und Fairness sowie gesellschaftliches und ökologisches Wohlergehen und Rechenschaftspflicht."

Insbesondere für KI-Anwendungen mit hohem Risiko (wie im Bereich Gesundheit) wird empfohlen, bei der Konformitätsbewertung zu prüfen

- Ob die Trainingsdaten adäquat sind für den geplanten Einsatzzweck;
- Die Ergebnisse bei der Nutzung nicht zu Diskriminierung führen;
- Datenschutz und Privatsphäre beachtet werden;
- Relevante Aufzeichnungen zu Datensätzen, Trainingsmethoden und Programmiermethoden vorliegen.

Damit folgen diese Empfehlungen im Grunde denen der FDA [3], die (ebenso wie die Bundesregierung in ihrer Stellungnahme zum COM Whitepaper) außerdem eine wiederholte Prüfung dieser Kriterien bei lernenden – also sich verändernden – KI-Systemen als notwendig erachtet.

Schlussfolgerungen für die Zuständigkeit der PTB

Die aktuelle Publikation der Plattform Lernende Systeme formuliert sehr klar:

„Für die Konformitätsbewertung sollte auf bestehende nationale Strukturen und Verfahren zurückgegriffen werden. Sofern es keine solche Behörden gibt, sollte es eine Pflicht zum Aufbau einer solchen Behörde oder zum Aufbau von Zuständigkeiten in bestehenden Behörden geben." [18]

Die PTB als wesentlicher Bestandteil der Qualitätsinfrastruktur und hoher Neutralität ist prädestiniert, diese Rolle zu übernehmen. Konkret ist die PTB bereits heute aufgefordert, ihren gesetzlichen Aufgaben auch für Messgeräte mit KI-Anteilen gerecht zu werden, indem sie entsprechende Kompetenzen kontinuierlich aufbaut. In ihrer Stellungnahme zum Weißbuch KI der COM legt die Bundesregierung dabei folgende Empfehlungen vor:

- Für Trainingsdaten von KI-Systemen sollten verbindliche rechtliche Anforderungen in Betracht gezogen werden. Dafür sollten konsistent auch Anforderungen für Test- und Evaluierungsdaten in Betracht gezogen werden.
- Auch rechtliche Anforderungen für Qualitätsparameter und -anforderungen für Trainings-, Test- und Evaluierungsdaten sind erforderlich, damit entsprechende KI-Systeme mit quantitativ ausreichenden und qualitativ hochwertigen Datensätzen entwickelt werden.
- Anforderungen bzgl. Robustheit und Genauigkeit sollten erkennbare und realistische Szenarien abdecken
- Wenn geeignete Verfahren zur Prüfung der KI-Ergebnisse auf Repräsentativität und Ausgewogenheit zur Verfügung stehen, kann auch auf den Zugriff auf die Trainings-/Testdaten verzichtet werden
- Zentral ist die Sicherstellung der Vertraulichkeit, Integrität und Verfügbarkeit des KI-Systems über dessen gesamten Lebenszyklus als solches
- Für KI-Systeme mit hohem Risiko sollte das Durchlaufen eines objektiven Konformitätsbewertungsverfahrens verbindlich vorgeschrieben werden. Dabei besteht eine Notwendigkeit wiederholter Bewertungen von sich weiterentwickelnden lernfähigen KI-Systemen.
- Entscheidend sind die Genauigkeit und Relevanz der Daten. Darüber hinaus ist es notwendig, die Bereitstellung von Referenzdaten, Benchmarktests und die Überprüfung von Algorithmen anhand von qualitätsgesicherten, vertrauenswürdigen Referenzdaten zur Verfügung zu stellen. Grundsätzlich wird die Aussage unterstützt, dass die Datenqualität während der gesamten Nutzungsdauer gewährleistet sein muss. Es ist aber zu beachten, dass dies nur vom

Betreiber geleistet werden kann, der wiederum kein Wirtschaftsakteur im Sinne des Produktsicherheitsrechts ist.
- Für den Fall, dass aufgrund einer Software-Änderung ein neues Produkt entstanden ist, muss dieses Produkt vollumfänglich dem Stand der Technik entsprechen, da ein neues Inverkehrbringen vorliegt. Dies muss bei der Betrachtung einer Software-Änderung mitberücksichtigt werden.

