Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme

Technologieprogramm des Bundesministeriums für Wirtschaft und Klimaschutz

LEITFADEN FÜR DAS QUALITÄTSMANAGEMENT BEI DER ENTWICKLUNG VON KI-PRODUKTEN UND -SERVICES

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

www.ki-innovationen.de

IMPRESSUM

Leitfaden im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK) im Rahmen der Begleitforschung zum Technologie-programm „Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme“ (KI-Innovationswettbewerb)

Herausgeber

Begleitforschung des Technologieprogramm KI-Innovationswettbewerb
des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK)

Institut für Innovation und Technik (iit) in der VDI/VDE Innovation + Technik GmbH
Dr. Steffen Wischmann
Steinplatz 1
10623 Berlin
wischmann@iit-berlin.de

Autoren

Dr. Nicole Wittenbrink
Dr. Tom Kraus
Dr. Stefanie Demirci
Sebastian Straub

Gestaltung

LHLK Agentur für Kommunikation GmbH
Hauptstraße 28
10827 Berlin
KI-Innovationswettbewerb@lhlk.de

Stand

Dezember 2022

Bilder

Murrstock – stock.adobe.com /Titel

MANAGEMENT SUMMARY

Der vorliegende Leitfaden bietet eine Orientierungshilfe für die in ein KI-Entwicklungsvorhaben involvierten Akteure, um ihre gemeinsame Produktvision auf ihre Umsetzbarkeit zu überprüfen und gegebenenfalls möglichst effizient in die Realität zu überführen. Er soll einen Beitrag dazu leisten, das Qualitätsmanagement bei der Entwicklung von KI-Produkten und -Services zu verbessern und insbesondere den Abbruch von Entwicklungsvorhaben aufgrund nicht erfüllbarer Anforderungen (u. a. durch den Zielmarkt, Nutzende, Gesetzgebung, Normen und Standards) oder einer unzureichenden Wirtschaftlichkeit zu vermeiden.

Die Verbreitung von KI-Produkten und -Services bleibt bisher insgesamt noch hinter den Erwartungen und Prognosen aus Vorjahren zurück (Billerbeck 2022). Hierzu trägt bei, dass viele KI-Entwicklungsvorhaben gegenwärtig nicht zu marktfähigen Produkten oder Services führen. Laut Berichten des IT-Beratungs- und Marktforschungsunternehmens Gartner aus dem Jahr 2020 schafft nur etwa die Hälfte aller entwickelten KI-Systeme überhaupt den Übergang vom Prototyp zum marktfähigen Produkt oder Service. Eine US-amerikanische Quelle geht zudem davon aus, dass knapp 80 Prozent aller Vorhaben noch vor der Inbetriebnahme des KI-Entwicklungsgegenstands abgebrochen werden, ein Drittel sogar bevor überhaupt ein Proof-of-Concept erbracht wurde[1].

Um die Marktdurchdringung von KI-Systemen zu fördern, ist es notwendig, ein auf KI-Produkte bzw. Services ausgerichtetes Qualitätsmanagement aufzusetzen, das KI-Systeme konsequent über alle ihre Lebensphasen begleitet – von der Idee bis zum operativen Einsatz. Dadurch werden die kritischen Faktoren, die maßgeblich zum Scheitern von KI-Vorhaben beitragen, besser beherrschbar. Diese kritischen Faktoren fallen insbesondere in die folgenden Kategorien: unrealistische Erwartungen, anwendungsfallbezogene Probleme (z. B. ein unausgewogenes Kosten-Nutzen-Verhältnis), organisatorische Randbedingungen (z. B. Regulierungshürden), Mangel an Schlüsselressourcen (z. B. Daten, Knowhow) und technologische Probleme (Westenberger et al. 2022).

Der Leitfaden staffelt daher das Qualitätsmanagement entlang von vier Lebensphasen von KI-Systemen: Charakterisierung, Design, Entwicklung und Betrieb. Er führt anschließend für jede Lebensphase Kriterien für Qualität an und unterlegt sie mit Indikatoren, anhand derer sich die Erfüllung der Kriterien nachverfolgen lässt. Die meisten der als Fragen formulierten Indikatoren sind mit Lösungshilfen verknüpft, die bei der Beantwortung unterstützen, indem sie auf bestehende und entstehende Normen und Standards, etablierte Werkzeuge, Best Practices sowie Forschungsansätze verweisen.

Für die Lebensphase Charakterisierung sieht der Leitfaden drei Kriterien vor: Kritikalität, Komplexität und Machbarkeit. Ein Indikator für die Machbarkeit umfasst dabei z. B. die Abwägung des Wertschöpfungspotenzials gegenüber dem für die Umsetzung zu erwartenden Aufwand. Die mit dem Indikator verknüpfen Lösungshilfen stellen Best Practices und Forschungsansätze für eine solche Kosteneffizienzbewertung vor. Die Lebensphase Design umfasst vier Kriterien, darunter die Verfügbarkeit von Ressourcen wie Daten. Der Leitfaden bietet an dieser Stelle u. a. einen Überblick über gängige Herangehensweisen und Ansätze aus der Literatur zur Abschätzung des minimal erforderlichen Datenvolumens. Zu den sechs Kriterien für die Phase Entwicklung zählt die Leistungsbewertung des KI-Systems bzw. der KI-Komponente. Der Leitfaden verweist hier u. a. auf vorhandene ISO-Leitlinien für das Testen von KI-basierten Systemen sowie zukünftige

1 https://content.alegion.com/download-dimensional-research-finds-ai-still-nascent

technische Spezifikationen, die sich aktuell noch in Bearbeitung befinden. Der Lebensphase Betrieb unterliegen vier Kriterien, darunter die Bedienbarkeit des KI-Systems. Die mit den Indikatoren verknüpften Lösungshilfen umfassen dabei u. a. allgemeine ISO-Standards für interaktive Systeme (u. a. Interaktionsprinzipien, Nutzendenführung). Insgesamt verzeichnet der Leitfaden 60 Indikatoren (Charakterisierung: 11, Design: 17, Entwicklung: 20, Betrieb: 12) und 78 Lösungshilfen (Charakterisierung: 12, Design: 18, Entwicklung: 39, Betrieb: 9).

Der Leitfaden schließt mit einem Ausblick, der die Aspekte des Qualitätsmanagements adressiert, für die bisher nur wenige, in der Regel unausgereifte, Lösungshilfen existieren. Dazu zählen die Überführung von derzeit horizontal ausgerichteten Normen und Standards in vertikale Branchennormen, das Testen von KI-Systemen unter Realbedingungen, die Entwicklung von Vorgehensmodellen für das Testen der Robustheit von KI-Komponenten sowie für das Monitoring und die Aufrechterhaltung der Leistung von KI-Systemen im Betrieb und letztlich auch die Entwicklung von Indikatoren für die Umweltwirkung von KI-Systemen.

INHALT

1 EINLEITUNG

1 EINLEITUNG

Mit der fortschreitenden Digitalisierung von Wirtschaft, Wissenschaft und Gesellschaft nimmt das pro Jahr global generierte Datenvolumen enorm zu. Das Marktforschungs- und Beratungsunternehmen IDC prognostiziert, dass es 2025 175 Zettabytes erreichen wird (IDC 2018). Dies entspricht einer Verfünffachung gegenüber dem Jahr 2018. Mit dem Volumen steigt auch die Vielfalt der erhobenen Daten – von Logdateien über technische Messreihen bis hin zu Bild-, Video- und Audioformaten. Technologien der Künstlichen Intelligenz (KI) wie das Maschinelle Lernen (engl. machine learning) sind mächtige Werkzeuge, mit denen sich Auswertungen von solch enorm großen Datenmengen erheblich erleichtern und Prozesse in praktisch allen Anwendungsdomänen gezielt verbessern oder automatisieren lassen (High-Level Expert Group on Artificial Intelligence (AI HLEG) 2018; Bitkom e. V. o. J.). Das Spektrum KI-basierter Produkte und Services ist dementsprechend weit und reicht von der individualisierten Empfehlung von Musiktiteln auf einer Unterhaltungsplattform über die automatische Erkennung von Verkehrsschildern und -situationen im Umfeld eines Kraftfahrzeugs bis hin zur Steuerung von Fertigungsprozessen in der Industrie.

Entwicklung und Inbetriebnahme von KI-Produkten und -Services

Hinter KI-Produkten und -Services verbergen sich Konstrukte, die sich in Aufbau und Komplexität stark unterscheiden können (z. B. Integration in eine Softwarearchitektur und gegebenenfalls Einbettung in ein physisches Gesamtsystem wie Fahrzeug, Produktionsanlage oder Roboter). In der Fachliteratur sind bislang noch keine einheitlichen bzw. verbindlichen Begrifflichkeiten für die Beschreibung des Aufbaus der Konstrukte verankert. Das Gesamtkonstrukt wird in der Regel als KI-System bezeichnet, wenn mindestens eine seiner Komponenten eine Funktionalität mittels Künstlicher Intelligenz umsetzt[2]. Im Rahmen des Prüfkatalogs des Fraunhofer IAIS für Vertrauenswürdige KI wird der formale Aufbau einer KI-Anwendung entlang der Begriffe KI-Modell (aus einem maschinellen Lernverfahren hervorgehendes mathematisches Objekt), KI-Komponente (KI-Modell einschließlich Verfahren zur Datenvor- und Datennachverarbeitung), Einbettung (Verknüpfung mit Software und weiteren technischen Modulen) sowie Interface/Schnittstelle (Interaktion mit der Außenwelt oder dem Gesamtsystem) spezifiziert (Poretschkin et al. 2021). Die vorliegende Publikation orientiert sich im Folgenden an diesen Begrifflichkeiten.

KI-Anwendungen bzw. KI-Systeme werden zunächst unter laborähnlichen, d. h. antizipier- und beeinflussbaren, Bedingungen entwickelt. Realbedingungen, mit denen KI-Systeme jedoch später im operativen Betrieb konfrontiert werden, unterscheiden sich von diesen Laborbedingungen unter Umständen erheblich. Tatsächlich versagen KI-Systeme, die das Entwicklungslabor verlassen, in der Praxis häufig. Dies kann mit für Menschen und Gesellschaft schädlichen bzw. nachteiligen Vorfällen verbunden sein[3]. Laut Berichten des IT-Beratungs- und Marktforschungsunternehmens Gartner schaffen nur 53 Prozent aller KI-Systeme überhaupt den Übergang vom Prototyp zum marktfähigen Produkt oder Service[4]. Laut einer vom Marktforschungsinstitut Dimensional Research für die Data-Labeling-Software-Plattform Alegion durchgeführten Befragung von US-amerikanischen KI-Expertinnen und Experten werden 78 Prozent aller KI-Projekte vor der praktischen Inbetriebnahme abgebrochen, 33 Prozent sogar vor dem Proof-of-Concept[5].

2 https://www.iese.fraunhofer.de/blog/dependable-ai
3 Die AI Incidents Database (https://incidentdatabase.ai/) gibt einen Überblick über KI-Anwendungen und Systeme, bei denen es im praktischen Einsatz zu Vorfällen kam. Mit der Datenbank wird das Ziel verfolgt, ein wiederholtes Scheitern in realen Einsatzsszenarien aufgrund gleicher Gründe zu verhindern bzw. einem Scheitern vorzubeugen (McGregor 2020).
4 https://www.gartner.com/en/newsroom/press-releases/2020-10-19-gartner-identifies-the-top-strategic-technology-trends-for-2021
5 Die Ergebnisse der Umfrage sind unter dem Titel „What data scientists tell us about AI model training today" zusammengefasst worden (https://content.alegion.com/download-dimensional-research-finds-ai-still-nascent).

Insgesamt schreitet der Einsatz von KI-Systemen in der Praxis daher langsamer voran als erwartet. Eine aktuelle Umfrage des Vereins Deutscher Ingenieure (VDI) unter seinen Mitgliedern unterstreicht dies, indem sie zeigt, dass die Erwartungen bzw. Prognosen der Mitglieder aus dem Jahr 2018 bis heute signifikant nicht erreicht wurden (Billerbeck 2022). Gerade kleine und mittelständische Unternehmen entscheiden sich noch oft gegen entsprechende Investitionen, weil eine Unsicherheit bleibt, ob diese sich letztlich auszahlen werden.

Kritische Faktoren für den Erfolg von KI-Vorhaben

Die ausschlaggebenden Faktoren für das Scheitern von KI-Vorhaben lassen sich laut einer Studie von Westenberger et al. 2022 in fünf Kategorien einteilen: unrealistische Erwartungen, anwendungsfallbezogene Aspekte (unausgewogenes Kosten-Nutzen-Verhältnis, hohe Komplexität), organisatorische Randbedingungen (niedriges Budget, rechtliche Hürden), Mangel an Schlüsselressourcen (Datenverfügbarkeit, Fachexpertise) und technologische Probleme (Modellinstabilität, mangelnde Transparenz, Anfälligkeit für Manipulation). Um die Erfolgsrate von KI-Vorhaben zu verbessern bzw. den erfolgreichen Einsatz von KI-Systemen in der Praxis voranzutreiben, müssen Instrumente entwickelt werden, mit denen sich diese vielfältigen technischen und nicht-technischen Faktoren bestmöglich beherrschen lassen. Ein auf KI-Produkte bzw. -Services ausgerichtetes Qualitätsmanagement, das die Entwicklung von der Idee bis zum Einsatz begleitet, stellt ein solches Instrument dar.

Unter Qualitätsmanagement versteht man allgemein die Planung, Steuerung und Überwachung der Qualität eines Prozesses bzw. Prozessergebnisses; dies umfasst die Qualitätsplanung, -lenkung, -prüfung, -verbesserung und -sicherung[6]. Das Qualitätsmanagement im Kontext von KI-Vorhaben sollte eine frühe Identifizierung unüberwindbarer Machbarkeitsprobleme und nicht erfüllbarer Anforderungen ermöglichen (d. h. bevor hohe finanzielle und personelle Ressourcen in die Entwicklung eines Produkts oder Services fließen) sowie nicht eingeplante, nachgelagerte Kosten möglichst niedrig halten. Um einem finanziellen Verlust sukzessive vorzubeugen, sollte das Qualitätsmanagement daher in Phasen gestaffelt sein, die Planungs-, Design-, Entwicklungs- und Betriebsaspekte voneinander trennen. Nicht alle der für den Erfolg eines KI-Vorhabens kritischen Faktoren können bereits in der Planungsphase hinreichend berücksichtigt werden. Beispielsweise ist es nahezu unmöglich, alle möglichen Quellen für Instabilitäten eines KI-Modells vor Aufnahme der aktiven Entwicklungsarbeiten vorherzusagen. Andere Faktoren können dagegen zumindest abgeschätzt werden, wie z. B. eine ausreichende Verfügbarkeit von Ressourcen. Darüber hinaus sollte das Qualitätsmanagement dazu beitragen, zukünftigen Problemen im Hinblick auf die Konformität der Produkte oder Services mit Rechts- und Sicherheitsvorschriften sowie Zertifizierungsanforderungen vorzubeugen. Letzteres steht im direkten Zusammenhang mit den kritischen Faktoren der Kategorie „organisatorische Randbedingungen". Diese Faktoren nehmen eine besondere Rolle ein, da die Schaffung geeigneter Rahmenbedingungen in Bezug auf KI aktuell noch Gegenstand laufender Gesetzgebungsvorhaben sowie laufender Normungs- und Standardisierungsverfahren ist, die im nächsten Abschnitt kurz skizziert werden.

6 https://wirtschaftslexikon.gabler.de/definition/total-quality-management-tqm-47755/version-271017

Gestaltung besserer Rahmenbedingungen für KI-Systeme

Um die Verbreitung von KI-Systemen zu begünstigen, betreiben aktuell sowohl Gesetzgeber als auch Normungs- und Standardisierungsgremien viel Aufwand, um geeignete Rahmenbedingungen herzustellen. Zu den zentralen Aufgaben zählen dabei die Schaffung eines rechtlich-regulatorischen Rahmens für den Einsatz und das Inverkehrbringen von KI-Systemen, die Etablierung von Normen und Standards für KI-Systeme sowie die Erarbeitung von angemessenen Kriterienkatalogen für die Prüfung und Zertifizierung von KI-Systemen durch unabhängige Stellen im Sinne einer Qualitätssicherung bzw. Konformitätsbewertung. Darüber hinaus müssen Anreize für die Entwicklung von KI-Systemen geschaffen und Maßnahmen etabliert werden, die die Wirtschaftlichkeit von KI-Projekten verbessern, z. B. durch Förderprogramme und KI-Recheninfrastrukturinitiativen. Das Policy Observatory Portal der OECD[7] (OECD.AI) verzeichnet dementsprechend aktuell mehr als 700 Initiativen aus über 60 Ländern, die einen Bezug zur Ausgestaltung des Rechtsrahmens und der Zertifizierung von KI-Systemen haben, Anreize für die KI-Entwicklung schaffen oder die Wirtschaftlichkeit von KI-Projekten verbessern. Davon entfallen 62 auf die Vereinigten Staaten von Amerika (USA) und 61 auf die Europäische Union. Viel Aktivität wird auch aus Australien (37 Initiativen) sowie China, Japan und Indien (jeweils 23 Initiativen) registriert.

Geht es um Regulierung und Zertifizierung von KI, gibt es derzeit keine einheitlichen, sektorenübergreifenden Metriken oder Verfahren, die konkrete Qualitätskriterien für KI-Systeme festlegen. Dennoch zeichnen sich bereits jetzt erste Ansätze zur Regulierung und Zertifizierung von KI-Systemen ab. Mit der KI-Verordnung soll erstmals ein EU-weit einheitlicher Rechtsrahmen für das Inverkehrbringen, die Inbetriebnahme und Verwendung von KI-Systemen auf den Weg gebracht werden. Das dort enthaltene Regelungsprinzip folgt einem risikobasierten Ansatz: Wo der Einsatz von KI mit einem hohen Risiko assoziiert ist, müssen hohe Mindestanforderungen umgesetzt werden. Bei einem geringen Risiko sind demgegenüber keine regulatorischen Pflichten vorgesehen. (Europäische Kommission 21.04.2021). Ein risikobasierter Ansatz wird auch in den USA verfolgt und dort insbesondere vom National Institute of Standards and Technology (NIST) des U.S. Departments of Commerce (DoC) vorangetrieben. Im März 2022 hat das NIST einen Erstentwurf für ein KI-Risikomanagement-Framework veröffentlicht, das eine Grundlage für die Bemessung des Risikos von KI-Systemen in ihrer jeweiligen Lebensphase schaffen soll (NIST 2022). Neben dem DoC erarbeiten in den USA aktuell auch das Weiße Haus, die Food and Drug Administration (FDA), die National Security Commission on Artificial Intelligence, das Government Accountability Office (GAO) und die Federal Trade Commission (FTC) entsprechende Richtlinien[8]. Die Regulierung von KI-Systemen nimmt somit auch in den USA Form an. Die Aufnahme von KI-spezifischen Fragestellungen in die Prioritätenliste des transatlantischen EU-US Trade and Technology Council (TTC) und die daraus im Mai 2022 hervorgegangene Absichtserklärung, gemeinsam Methoden zur Bemessung der Risiken und Zuverlässigkeit von KI-Systemen zu entwickeln[9], deutet an, dass künftig eine gemeinsame Ausrichtung in Regulierungs- und Konformitätsbewertungsfragen angestrebt wird.

7 Mit dem OECD Policy Observatory steht seit 2020 ein Portal zur Verfügung, das die weltweiten Ansätze zur Gestaltung und Regulierung von KI-Systemen dokumentiert und über Dashboards einsehbar und vergleichbar macht (https://oecd.ai/en/dashboards). Grundlage des Portals ist eine in Zusammenarbeit mit der Europäischen Kommission erstellte Echtzeit-Datenbank zu KI-Initiativen aus aktuell über 60 Ländern.

8 Für eine Übersicht zu Aktivitäten in den USA mit Bezug zu KI-Regulierung siehe: https://www.orrick.com/en/Insights/2021/11/US-Artificial-Intelligence-Regulation-Takes-Shape

9 Second Ministerial Meeting (15-16 May 2022) des TTC: https://ec.europa.eu/info/strategy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en

Im Hinblick auf die Konformitätsbewertung und Zertifizierung von KI-Systemen werden derzeit viele sowohl ethische als auch informatisch-technologische Aspekte diskutiert, darunter ihre Sicherheit, Transparenz, Nachprüfbarkeit, Diskriminierungsfreiheit, Menschenzentrierung, Nutzerfreundlichkeit, Systemoperabilität sowie die Begrenzung ihrer Systemfunktionalität (Heesen et al. 2020). Auf nationaler Ebene liegen erste Vorschläge für entsprechende Prüf- bzw. Kriterienkataloge vor, darunter der Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz des Fraunhofer IAIS (Poretschkin et al. 2021), der AIC4-Katalog für cloudbasierte KI-Systeme des Bundesamts für Sicherheit in der Informationstechnik (BSI) (Bundesamt für Sicherheit in der Informationstechnik 2021), die VDE SPEC 90012 V1.0 für die Charakterisierung der Vertrauenswürdigkeit von KI-Systemen (VDE 2022) und das Rahmenwerk für die Bewertung der ethischen Konformität der vom VDE und der Bertelsmann Stiftung geleiteten AI Ethics Impact Group (AI Ethics Impact Group 2020). Wie die Einhaltung der veranschlagten Kriterien konkret messbar und somit überprüfbar gemacht werden soll, ist in Bezug auf viele potenziell prüfrelevante Aspekte noch ungeklärt..