Die Forschungsaktivitäten der PTB im Bereich KI müssen demnach sicherstellen, dass diese Anforderungen und Erwartungen erfüllt werden können. Demzufolge sollte insbesondere weniger die Entwicklung neuer KI-Methoden im Vordergrund stehen, sondern ein Fokus gelegt werden auf die Entwicklung von Bewertungsmethoden. Ein wichtiges Alleinstellungsmerkmal der PTB ist dabei ihr Domänenwissen sowie ihre Neutralität.

Forschungskooperationen

Die PTB wird auch auf lange Sicht nicht mit dem Umfang der Arbeiten und den Möglichkeiten anderer großer Forschungsverbünde konkurrieren können – und das auch nicht müssen. Für einen möglichst effektiven Einsatz der verfügbaren Ressourcen wird die PTB daher gezielte Kooperationen mit Forschungspartnern eingehen.

Eine gute Übersicht über die aktuelle Forschungslandschaft im Bereich KI bietet die „Landkarte KI“ der Plattform Lernende Systeme[1]. Einige prominente Beispiele sind dabei

- Deutsches Forschungszentrum KI (DFKI)
- Fraunhofer Gesellschaft
- Deutsches Zentrum für Luft- und Raumfahrt (DLR)
- Helmholtz Gemeinschaft (insbesondere Helmholtz KI-Kooperationseinheit, s.u.)
- Max-Planck-Institut für Intelligente Systeme
- Mevis Medical Solutions (Bremen)
- Uni-Kliniken (z. B. Charité) und Medizinische Fakultäten, bspw. der Uni Duisburg-Essen und Universitätsmedizin Essen (Institut für Künstliche Intelligenz in der Medizin)
- Einrichtungen im Cyber Valley in Tübingen als Europas größtes Forschungskonsortium im Bereich KI mit Partner aus Wissenschaft und Industrie
- Robert Koch Institut (RKI) (Aufbau eines Zentrums zu Validierung von KI-Algorithmen in der Gesundheitsforschung)

Teilweise orientiert sich die PTB auch für organisatorische und strukturelle Entscheidungen an großen Forschungseinrichtungen. So hat bspw. die Helmholtz Gemeinschaft die Helmholtz-Künstliche Intelligenz-Kooperationseinheit (Helmholtz AI) gegründet. Diese ist eine von fünf Plattformen, die vom Helmholtz-Inkubator für Informations- und Datenwissenschaften initiiert wurde.

„Ihr Hauptziel ist es, durch die Entwicklung und Verbreitung von KI-Methoden in allen Helmholtz-Zentren zu einem Motor für angewandte künstliche Intelligenz (KI) zu werden, indem KI-basierte Analytik mit den einzigartigen Forschungsfragen und Datensätzen von Helmholtz effektiv kombiniert wird“. (Quelle: https://helmholtz.ai*)*

In dieser Plattform werden Wissenschaftler aller Zentren zusammengeführt und fördern so transdisziplinäre Forschung.

[1] https://www.plattform-lernende-systeme.de/ki-landkarte.html

2. Forschungsaspekte zu Metrologie für KI

Die PTB versetzt sich mit ihren Forschungsaktivitäten im Bereich KI in die Lage, ihrem gesetzlichen Auftrag auch in Zukunft gerecht werden zu können. Dazu gehören die durch Normen und Standards gesetzten Vorgaben für Qualitätsmerkmale von KI ebenso wie Anforderungen aus Verordnungen und Gesetzen für die Zertifizierung und Konformitätsbewertung von Qualitätssicherungsmethoden für Trainings- und Testdaten.

Bewertung von KI

Die PTB führt bereits Grundlagenuntersuchungen und Anwendungsstudien für die Ermittlung von Bewertungsverfahren für KI-Methoden durch. Dabei steht die Entwicklung quantitativer Maße für die Bewertung von *Erklärbarkeit*, *Unsicherheit*, *Generalisierbarkeit* und *Robustheit* im Mittelpunkt.