Ziel des hier vorgestellten Leitfadens für das Qualitätsmanagement bei der Entwicklung von KI-Produkten und -Services ist es, die in die Entwicklung eines KI-Systems involvierten Akteure mithilfe eines Fragenkatalogs dabei zu unterstützen, ihre gemeinsame Produktvision auf ihre Umsetzbarkeit zu überprüfen und sie gegebenenfalls möglichst effizient in die Realität zu überführen. Grundlegend für den Aufbau des Leitfadens ist dabei die Staffelung des Qualitätsmanagements entlang von vier Lebensphasen eines KI-Systems (Charakterisierung, Design, Entwicklung, Betrieb) und die Festlegung von Kriterien für die Qualität der individuellen Lebensphasen sowie von Indikatoren, anhand derer sich die Erfüllung der jeweiligen Kriterien verfolgen lässt. Die Indikatoren sind dabei durchgängig als Fragen formuliert und mit Lösungshilfen verknüpft, die ihre Bewertung bzw. Beantwortung erleichtern sollen. Die Lösungshilfen verweisen auf bestehende und in der Entwicklung befindliche Normen und Standards, etablierte Werkzeuge, Best Practices und Forschungsansätze.

Für den Entwurf des Leitfadens wurde zunächst eine Literaturrecherche durchgeführt, um eine fachwissenschaftliche Grundlage zu den Themen Qualität und Qualitätsbewertung von KI-Systemen zu schaffen. Der Aufbau des Leitfadens erfolgte danach strukturell angelehnt an das sogenannte WKIO-Modell (Hubig 2016). Der initiale Entwurf des Leitfadens wurde im Juni/Juli 2022 im Rahmen von elf Interviews mit Expertinnen und Experten aus Wirtschaft (fünf Interviews) und Forschung (sechs Interviews) validiert, die ein weites Anwendungsspektrum für KI-Systeme abdecken (Gesundheitswesen, Produktion, Prozessindustrie, Landwirtschaft, Automobilbranche, Smart Living, Krisenmanagement, Emotion Analytics und Logistikroboter). Die Resultate der Interviews wurden anschließend für die Finalisierung des Leitfadens konsolidiert. Weitere Hinweise zum Aufbau und zur Verwendung des Leitfadens werden in Kapitel 2 beschrieben. Der Leitfaden selbst wird in Kapitel 3 vorgestellt, Kapitel 4 fasst aktuelle Bedarfe im Hinblick auf das Qualitätsmanagement von KI-Systemen zusammen und benennt zukünftige Handlungsfelder.

Das Autorenteam möchte sich an dieser Stelle noch einmal ganz herzlich bei allen Interviewpartnerinnen und -partnern für ihre wertvollen Beiträge, Zeit und Offenheit bedanken:

- Frau Dr. Klaudia Dussa-Zieger, imbus AG (Unterauftragnehmer im Projekt Agri-GAIA)
- Herr Dr. Hüseyin Erdogan, Continental AG
- Herr Dr. Hilko Hoffmann, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH (Projekt ForeSight)
- Herr Prof. Dr. Martin Leucker, Universität zu Lübeck (Projekt KI-SIGS)
- Herr Dr. Marco Maier, Tawny GmbH
- Herr Dr. Andreas Schmidt, Fraunhofer IESE (Projekt FabOS)
- Herr Dr. Daniel Schneider, Fraunhofer IESE (Projekt SPELL)
- Frau Andrea Suckro, slashwhy GmbH & Co. KG (Projekt IIP-Ecosphere)
- Herr Dr. Moritz Tenorth, Magazino GmbH
- Herr Prof. Dr. Leon Urbas, Technische Universität Dresden (Projekt KEEN)
- Herr Norman Zerbe, Charité – Universitätsmedizin Berlin (Projekt EMPAIA)

Die Verantwortung für den Inhalt dieses Leitfadens liegt ausschließlich bei den Autorinnen und Autoren.

Der Leitfaden wurde im Rahmen der Begleitforschung zum Innovationswettbewerb „Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme“ im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz erstellt. Das Programm umfasst aktuell 26 Projekte, die neue Formen KI-basierter Plattformökonomie in wichtigen Sektoren der deutschen Wirtschaft entwickeln. Mehrere der Interviewpartnerinnen und Interviewpartner sind in diesen Projekten aktiv.

2 AUFBAU UND VERWENDUNG DES LEITFADENS

2 AUFBAU UND VERWENDUNG DES LEITFADENS

Der Leitfaden ist anwendungsagnostisch konzipiert. Er führt an, welche Aspekte in den jeweiligen Lebensphasen eines KI-Systems im Hinblick auf das Qualitätsmanagement grundsätzlich berücksichtigt werden sollten und unterstützt dabei, dies umzusetzen. Der Leitfaden formuliert keine technischen Anforderungen für zukünftige Konformitätsbewertungen, Zulassungs- oder Zertifizierungsprozesse. Herausforderungen in Bezug auf die hardwarebezogene Systemintegration und die Cyber-Sicherheit werden in diesem Leitfaden nicht oder nur in groben Ansätzen berücksichtigt. Aus Gründen der Kompaktheit werden zudem Aspekte der allgemeinen Softwarequalität und der passenden Nutzung/Eignung von Machine-Learning-Frameworks bzw. -Entwicklungsumgebungen zwar implizit berücksichtigt, aber nicht eigenständig aufgeschlüsselt.

Im Folgenden wird zunächst der grundsätzliche Aufbau des Leitfadens vorgestellt. Das Kapitel schließt mit Hinweisen zur seiner Verwendung.

2.1 Aufbau des Leitfadens

Im Aufbau lehnt sich der Leitfaden an das sogenannte WKIO-Modell von Hubig 2016 an (engl. VCIO model). Dieses Modell wurde in jüngster Vergangenheit bereits im Kontext der Bewertung der ethischen Konformität von KI-Systemen (AI Ethics Impact Group 2020) sowie der Bewertung der Zuverlässigkeit von KI-Systemen (VDE 2022) hinzugezogen. Aus Gründen der Vereinbarkeit und Kohärenz der Konzepte orientiert sich auch der hier vorgestellte Leitfaden an diesem Modell (siehe Unterkapitel 5.1 und 5.2 im Anhang für Erläuterungen zum klassischen WKIO-Modell bzw. zu seiner modifizierten Umsetzung im Rahmen des Leitfadens).

Grundlegend für den Aufbau des Leitfadens ist die Staffelung des Qualitätsmanagements entlang der Lebensphasen eines KI-Systems (Charakterisierung, Design, Entwicklung, Betrieb; siehe Abbildung 1) und die Festlegung von Kriterien für die Qualität der individuellen Lebensphasen sowie von Indikatoren, anhand derer sich die Erfüllung der jeweiligen Kriterien verfolgen lässt (Abbildung 2). Die Indikatoren sind dabei durchgängig als Fragen formuliert und mit Lösungshilfen verknüpft, die die Bewertung bzw. Beantwortung erleichtern können (Abbildung 2). Abbildung 3 gibt einen Überblick über die Zusammensetzung des Leitfadens in Zahlen (Anzahl der Kriterien, Anzahl der Indikatoren und Anzahl der Lösungshilfen je Lebensphase).

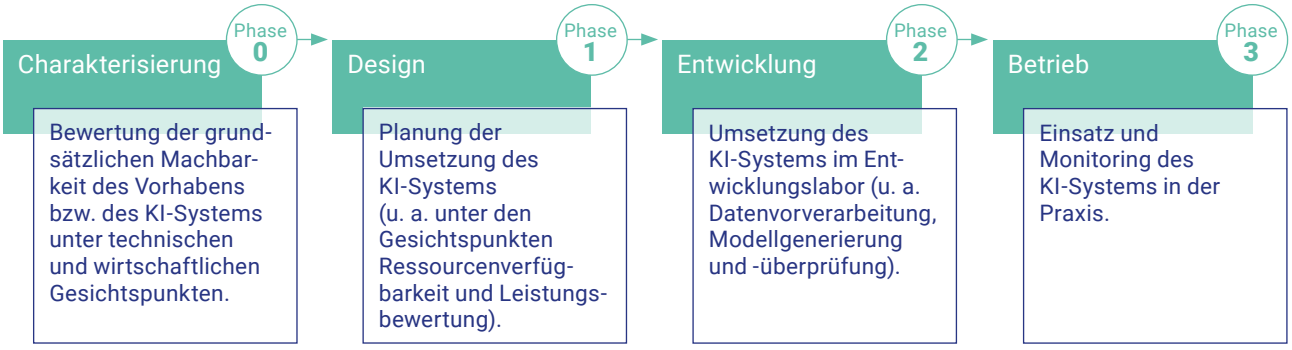


Abbildung 1: Einteilung der Lebensphasen von KI-Systemen im Rahmen des Leitfadens. Die Einteilung orientiert sich grundsätzlich am Vorschlag der OECD zur Einteilung der Lebensphasen von KI-Systemen (OECD 2020). Um Planungs-, Design- und Entwicklungsaspekte schärfer zu trennen, wurde der Vorschlag jedoch entsprechend angepasst.

Die Lösungshilfen umfassen bestehende und in der Entwicklung befindliche Normen und Standards, etablierte Werkzeuge, Best Practices sowie Forschungsansätze. Bei den verzeichneten Hilfen handelt es sich um eine von den Autorinnen und Autoren getroffene Auswahl, die keinen Anspruch auf Vollständigkeit erhebt. Für ausgewählte Indikatoren bzw. komplexere Fragestellungen haben die Autorinnen und Autoren zusätzlich Orientierungshilfen verfasst, die die Auseinandersetzung mit der zugrunde liegenden Thematik erleichtern sollen.

Die Lösungshilfen wurden für den Leitfaden entsprechend ihrer Kategorie bzw. ihres Reifegrads farblich kodiert; die Kodierung enthält zudem eine inhaltliche Kurzbezeichnung sowie einen Quellenverweis (Abbildung 2). Der Leitfaden wird von einem Katalog aller Lösungshilfen (siehe Anhang A.3) begleitet, dem zusätzliche Informationen wie der vollständige Titel sowie der Hyperlink, über den sich die jeweilige Hilfe direkt öffnen lässt, entnommen werden können.

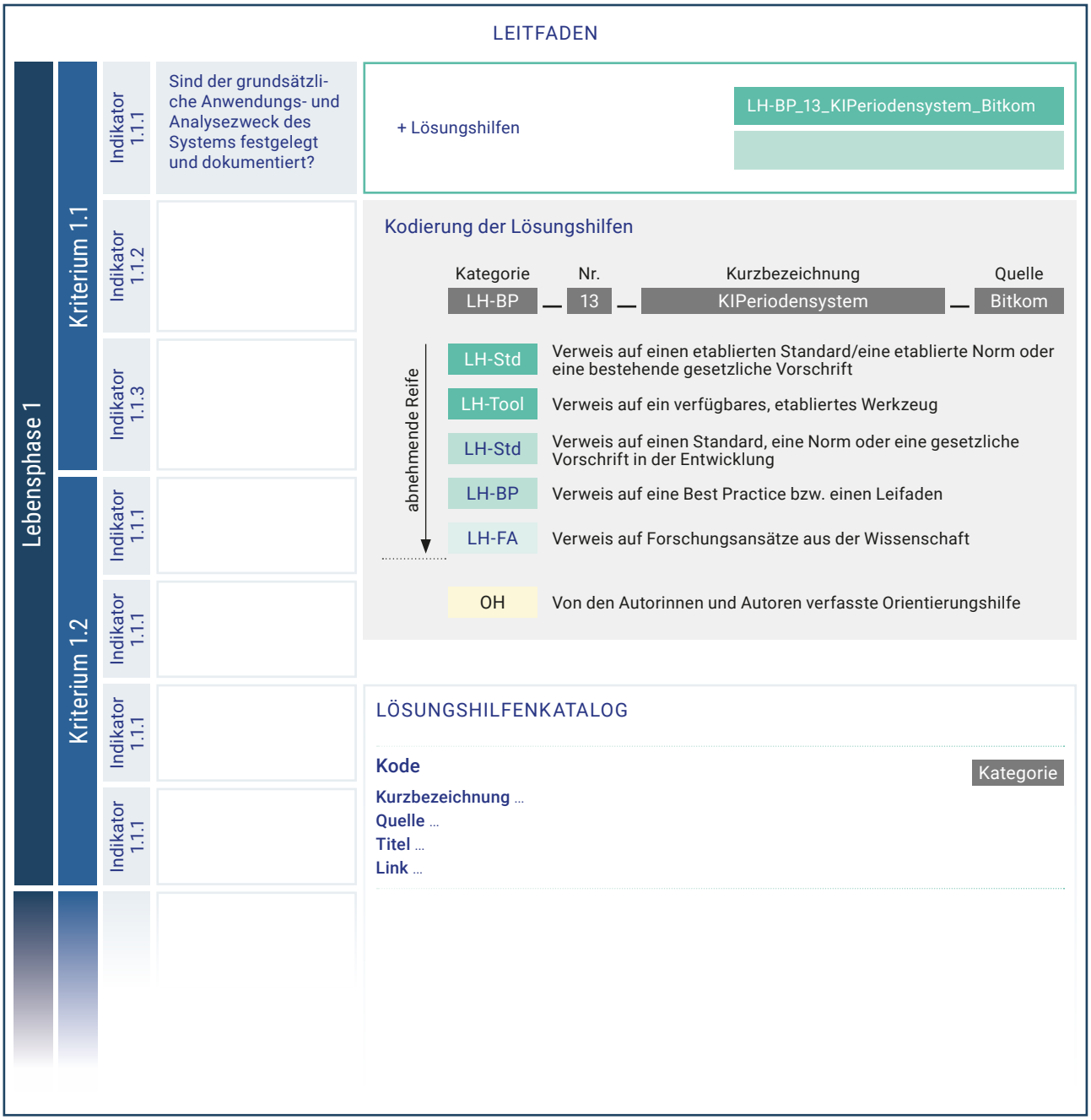


Abbildung 2: Aufbau des Leitfadens, Kodierung der Lösungshilfen und Verknüpfung des Leitfadens mit dem Lösungshilfenkatalog.

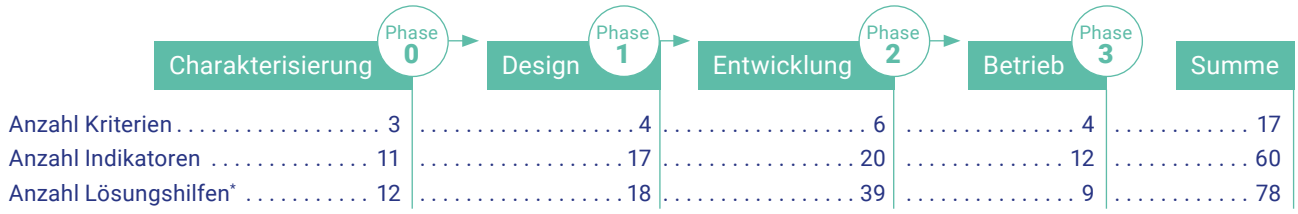


Abbildung 3: Zusammensetzung des Leitfadens. *Eine Lösungshilfe kann im Kontext mehrerer Werte verankert sein; für die Darstellung hier ist jede Lösungshilfe nur dem ersten Wert zugeordnet worden, für den sie angeführt wird.

2.2 Hinweise zur Verwendung des Leitfadens

Die grafische Darstellung des Leitfadens im folgenden Kapitel 3 (auf den Seiten 17 bis 24) soll als handliche Übersicht zu den Lebensphasen, ihren Kriterien und jeweiligen Indikatoren sowie den zugehörigen Kodes der Lösungshilfen dienen. Aus dem Kode sind direkt der Inhalt und die Quelle sowie die Kategorie bzw. der Reifegrad der Hilfen ersichtlich. Der mit dem Leitfaden verknüpfte Lösungshilfenkatalog (Anhang A.3) ermöglicht über entsprechende Hyperlinks zudem einen direkten Zugang zu den Hilfen.

In den Unterkapiteln 3.1, 3.2, 3.3 und 3.4 wird jeweils die Bedeutung der Lebensphasen sowie die Relevanz und der inhaltliche Kontext der für sie festgelegten Qualitätskriterien näher erläutert. Im Rahmen der Erläuterungen zu den Kriterien wird auf die Indikatoren verwiesen, mit denen sich ihre Erfüllung verfolgen lässt. Die bereits in der grafischen Darstellung des Leitfadens in Form von Fragen verankerten Indikatoren werden dabei aus Gründen der Kompaktheit nicht noch einmal wortwörtlich wiedergegeben. Die Erläuterungen schließen jeweils mit einer Vorstellung der im Leitfaden verankerten Lösungshilfen und Verweisen auf die Indikatoren, für deren Betrachtung sie hinzugezogen werden können.

Der Leitfaden soll als eine pragmatische Hilfe für das Qualitätsmanagement im Rahmen der Umsetzung von KI-Vorhaben dienen. An einer solchen Umsetzung können je nach Konstellation verschiedene Organisationen und Stakeholder aus Wirtschaft, Wissenschaft und dem öffentlichen Sektor sowie Personen mit unterschiedlichen beruflichen Hintergründen und somit unterschiedlicher KI-Kompetenz beteiligt sein. Der Leitfaden soll für all diese Gruppen zugänglich sein, daher werden die Erläuterungen zu den Kriterien und ihren Indikatoren durch veranschaulichende Elemente wie fiktive und reale Beispielszenarien sowie Orientierungshilfen ergänzt. Die veranschaulichenden Elemente stellen ein optionales Angebot dar, das je nach individuellem Bedürfnis genutzt werden kann.

Aufgrund seiner anwendungsagnostischen Konzipierung umfasst der Leitfaden Kriterien und Indikatoren, die vom Team der Autorinnen und Autoren sowie der Mehrzahl der interviewten Expertinnen und Experten als potenziell relevant für das Qualitätsmanagement in der jeweiligen Lebensphase eines KI-Systems eingestuft werden. Für das eigene konkrete Umsetzungsziel heißt das, dass die Kriterien bzw. ihre Indikatoren relevant sein können, aber nicht zwingend relevant sein müssen. Zunächst sollten die an der Umsetzung beteiligten Organisationen und Stakeholder daher zusammen eine entsprechende Abwägung und gegebenenfalls Priorisierung im Hinblick auf ihr gemeinsam angestrebtes anwendungsspezifisches Umsetzungsziel vornehmen. Je nach Ziel und Anwendung können gegebenenfalls bestimmte Kriterien bzw. Indikatoren nicht oder nur von sehr geringer Bedeutung sein.