Für die in [1] geforderte Bewertung der Funktionalität und Performance von KI als ein Maß für die Qualität wird die quantitative Bestimmung der *Unsicherheit* der Vorhersagen der KI benötigt. Wesentlich ist, dass das „Maß“, mit dem die Unsicherheit gemessen wird, standardisiert ist, da nur dann Unsicherheiten verschiedener KI- Methoden in ihren Vorhersagen überhaupt verglichen werden können wie in [4] gefordert. Methoden zur quantitativen Ermittlung von Messunsicherheiten spielen eine zentrale Rolle in der Metrologie, wo es mittlerweile mit dem GUM einen weltweit anerkannten Standard gibt [7]. Eine solche Standardisierung fehlt bisher im Bereich der KI, wo es eine Vielzahl unterschiedlicher Ansätze zur Quantifizierung der Unsicherheit gibt [8-10, 16]. Die PTB untersucht derzeit die Eignung aktueller Ansätze zur Quantifizierung der Unsicherheit von KI-Methoden mit dem Ziel, eine Empfehlung für eine mögliche Standardisierung zu erarbeiten. Die Untersuchungen beinhalten grundlegende Untersuchungen sowie Anwendungsbeispiele [11]. Aus Sicht der Metrologie wäre es erstrebenswert, wenn eine Standardisierung der Unsicherheit im Einklang mit den Prinzipien der Unsicherheitsermittlung in der Metrologie stehen würde und so in Anwendungen, bei denen KI Methoden und klassische Verfahren in ähnlicher Weise operieren, auch gleiche Unsicherheiten zugeordnet werden könnten.

Um Vertrauen in die KI-Methoden zu gewährleisten ist es wichtig, deren Verhalten zu verstehen und sicherzustellen, dass diese nicht etwa nur auf spezielle Aspekte der Trainingsdaten reagieren [12], sondern die relevante Information in den Daten verwenden. Ähnlich wie bei der Unsicherheit gibt es auch bei der *Erklärbarkeit* mittlerweile eine Vielzahl an Ansätzen, siehe z. B. [13, 14] und die Referenzen darin. Ein Ziel der PTB in diesem Bereich ist es letztlich auch, für die Quantifizierung der Erklärbarkeit ein standardisiertes Maß festzulegen. In diesem Bereich ist eine enge Kooperation der PTB mit dem HHI geplant, die im Jahr 2021 startet. Die Kooperation ist Teil eines an der PTB durchgeführten Projekts zur Untersuchung von KI-Methoden bei der medizinischen Bildgebung aus Sicht der Metrologie.

Die *Robustheit* und *Generalisierbarkeit* von KI-Methoden gegenüber Eingangsdaten, die von den zum Trainieren der Methode benutzten Daten abweichen, spielt insbesondere in der Medizintechnik eine große Rolle. Von Bedeutung sind hierbei zum Beispiel „out-of-distribution“ Fehler, die dadurch entstehen, dass gewisse Merkmale nicht in den Trainingsdaten abgebildet sind. Eine große Bedeutung kommt auch den sog. „adversarial attacks“ zu, bei denen „gutartige“ Eingangsdaten gezielt geringfügig so geändert werden, dass eine KI-Methode versagt. Um die Bewertung der Robustheit bezüglich dieser Einflussfaktoren quantitativ vergleichbar zu machen, sind mehrere Bewertungskriterien vorgeschlagen worden. Die PTB untersucht diese Kriterien, und hat auf Basis statistischer Ansätze Alternativen entwickelt, die in bisherigen Untersuchungen sehr gute Eigenschaften aufweisen [15, 16].