3 LEITFADEN FÜR DAS PHASEN-GESTAFFELTE QUALITÄTS-MANAGEMENT BEI DER ENTWICKLUNG VON KI-PRODUKTEN UND -SERVICES

3 LEITFADEN FÜR DAS PHASENGESTAFFELTE QUALITÄTSMANAGEMENT BEI DER ENTWICKLUNG VON KI-PRODUKTEN UND -SERVICES

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
0. Charakterisierung	0.1 Kritikalität	I0.1.1	Liegt eine Einordnung des potenziellen Risikos des angestrebten Produktes oder Services vor, die den Einsatzkontext und das zukünftige Gesamtsystem berücksichtigt?	LH-Std_1_ISO 14971:2019:RiskManagement_Medizinprodukt_ISO LH-Std_2_ISO/TR22100-5:2021Maschinensicherheit_ISO LH-Std_3_AIRiskManagement_NIST LH-Std_4_ISO/IEC_FDIS23894_RisikomanagementKI_ISO LH-FA_5_RiskClassificationIEAI_IEAI
	0.2 Komplexität	I0.2.1	Ist der Einsatzkontext hinreichend bekannt und sind veränderliche Wechselwirkungen mit anderen Entitäten oder Umwelteinwirkungen in der Einsatzumgebung zu berücksichtigen? (z. B. autonome Teilsysteme, menschlich gesteuerte Prozesse, Temperaturänderung)	Anwendungswissen OH
		I0.2.2	Bleibt das betrachtete Analyseobjekt bzw. das gegebene Realsystem über die Zeit unverändert? Falls nein, ist es erforderlich, die Zeit bzw. das Zeitverhalten des Analyseobjekts/ Realsystems als Ordnungsstruktur bei der Produkt- oder Serviceentwicklung zu berücksichtigen?	Anwendungswissen OH
		I0.2.3	Sind (z. B. physische oder sicherheitsrelevante) Grenzen des zukünftigen Gesamtsystems zu berücksichtigen, die den zulässigen Wertebereich von Ausgaben (Lösungsraum) einschränken?	Anwendungswissen
	0.3 Machbarkeit	I0.3.1	Sind die relevanten Größen des Analyseobjekts/ Realsystems zugänglich bzw. direkt messbar oder über andere zugängliche oder messbare Größen abschätzbar bzw. beobachtbar? Liegt ein entsprechender Machbarkeitsnachweis vor (z.B. Datenakquisition in kontrollierter Laborumgebung)?	LH-BP_6_Beobachtbarkeit_GoogleCloud LH-FA_7_ObservabilityComplexSystems_Stigter Anwendungswissen
		I0.3.2	Können die Daten selbst erzeugt bzw. mit vertretbarem Aufwand akquiriert werden (z. B. durch Experimente, Simulation) oder ist ihre Beschaffung von begrenzt verfügbaren, teuren oder riskanten Messungen abhängig?	Anwendungswissen
		I0.3.3	Ist die Nutzung der verfügbaren Daten rechtens (z. B. im Hinblick auf Datenschutz, Informationssicherheit oder Lizenzen)?	LH-BP_8_PrivacyGovernance_HLEG
		I0.3.4	Ist der Prozess der Datengenerierung dokumentiert (z. B. Messapparaturaufbau, Messvorgang, Messprotokoll)? Wenn ja, liegen Informationen zu Fehlerquellen, auftretenden Messfehlern/Störungen und deren statistischer Verteilung vor?	LH-BP_9_SOP_Hollmann
		I0.3.5	Ist das Format, in dem die Daten zur Verfügung gestellt werden, dokumentiert (unstrukturiert, struktuiert, oder semi-strukturiert)?	LH-MW-10_DataFormats_IBM
		I0.3.6	Ist es notwendig, auf zusätzliches (nicht in Form von Daten vorliegendes) Anwendungs-/Methodikwissen zurückzugreifen, um Teilaufgaben zu realisieren?	Anwendungswissen/Methodenwissen
		I0.3.7	Werden das Wertschöpfungspotenzial (z. B. Kostenreduzierung im Vergleich zum Stand der Technik) und die Komplexität der Lösungsumsetzung (z. B. erforderliche Ressourcen, Risiken) im Rahmen einer Kosten-Nutzen-Analyse gegeneinander abgewogen?	LH-BP_11_AIBusinessModelCanvas_Kerzel LH-FA_12_CostBenefitMedicine_Ziegelmayer Anwendungswissen/Methodenwissen

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
1. Design	1.1 Zweckbestimmung des Systems	I1.1.1	Sind der grundsätzliche Anwendungs- und Analysezweck des Systems festgelegt und dokumentiert?	LH-BP_13_KIPeriodensystem_Bitkom
		I1.1.2	Ist die Art des Systems eindeutig festgelegt und dokumentiert? Wenn ja, ist der Autonomiegrad des Systems definiert und dokumentiert?	LH-BP_14_AutonomiestufenIndustrie_Plattform Industrie 4.0
		I1.1.3	Sind die funktionalen Anforderungen an das System festgelegt, priorisiert und dokumentiert?	LH-Std_15_ISO/IEC25010:2011_ISO LH-Std_16_ISO/IEC DIS 25059_ISO LH-FA_17_RequirementsEngineering_Vogelsang LH-FA_18_RequirementsEngineeringSafetyCritical_Martins Anwendungswissen/Methodenwissen
		I1.1.4	Sind die jeweiligen ggf. einzuhaltenden Nebenbedingungen festgehalten (z. B. in Hinblick auf funktionale Sicherheit, mechanische Limitierungen, Rechenzeit)? Unterliegt das System z. B. aufgrund des Anwendungskontexts besonderen regulatorischen Anforderungen?	LH-Std_15_ISO/IEC25010:2011_ISO LH-Std_16_ISO/IEC DIS 25059_ISO LH-FA_17_RequirementsEngineering_Vogelsang LH-FA_18_RequirementsEngineeringSafetyCritical_Martins Anwendungswissen/Methodenwissen
	1.2 Verfügbarkeit von Ressourcen	I1.2.1	Sind die Quellen für die Bereitstellung aller erforderlichen Daten bekannt und dokumentiert? Falls ja, ist sichergestellt, dass alle erforderlichen Daten rechtzeitig zur Verfügung gestellt werden können?	LH-BP_19_FAIRPrinciples_GO FAIR Initiative
		I1.2.2	Stellen die Quellen die Daten in einer dem Anwendungskontext angemessenen Sprache bzw. dem Anwendungskontext angemessenen Symbolen und Einheiten bereit?	LH-BP_19_FAIRPrinciples_GO FAIR Initiative Anwendungswissen
		I1.2.3	Stellen die Quellen Kontextinformationen bzw. Metadaten bereit, die die Daten interpretierbar und ihre Erfassung nachvollziehbar machen?	LH-BP_19_FAIRPrinciples_GO FAIR Initiative LH-BP_20_StandardsMetadaten_DCC Anwendungswissen
		I1.2.4	Liegt eine Abschätzung des minimal erforderlichen Datenvolumens für die Entwicklung des Systems vor? Wenn ja, kann es bereitgestellt werden?	Anwendungswissen/Methodenwissen OH
		I1.2.5	Liegt eine Abschätzung der angesichts der Entwicklungszeit minimal erforderlichen Rechenressourcen, Kommunikationsresssourcen und Bandbreiten vor? Wenn ja, können sie bereitgestellt werden?	Anwendungswissen/Methodenwissen

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
1. Design	1.3 Konzept für Leistungsbewertung	I1.3.1	Sind die technischen Gütekriterien des geplanten Systems definiert, priorisiert und dokumentiert?	LH-Std_15_ISO/IEC25010:2011_ISO LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftware-Test_ISO/IEC LH-Std_16_ISO/IEC DIS 25059_ISO LH-FA_22_MLQualityModel_Siebert
		I1.3.2	Sind ggf. weitere, nicht-technische Gütekriterien des Systems definiert, priorisiert und dokumentiert (z. B. Benutzerfreundlichkeit, Fairness, Erklärbarkeit)?	LH-Std_15_ISO/IEC25010:2011_ISO LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftware-Test_ISO/IEC LH-Std_23_ISO9241-1:1997_ISO LH-Std_16_ISO/IEC DIS 25059_ISO LH-FA_22_MLQualityModel_Siebert
		I1.3.3	Existiert ggf. ein etablierter Mindest- bzw. Referenzstandard für die Leistungsbewertung des Systems im Einsatzkontext? Wenn ja, wird er im Konzept für die Leistungsbewertung berücksichtigt?	LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftware-Test_ISO/IEC LH-FA_24_ReferenceStandardMedicine_Chen Anwendungswissen
		I1.3.4	Sind ggf. Community-Standarddatensätze oder Benchmark-Suites für den Leistungsvergleich des angstrebten Systems mit anderen Systemen vorhanden?	LH-Tool_25_PennMLBenchmarks_EpistasisLab LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftware-Test_ISO/IEC LH-FA_26_MLPerfTrainingBenchmark_Mattson
		I1.3.5	Werden für den Leistungsvergleich ggf. eigene problemspezifische Datensätze zusammengestellt, gepflegt und zur Verfügung gestellt?	LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftware-Test_ISO/IEC Anwendungswissen/Methodenwissen
	1.4 Problemformulierung	I1.4.1	Ist die geplante Zielfunktion/-setzung geeignet bzw. angemessen (z. B. Verlust- oder Kostenfunktion des Regressions- oder Klassifikationsproblems)?	LH-BP_27_LossFunctionOverview_Wang
		I1.4.2	Ist das geplante Lernverfahren geeignet bzw. angemessen?	LH-BP_28_AlgorithmOverview_Sarker LH-Tool_29_MSAzureMLSpickzettel_Microsoft LH-Tool_30_ScikitLearnFlowchart_scikit-learn
		I1.4.3	Ist der geplante Modelltyp bzw. die geplante Modellinstanz grundsätzlich geeignet bzw. angemessen?	LH-BP_28_AlgorithmOverview_Sarker LH-Tool_29_MSAzureMLSpickzettel_Microsoft LH-Tool_30_ScikitLearnFlowchart_scikit-learn

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
2. Entwicklung	2.1 Dokumentation der Entwicklungsziele	I2.1.1	Liegt eine Dokumentation der Entwicklungsziele für das System vor (z. B. in Form eines Lastenheftes oder eines Product Backlog)? Wenn ja, geht aus ihr eindeutig hervor, welchen Entwicklungsstand das System erreichen und wie es bereitgestellt werden soll? Wird die ggf. notwendige funktionale Sicherheit des Systems berücksichtigt?	LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE LH-Std_32_ISO/IEC/IEEE29148:2018_RequirementsEngineering_ISO/IEC/IEEE OH
	2.2 Datenvorverarbeitung und -exploration	I2.2.1	Ist dokumentiert, welche Schritte die Datenvorverarbeitung umfasst (Bereinigung, Integration, Transformation, Reduktion)? Wenn ja, sind die jeweils angewandten Verfahren und Operationen dokumentiert?	LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE LH-Tool_36_Luigi_SpotifyDataTeam LH-Tool_37_ApacheAirflow_ASF
		I2.2.2	Werden ggf. Maßnahmen ergriffen bzw. Werkzeuge eingesetzt, um Verzerrungen oder Unstimmigkeiten in den Daten zu identifizieren und zu beheben? Wenn ja, sind das allgemeine Vorgehen, die angewandten Verfahren und die Ergebnisse nachvollziehbar dokumentiert?	LH-Tool_38_AIFairness360Toolkit_IBM LH-BP_39_Fairlearn_FairlearnProject LH-BP_40_ProtocolDataExploration_Zuur LH-FA_41_SurveyBiasInML_Mehrabi LH-FA_42_SourcesOfHarm_Suresh
		I2.2.3	Werden ggf. Maßnahmen ergriffen bzw. Werkzeuge eingesetzt, um Störvariablen (engl. confounder) oder Stellvertretervariablen (engl. proxies) in den Daten zu identifizieren? Wenn ja, sind das allgemeine Vorgehen, die Verfahren und die Ergebnisse nachvollziehbar dokumentiert?	LH-BP_40_ProtocolDataExploration_Zuur LH-FA_43_ConfounderDiscovery_Rogozhnikov LH-FA_44_ConfoundingControlling_Dinga LH-FA_42_SourcesOfHarm_Suresh OH
		I2.2.4	Werden ggf. Maßnahmen ergriffen, um zu verhindern, dass aus dem Datensatz auf persönliche oder sensible Informationen geschlossen werden kann? Wenn ja, sind sie nachvollziehbar dokumentiert?	LH-Tool_45_OpenDataAnonymizer_ArtLabs LH-BP_46_AnonymisierungVerfahren_Dewes
		I2.2.5	Werden inhärente Qualitätsattribute der Daten wie Genauigkeit, Vollständigkeit, Konsistenz oder Aktualität erfasst und dokumentiert? Wenn ja, sind auch die jeweilig verwendeten Metriken spezifiziert?	LH-Std_47_ISO/IEC 25012:2008_DataQualityModel_ISO/IEC LH-BP_48_DimensionenDatenqualität_DAMA-NL LH-BP_49_DatenqualitätMetrikenDatenwirtschaft_Rohde

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
2. Entwicklung	2.3 Modellgenerierung- und überprüfung	I2.3.1	Ist der Prozess der Modellgenerierung (Training) einschließlich Modellselektion, -optimierung und Hyperparameter-Tuning dokumentiert? Wenn ja, liegen alle Informationen vor, um die Modellgenerierung reproduzieren zu können?	LH-Tool_50_MLflow_MLflowProject LH-BP_51_ModelEvaluationSelection_Raschka LH-FA_52_HyperparmeterOptimization_Yang LH-FA_53_EarlyStopping_Dodge OH
		I2.3.2	Ist der Prozess der Modellüberprüfung (interne Validierung) dokumentiert? Wenn ja, geht aus der Dokumentation hervor, dass die eingesetzten Methoden sowie die Partitionierung des Datensatzes in Trainings- und Testdaten angemessen und Datenleckagen ausgeschlossen sind?	LH-Tool_50_MLflow_MLflowProject LH-BP_51_ModelEvaluationSelection_Raschka LH-FA_54_OptimalDataSplit_Joseph OH
		I2.3.3	Werden im Rahmen der Modellgenerierung ggf. Maßnahmen ergriffen und dokumentiert, um unausgewogene, möglicherweise ungerechte bzw. diskriminierende Entscheidungen des Modells zu identifizieren oder zu vermeiden?	LH-Tool_50_MLflow_MLflowProject LH-Tool_38_AIFairness360Toolkit_IBM LH-BP_55_ErklärbareKI_Kraus LH-FA_41_SurveyBiasInML_Mehrabi LH-FA_56_FairMLTool_Adebayo
		I2.3.4	Liegt eine nachträgliche Analyse vor, die belegt, dass die Größe des zum Trainieren des Modells verwendeten Datensatzes ausreichend ist?	LH-FA_57_LearningCurves_Viering OH
		I2.3.5	Werden im Rahmen der Modellgenerierung ggf. gängige Best Practices aus der Softwareentwicklung wie Code Review, Continuous Integration und Versionskontrolle berücksichtigt?	LH-Std_15_ISO/IEC25010:2011_ISO LH-Tool_58_GitLab_GitLabInc. LH-BP_59_BestPracticesScientificComputing_Wilson LH-BP_60_HowToCodeReview_GoogleGithub
	2.4 Leistungsbewertung	I2.4.1	Ist die grundsätzliche Generalisierbarkeit des finalen Modells durch eine Überprüfung mit einem unabhängigen Datensatz belegt (externe Validierung)? Wenn ja, sind der Datensatz, das Vorgehen und das Ergebnis nachvollziehbar dokumentiert?	LH-FA_61_ExternalValidationProcess_Ho
		I2.4.2	Werden ggf. weitere anwendungskontext-spezifische Überprüfungen durchgeführt (z. B. Softwaretest, klinische Validierung)? Wenn ja, sind sie nachvollziehbar dokumentiert?	LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftware-Test_ISO/IEC
		I2.4.3	Werden Maßnahmen ergriffen bzw. Werkzeuge eingesetzt, um die Robustheit des Modells, (z. B. gegenüber kleinen Änderungen in den Eingabedaten) bewerten zu können? Wenn ja, werden die gegebenenfalls identifizierten Anfälligkeiten des Modells behoben und wird das Vorgehen dabei nachvollziehbar dokumentiert?	LH-Tool_62_AdversarialRobustnessToolkit_IBM LH-FA_63_RobustnessBenchmarks_Hendrycks
		I2.4.4	Unterliegt der Einsatz des Modells bzw. des Gesamtsystems im definierten Einsatzkontext Beschränkungen (z. B. aufgrund von identifizierten Schwachstellen)? Wenn ja, sind die Beschränkungen nachvollziehbar dokumentiert?	
		I2.4.5	Werden ggf. Usability-Tests durchgeführt, um die Gebrauchstauglichkeit des Gesamtsystems und das Produkterlebnis zu überprüfen?	LH-Std_23_ISO9241-1:1997_ISO LH-FA_64_AIandUX_Lew

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
2. Entwicklung	2.5 Funktionalität und Verlässlichkeit	I2.5.1	Ist die Funktionaliät des Systems im realen, praktischen Einsatzkontext belegt (z. B. durch betriebsnahes Testen)? Wenn ja, erfüllt das System alle in der Designphase festgelegten funktionalen Anforderungen?	LH-Std_65_AssuranceMetamodel_OMG LH-BP_66_SafetyAssuranceCases_JohnerInstitut LH-BP_67_ConditionalSafetyCertificates _FraunhoferIESE LH-FA_68_NeuralNetworkMethodsSafetyCritcialApplications_Adler
		I2.5.2	Ist die Verlässlichkeit des Systems im realen, praktischen Einsatzkonzept belegt (z. B. durch betriebsnahes Testen)? Wenn ja, erfüllt das System alle in der Designphase festgelegten Nebenbedingungen?	LH-Std_65_AssuranceMetamodel_OMG LH-BP_66_SafetyAssuranceCases_JohnerInstitut LH-BP_67_ConditionalSafetyCertificates _FraunhoferIESE LH-FA_68_NeuralNetworkMethodsSafetyCritcialApplications_Adler
	2.6 Dokumentation des Entwicklungsprozesses	I2.6.1	Liegt eine durchgängige Dokumentation der Entwicklungsphasen vor (einschließlich Datensammlung, Datenmanagement , Meilensteinen, Unterstützung durch Dritte und Qualitätskontrolle)?	LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE
		I2.6.2	Liegt ein Gebrauchshandbuch bzw. eine Anwendungsdokumentation vor?	LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE OH

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
3. Betrieb	3.1 Bedienbarkeit	I3.1.1	Ist definiert und dokumentiert, wer das System nutzen kann bzw. darf? Wenn ja, gehen etwaige Einschränkungen aus der Dokumentation eindeutig hervor?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI
		I3.1.2	Ist eine Interaktion der Nutzenden mit dem System zwingend erforderlich, um bei oder nach Inbetriebnahme seine Funktionalität sicherzustellen? Wenn ja, werden den Nutzenden Hilfen bzw. Hilfefunktionen zur Verfügung gestellt, die sie beim Umgang mit dem System unterstützen?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI
		I3.1.3	Sind optionale Interaktionen der Nutzenden mit dem System vorgesehen (z. B. manuelle Korrektur von Eingabedaten oder Zwischenergebnissen)? Wenn ja, werden den Nutzenden Hilfen bzw. Hilfefunktionen zur Verfügung gestellt, die sie beim Umgang mit dem System unterstützen?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI
		I3.1.4	Wird den Nutzenden eine Anleitung zur Problembehebung zur Verfügung gestellt? Wenn ja, ist sie verständlich?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI
		I3.1.5	Sieht das System ggf. eine Absicherung gegen übliche Fehler der Nutzenden vor?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI
		I3.1.6	Sieht das System bei Fehlfunktionen ggf. die Möglichkeit einer Wiederherstellung vorheriger System- bzw. Systemkonfigurationszustände durch die Nutzenden vor?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI
		I3.1.7	Kann die KI-Komponente des Systems im Notfall oder bei Bedarf sicher außer Betrieb genommen werden?	LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-MW_72_SOP_BIT.AI

Phase	Kriterium	Indikator-ID	Indikatoren	Lösungshilfen
3. Betrieb	3.2 Leistungsmonitoring	I3.2.1	Wird die Leistung des Systems im Betrieb kontinuierlich oder zumindest in regelmäßigen Abständen erfasst und dokumentiert? Wenn ja, wird den Nutzenden die Dokumentation zur Verfügung gestellt und werden sie ggf. über einen kritischen Abfall der Leistung informiert?	LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut LH-BP_74_EvaluationClinicalDecisionSupport_Duke-MargolisCenter LH-Std_75_AIModificationProposedRegulation_FDA LH-FA_76_ConceptDrift_Lu LH-FA_77_UncertaintyHandling_Bandyszak OH
		I3.2.2	Sieht das System eine automatisierte Aufzeichnung von im Betrieb eintretenden Ereignissen wie Fehlfunktionen, Problemen oder Statusabfragen vor?	LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut LH-BP_74_EvaluationClinicalDecisionSupport_Duke-MargolisCenter LH-Std_75_AIModificationProposedRegulation_FDA LH-FA_76_ConceptDrift_Lu LH-FA_77_UncertaintyHandling_Bandyszak
	3.3 Instandhaltung d. KI-Komponente	I3.3.1	Ist bei kritischem Leistungsabfall ggf. ein Re-Trainieren der KI-Komponente vorgesehen? Wenn ja, wird der Vorgang automatisch eingeleitet oder muss er von den Nutzenden angestoßen werden?	LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut LH-Std_75_AIModificationProposedRegulation_FDA LH-FA_76_ConceptDrift_Lu LH-FA_78_AggregationSchemes_Albarqouni
		I3.3.2	Wird die Leistung der re-trainierten KI-Komponente erfasst und fortlaufend dokumentiert? Wenn ja, gehen die Bewertungskriterien aus der Dokumentation hervor und sind sie angemessen? Sind für die Bewertung ggf. Eingaben der Nutzenden erforderlich?	LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut LH-Std_75_AIModificationProposedRegulation_FDA LH-FA_76_ConceptDrift_Lu LH-FA_78_AggregationSchemes_Albarqouni
	3.4 Dokumentation des Systems im Betrieb	I3.4.1	Ist die Dokumentation und Pflege einer Betriebshistorie vorgesehen oder ggf. aufgrund regulatorischer Rahmenbedingungen verpflichtend vorgeschrieben? Wenn ja, ist festgehalten, wer dafür verantwortlich ist und welche Aspekte sie umfasst (z. B. Qualitätskontrolle, Produktanpassung)?	OH

3.1 Kriterien Phase 0: Charakterisierung

Der Planung eines KI-Systems sollte stets eine Charakterisierung des Anwendungsumfeldes vorausgehen, um die grundsätzliche Machbarkeit eines Entwicklungsprojektes unter ethischen, technischen und wirtschaftlichen Gesichtspunkten bewerten zu können. Der Begriff Anwendungsumfeld umfasst dabei sowohl alle rechtlich-regulatorischen Gesichtspunkte, die sich durch den spezifischen Einsatzkontext eines KI-Systems ergeben, als auch die spezifischen Herausforderungen, die ein dem KI-System zugrunde liegendes Realsystem ggf. charakterisieren. Als Realsystem wird dabei ein gegebener, potenziell zeitabhängiger Prozess bezeichnet, der durch ein angestrebtes KI-System technisch nachgebildet werden soll bzw. für den das angestrebte KI-System möglichst genaue oder optimale Ausgaben in Form von z. B. Vorhersagen oder Aktionen generieren soll.