Referenzdaten und Bewertung von Datenqualität

In allen Quellen zur Bewertung, Zertifizierung und Konformitätsbewertung von KI-Anwendungen oder Produkten mit KI-Anteilen wird die Notwendigkeit von Referenzdaten sowie allgemein anerkannten Kriterien für Datenqualität und Datenhandling genannt. Für die Wahrnehmung ihrer Aufgaben muss die PTB demnach Kompetenzen zu diesen Fragen aufbauen. Dabei ist die bspw. auch in [4] genannte Notwendigkeit von Domänenwissen wichtig bei der Entscheidung für geeignete Forschungsvorhaben. So ist insbesondere die Repräsentativität von Referenzdaten „aus sich selbst" nicht möglich, sondern immer nur kontextbezogen vor dem Hintergrund einer Grundpopulation. Stattdessen könnten statistische Kriterien (z. B. Test auf Gleichheit der Verteilungen) zum Zuge kommen. Hier könnten auch Anleitungen zur Konstruktion der (synthetischen) Referenzdaten als Aufgabe für die PTB hinzukommen. Die Metrologie beschäftigt sich bereits mit der Beurteilung von Daten, aber tut das bisher eher auf dem bottom-up level (GUM-like), basierend auf dem Verständnis der zugrundeliegenden Physik, als top-down über die Eigenschaften der Daten selbst. In einigen Bereichen stellt die PTB bereits auch physikalische/chemische Referenzdaten zur Verfügung. In Zukunft könnte dies weiter ausgebaut werden mit dem Ziel, Referenzdaten gezielt für die Bewertung von KI-Methoden zu entwickeln. Dabei sollte auch die Entwicklung von Methoden für synthetische Daten berücksichtigt werden.

Inzwischen existieren erste Beispiele für die automatische Annotation von Trainingsdaten durch die Kombination verschiedener Modalitäten. So wurde in [18] in einem ersten Schritt ein ML-Verfahren darauf trainiert, Tomographie-Aufnahmen der Retina und co-registrierte Fundus-Aufnahmen zu einer Prädiktion der Retina-Dicke zu kombinieren. Als Ergebnis wurde das trainierte ML-Verfahren dazu verwendet, einen Datenbestand von 120 000 Datensätzen automatisch zu annotieren. Diese dienen dann wiederum als Trainingsdatensatz für ML-Verfahren zur Detektion von durch Diabetes hervorgerufenen Augenschädigungen mit drohender Blindheit. In einer Zulassung solch eines ML-Verfahrens sind dann nicht mehr nur die reinen Rohdaten zu bewerten, sondern auch der gesamte Workflow zur Verwendung dieser Daten. Entsprechend müsste die PTB auch Kompetenzen im Bereich des Datenhandling aufbauen, um bspw. die Anforderungen aus [3] und [18] abbilden zu können.

3. Referenzen

[1] DIN SPEC 92001-1: Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Metamodel

[2] DIN SPEC 92001-2: Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 2: Robustheit

[3] FDA "Proposed Regulatory Framework for Modifications to Artificial lntelligence/Machine Learning [Al/ML] Based Software as a Medical Device [SaMD]"

[4] Fraunhofer IAIS Whitepaper „Vertrauenswürdiger Einsatz von Künstlicher Intelligenz"

[5] EUROLAB „Position paper in response to EC report COM(2020) 65 final" Henriksen & A. Bechmann "Building truths in AI: Making predictive algorithms doable in healthcare

[6] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML. Evaluation of measurement data – Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008.

[7] Gal Y. Uncertainty in deep learning. University of Cambridge, 2016.

[8] Kingma D, Salimans T, Welling M. Variational dropout and the local reparameterization trick. Advances in neural information processing systems pp. 2575-2583, 2015.

[9] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Advances in neural information processing systems pp. 5574-5584, 2017.

[10] Kretz T, Müller K, Schaeffter T, Elster C. Mammography Image Quality Assurance Using Deep Learning. IEEE Transactions on Biomedical Engineering, 2020.

[11] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1), 1-8.
[12] Goodfellow I, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations, 2015.
[13] Muller R, Kornblith S, Hinton G. When does label smoothing help? Advances in neural Information processing systems pp 4694-4703, 2019.
[14] Martin J, Elster C. Inspecting adversarial examples using the Fisher information. Neurocomputing 382: 80-86, 2020.
[15] Martin J, Elster C. Detecting unusual input to neural networks. Applied Intelligence (2020), in press.
[16] DIN Normungsroadmap KI
[17] Whitepaper “Zertifizierung von KI-Systemen“, Plattform Lernende Systeme (2020)
[18] Impulspapier "Zertifizierung von KI-Systeme“ Plattform Lernende Systeme (2020)