Wenn es sich bei den an einem KI-Projekt beteiligten Parteien um unabhängige Organisationen handelt, findet der Austausch hierzu zumeist unter einem gewissen Zeitdruck während der Anbahnung eines möglichen Geschäftsverhältnisses statt. Das heißt, es besteht zu diesem Zeitpunkt zumeist noch kein Vertragsverhältnis zwischen den Parteien. Es muss folglich zügig geklärt werden, ob eine Umsetzung eines KI-Systems grundsätzlich möglich ist und ob die entwickelnde Organisation auch tatsächlich über das methodische Portfolio verfügt, um die wesentlichen Anforderungen des jeweiligen Anwendungsumfelds auf angemessene Weise zu adressieren. Hierfür muss besonders die potenziell beauftragende Organisation deutlich herausstellen können, wo im Hinblick auf ein Entwicklungsprojekt die Besonderheiten und Herausforderungen liegen. Eine Vorstrukturierung dieses Dialogs in der Charakterisierungsphase ist sinnvoll, da Personen mit ausgewiesener Domänenexpertise und solche mit methodischer Expertise eine sehr unterschiedliche Auffassung davon haben können, was in diesem Zusammenhang als Besonderheit oder Herausforderung anzusehen ist. Für die Charakterisierung des Anwendungsumfelds eines KI-Systems wird daher ein Dialog entlang folgender Kriterien empfohlen:

i) Kritikalität, d. h. das Maß für potenzielle Gefahren, die vom Einsatz eines KI-Systems in einem spezifischen Anwendungskontext ausgehen können (Heesen et al. 2020)

ii) Einordnung der Komplexität (Einsatzkontext, Zeitverhalten des betrachteten Realsystems, Einschränkungen des Entscheidungsraums)

iii) Machbarkeit (Messbarkeit/Beobachtbarkeit, Akquirierbarkeit, Verwendbarkeit, Qualität, Formate von Informationen bzw. Daten, gegebenenfalls notwendige zusätzliche Wissensbasen, Kosten-Nutzen-Analyse)

3.1.1 Kritikalität

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	0.1 Kritikalität	0.2 Komplexität		0.3 Machbarkeit
Indikator	I0.1.1	I0.2.1–I1.2.3		I0.3.1–I0.3.7

Der von der EU-Kommission vorgelegte Entwurf für eine KI-Verordnung (Europäische Kommission 21.04.2021) enthält Vorschriften für das Inverkehrbringen, die Inbetriebnahme und die Verwendung von Systemen der künstlichen Intelligenz. Der Gesetzgeber folgt dabei einem risikobasierten Ansatz: Es werden nur dort regulatorische Belastungen festgelegt, wo Risiken für die Grundrechte von natürlichen Personen oder die Gefährdung von sensiblen Rechtsgütern erwartet werden. Bei der Klassifizierung als Hochrisiko-KI-System wird u. a. darauf abgestellt, in welchem Anwendungsumfeld das KI-System eingesetzt werden sollen. Der Einsatz von KI-Systemen in besonders sensiblen Bereichen führt dazu, dass diese KI als hochriskant eingestuft wird (z. B. Einsatz in kritischen Infrastrukturen oder in den Bereichen Bildung, Beschäftigung oder Strafverfolgung). Die vom Gesetzgeber vorgegebene Risikoklassifizierung und die Festlegung von konkreten Anwendungsszenarien können beispielgebend als Indikator für die Kritikalität des KI-Systems herangezogen werden (Indikator I0.1.1).

FIKTIVES BEISPIELSZENARIO

Ein Start-up, das bereits ein KI-gestütztes Entertainment-Analysesystem vertreibt und darüber kundenspezifische Empfehlungen für Musikstücke oder Spielfilme bereitstellt (Anwendungsfeld mit minimalem Risiko), sucht nach neuer Kundschaft in neuen Anwendungsdomänen.

Dabei zeigt sich: Mit dem unternehmenseigenen technischen Ansatz sind beispielsweise auch Eigenschaften von chemischen Wechselwirkungen unterschiedlicher Stoffe vorhersagbar, die ansonsten durch teure Experimente aufwendig bestimmt werden müssten. Eine Nutzung des KI-Ansatzes in der Produktentwicklung, z. B. bei der Entwicklung neuer Kunststoffe bzw. Polymermischungen, stellt kein erhöhtes Risiko dar. Das liegt daran, dass die Qualitätsprüfung der Materialien und Produkte unverändert nach etablierten Verfahren der Kunststoffprüfung und anwendungsspezifischen Prüfverfahren (z. B. im streng regulierten Flugzeugbau) durchgeführt wird.

Eine weitere Überlegung des Start-ups ist es, mit dem technischen Ansatz medizinisch nicht vorgebildeten Personen ein System bereitzustellen, das es auf Basis von selbsterfassten Krankheitssymptomen ermöglicht, Krankheitsbilder zuzuordnen. Nach einer Rechtsberatung verwirft das Start-up diese Überlegung, da es sich bei dieser sogenannten Selbstdiagnose (Aboueid et al. 2019) um ein Anwendungsumfeld mit zumindest begrenztem Risiko handelt. Das Start-up befindet, dass die verfügbaren Kapital- und Personalressourcen nicht ausreichend sind, um das Produkt rechtssicher zu gestalten (z. B. rechtssichere Ausformulierung der AGB, Umgang mit Haftungsrisiken oder gegebenenfalls notwendige Softwareanpassung bei Änderungen der Rechtslage).

Eine Nutzung des technischen Ansatzes zur KI-gestützten Prognose der Rückfallwahrscheinlichkeit von mutmaßlichen Straftäterinnen und Straftätern wird trotz theoretischer technischer Umsetzbarkeit von dem Start-up nicht erwogen. Der Umstand, dass der „AI-Act“ die Nutzung von KI zum Zweck der Strafverfolgung als Hochrisiko-Anwendung ansieht und besondere Prüfanforderungen vorsieht, wird vom Unternehmen als wesentliches Hemmnis empfunden.

LÖSUNGSHILFEN

- **LH-FA_5_RiskClassificationIEAI_IEAI**
  Das Institut für Ethik in der Künstlichen Intelligenz der Technischen Universität München stellt in einem Whitepaper einen ersten Ansatz für die Risikoklassifizierung nach KI-Verordnung der EU vor (Institute for Ethics in Artificial Intelligence (IEAI) 2022). Die Publikation bietet darüber hinaus eine umfassende Übersicht zu Risikobewertungsansätzen aus Wissenschaft und Industrie (Indikator I0.1.1).

- **LH-Std_3_AIRiskManagement_NIST**
  Das Nationale Institut für Standards und Technologie (NIST) der USA propagiert in einem aktuellen Entwurf seines Rahmenwerks für KI-Risikomanagement (engl. AI Risk Management Framework) eine dreigliedrige Taxonomie aus technischen sowie sozio-technischen Charakteristiken und Leitprinzipien für Identifikation und Management von KI-bezogenen Risikofaktoren (NIST 2022) (Indikator I0.1.1). Es werden direkte Vergleiche u. a. zur KI-Verordnung der EU (Europäische Kommission 21.04.2021) und den OECD-Empfehlungen zum Umgang mit KI (OECD 2019) gezogen.

- **LH-Std_1_ISO 14971:2019:RiskManagement_Medizinprodukt_ISO und LH-Std_2_ISO/TR22100-5:2021Maschinensicherheit_ISO**
  Gegebenenfalls können auch die folgenden Normen als Lösungshilfe für den Indikator I0.1.1 herangezogen werden:

  - ISO 14971:2019[10] (Anwendung des Risikomanagements auf Medizinprodukte) beschreibt einen Risikomanagementprozess, mit dem sichergestellt werden soll, dass die Risiken durch Medizinprodukte sowohl bekannt und beherrscht als auch im Vergleich zum Nutzen akzeptabel sind.

  - ISO/TR 22100-5:2021 (Sicherheit von Maschinen – Zusammenhang mit ISO 12100 – Teil 5: Auswirkungen des maschinellen Lernens mit künstlicher Intelligenz) beschreibt, wie Gefährdungen im Zusammenhang mit KI-Anwendungen innerhalb von Maschinensystemen im Rahmen eines Risikobeurteilungsprozess berücksichtigt werden können.

- **LH-Std_4_ISO/IEC_FDIS23894_RisikomanagementKI_ISO**
  Im Entwurf befindet sich aktuell die Norm ISO/IEC FDIS 23894 (Titel: Informationstechnik, Künstliche Intelligenz, Risikomanagement), die nach ihrer Fertigstellung und Etablierung als Orientierungshilfe für den Indikator I.0.1.1 dienen kann.

10 Erläuterungen zur ISO 14971:2019 siehe auch: https://www.johner-institut.de/blog/category/iso-14971-risikomanagement/

3.1.2 Komplexität

Das zweite Kriterium für die Charakterisierung der generellen Aufgabenstellung ist eine Einordnung der Komplexität des Anwendungsumfelds. Das Hauptziel ist dabei, alle besonderen potenziellen technischen Schwierigkeiten zu identifizieren, die sich aus dem spezifischen Einsatzkontext eines KI-Systems und dem ihm zugrunde liegenden Realsystem bzw. den Umwelteinflüssen, denen es unterliegt, ergeben. Dafür sollte zunächst betrachtet werden, ob im Hinblick auf den „Agenten" (z. B. KI-gestützte Software oder Mensch) folgende Aspekte zu berücksichtigten sind – gegebenenfalls auch getrennt voneinander:

i) Unsicherheiten des Einsatzkontexts (z. B. nicht vorhersehbare Witterungsänderungen, Eingaben unabhängiger Systeme oder menschlicher Entscheiderinnen und Entscheider) (Indikator I0.2.1)

ii) Inhärentes Zeitverhalten des betrachteten Realsystems (z. B. dynamische Wirkzusammenhänge, Alterungsprozesse von mechanischen Komponenten) (Indikator I0.2.2)

Welche Einflüsse auf ein Realsystem im konkreten Anwendungsfall, zumindest teilweise, antizipiert werden können und welche als unvorhersehbare Einflüsse angesehen werden, ist in der Praxis häufig eine Frage der Datenverfügbarkeit sowie der späteren Gesamtsystemauslegung und Modellierung[11]. Wenn beispielsweise mittels Kamera erkannt werden kann, dass eine Person, die nominell ein Kraftfahrzeug steuert, längere Zeit gar nicht auf die Straße schaut, kann ein verbundenes Fahrassistenzsystem entsprechende Warnungen geben oder Maßnahmen einleiten, um das Fahrzeug am Straßenrand zum Stehen zu bringen. Die Nutzung derartiger Sensorinformationen kann die Unsicherheit im Einsatzkontext zu einem gewissen Grad, jedoch nicht vollständig reduzieren. Der Umstand, dass mit Unsicherheiten im Einsatzkontext (im Beispiel: unvorhersehbare Entscheidungen der fahrzeugführenden Person oder von anderen Verkehrsteilnehmerinnen und Verkehrsteilnehmern) oder dynamischen Effekten (im Beispiel die Verlängerung des Bremswegs in Abhängigkeit von der Fahrzeuggeschwindigkeit und der Verkehrssituation) zu rechnen ist, ist aus Entwicklungssicht entscheidend und aus Anwendungssicht jedoch in den meisten Fällen eindeutig beurteilbar, ungeachtet der späteren Modellierung bzw. Untergliederung (siehe auch Orientierungshilfe zur Charakterisierung des Einsatzkontexts und des Zeitverhaltens des Realsystems).

Ebenfalls bedeutend für die systembezogene Einordnung des Anwendungsumfelds ist die Frage, ob der „Agent" (z. B. KI-gestützte Software oder Mensch) Einschränkungen des Entscheidungsraums berücksichtigen muss (Indikator I0.2.3). Solche Einschränkungen können etwa durch mechanische oder physikalische Grenzen gegeben sein, wie dem maximalen Lenkwinkel eines Fahrzeugs oder der Lichtgeschwindigkeit als maximal erreichbare Geschwindigkeit eines Partikels. Darüber hinaus können auch sicherheitsrelevante oder gesetzliche Vorgaben den Entscheidungsraum einschränken, wie z. B. die Maximaltemperatur eines chemischen Reaktors oder eine festgesetzte Höchstgeschwindigkeit im Straßenverkehr.

11 In der Systemtheorie werden äußere Einwirkungen von inneren Wirkungsverknüpfungen darüber unterschieden, dass Erstere über eine festzulegende „Systemgrenze" hinweg das System beeinflussen, siehe z. B. Bossel 1994, die gleichzeitig auch die Systemelemente von der umgebenden Umwelt abgrenzt.

Je nach Anwendungsfall kann sich das Ignorieren entsprechender Grenzen bei der Umsetzung eines KI-Systems sehr unterschiedlich auswirken. Im besten Falle hat es keine Auswirkungen, wenn ein KI-Modell mangels entsprechender Vorkehrungen unzulässige Zwischenwerte oder Ausgaben erlaubt. Das kann z. B. der Fall sein, wenn möglichst hohe Werte einer bestimmten Ausgabegröße stets zum optimalen Ergebnis führen und eine unpassend festgelegte Untergrenze nicht ins Gewicht fällt. Es kann auch tolerierbar sein, wenn unzulässige Systemausgaben bei einer automatisierten Prüfung oder der Nachprüfung durch eine reale Person stets entdeckt werden können und das vorläufige „Ausbleiben" der Weitergabe der Ausgabewerte keine Konsequenz auf nachgelagerte Prozesse hat. Handelt es sich jedoch um ein autonomes System, das ganz ohne das Zutun bzw. die Prüfung durch eine fachlich qualifizierte Person Entscheidungen trifft, sollte technisch sichergestellt werden, dass die Vorgaben eingehalten werden. Dies gilt insbesondere, wenn weitere Berechnungen von diesen Ergebnissen abhängen oder es eine Rückkopplungsschleife gibt (wie z. B. bei autonomen Fahrzeugen). Im Hinblick auf die Auswahl eines geeigneten technischen Ansatzes innerhalb von Regelkreisen bzw. zur Regelung von Prozessen kann es dann auch empfehlenswert sein, die Systemstabilität zu klassifizieren.

Orientierungshilfe zur Charakterisierung des Einsatzkontextes (Indikator I0.2.1) und des Zeitverhaltens des Realsystems (Indikator I0.2.2)

Der Einsatzkontext charakterisiert im Requirements Engineering oder Safety Engineering die Eigenschaften der unmittelbaren Umgebung eines (Real-)Systems, welche für dessen Verhalten relevant sind. Dabei setzt man implizit voraus, dass zwischen den inneren Vorgängen eines (Real-)Systems und äußeren Einwirkungen unterschieden wird. Laut Systemtheorie können äußere Einwirkungen zwar die Ausgaben eines betrachteten (Real-)Systems oder dessen Verhalten beeinflussen, es ist umgekehrt aber für das (Real-)System nicht möglich, die äußeren Bedingungen gezielt zu manipulieren (Bossel 1994).

Falls sich die Umgebungsbedingungen durch eine hohe Unsicherheit (eng. uncertainty) auszeichnen, beispielsweise aufgrund nicht antizipierbarer Aktionen, die gegebenenfalls von externen Entitäten (z. B. interagierende unabhängige Software-Systeme, Menschen) ausgelöst werden, sollte der Einsatzkontext in der Charakterisierungsphase besonders intensiv betrachtet werden. Wenn die Strategien der externen Entitäten nicht bekannt bzw. nachvollziehbar sind oder die von ihnen ausgelösten Aktionen nicht vorhergesagt werden können, stellt die Bewertung oder Quantifizierung der äußeren Einwirkungen jedoch eine große Herausforderung dar.

Im Safety Engineering spielt die Analyse von Unsicherheiten bei der Charakterisierung des Einsatzkontextes eine zentrale Rolle. Dabei ist es unerlässlich, bestimmte Annahmen über den Einsatzkontext zu treffen, damit auf dieser Basis Systeme entwickelt werden können, die nachweislich verlässlich arbeiten und insbesondere keine Personen- und Sachschäden verursachen. Falls der Einsatzkontext hinsichtlich gewisser Aspekte als ungewiss („uncertain") angenommen werden muss, ist es für das Safety Engineering zumeist notwendig, „Worst-Case"-Annahmen zu treffen (was in der Regel jedoch die Performanz des Systems beeinflusst). Es könnte beispielsweise für zukünftige autonom fahrende Fahrzeuge eine sinnvolle Strategie sein, in der unmittelbaren Nähe von Schulen und Kindergärten die Worst-Case-Annahme zu treffen, dass Kinder aus jedem nicht einsehbaren Bereich heraus mit einer vergleichsweise hohen Geschwindigkeit auf die Straße laufen. In diesen Bereichen sollte also sehr langsam gefahren werden und es sollte sichergestellt werden, dass im Notfall stets eine rechtzeitige Bremsung möglich ist. Auf der Autobahn oder auf Schnellstraßen ist hingegen nur in Ausnahmefällen mit Personen zu rechnen, die aus nicht einsehbaren Bereichen mit hoher Geschwindigkeit auf die Fahrbahn laufen. Daher ist hier die Annahme für den permanent möglichen Eintritt des entsprechenden Worst-Case-Szenarios deutlich weniger geeignet.

Man spricht von einem offenen (Real-)System, wenn der Zustand eines Systems mit der Systemumgebung verknüpft ist, d. h. wenn er von außen beeinflussbar ist (z. B. durch Eingaben unabhängiger (Software-)Systeme oder durch menschlich gesteuerte Prozesse). Kann keine Beeinflussung von außen stattfinden, handelt es sich um ein geschlossenes System.

Wenn dynamische Effekte im Hinblick auf das gegebene Realsystem eine wesentliche Rolle spielen, kann dies in der Konsequenz die Qualität und Nutzbarkeit von Daten sowie die technologischen Ausgestaltungsmöglichkeiten eines funktionalen KI-Systems deutlich einschränken. Es kann dann möglicherweise erforderlich werden, die Zeit oder das Zeitverhalten des Analyseobjekts bzw. des Realsystems ebenfalls als ordnende Struktur bei der Nachbildung der Prozesse und der entsprechenden Zuordnung der Daten zu berücksichtigen (u. a. auch über die Betrachtung von Zeitreihen anstatt zeitunabhängiger Daten). Wenn sich beispielsweise Wirkbeziehungen in einem Realsystem maßgeblich verändern und dies auf nicht messbare Alterungs- oder Abnutzungsprozesse von mechanischen Komponenten zurückzuführen ist, sollte man sich überlegen, ob man diese Effekte modellseitig ignorieren und das Modell diesbezüglich als statisch annehmen sollte. Dies kann nämlich zur Folge haben, dass sich die Modellqualität dadurch stark verringert. Andererseits kann das Berücksichtigen von dynamischen Effekten bei rein datengetriebenen Verfahren gegebenenfalls dazu führen, dass auch sehr große Datenmengen für eine Lösungsumsetzung nicht mehr ausreichen, obwohl sie für Lösungen unter Vernachlässigung der Dynamik (z.B. auch in „stationären" Prozessphasen) möglicherweise eine akzeptable Basis dargestellt hätten.

Andererseits kann es auch vertretbar sein, dynamische Prozesse gänzlich zu vernachlässigen, wenn diese z. B. sehr langsam ablaufen oder sie für die gestellte Aufgabe unwesentlich sind. Man spricht von einem statischen Verhalten eines Realsystems, wenn das System kein Zeitverhalten aufweist bzw. jeder Ausgangswert des Systems ausschließlich von dem zum gleichen Zeitpunkt anliegenden Eingangswert abhängt. Ein anschauliches Beispiel für ein System, was beispielsweise als statisch und zudem als geschlossen angesehen werden kann, ist ein gelöstes Sudoku-Rätsel. Zahlen dürfen in keiner Reihe, keiner Spalte und in keinem Block doppelt auftreten, weshalb hier über die Zeit keinerlei Veränderung möglich ist, auch unter der Berücksichtigung kosmologischer Zeitkonstanten. Weitere anschauliche Beispiele für Systeme, die dynamische bzw. statische Eigenschaften mit Offenheit bzw. Geschlossenheit kombinieren, sind etwa Compiler-Programme (dynamisch, geschlossen) oder Fahrzeuge im Stadtverkehr (dynamisch, offen)[12].

LÖSUNGSHILFEN:

- Für die Charakterisierung des individuellen Anwendungsumfelds ist eine Analyse des Einsatzkontextes, des betrachteten Analysegegenstands bzw. des gegebenen Realsystems und der zu beachtenden Einschränkungen des Entscheidungsraums erforderlich. Diese ist allerdings im höchsten Maße anwendungsspezifisch, sodass hier auf keine universelle Lösungshilfe außer dem Zurückgreifen auf Anwendungswissen verwiesen werden kann.

12 Die Beispiele sind den Folien einer Lehrveranstaltung des Software and Computational Systems Lab der Ludwig-Maximilians-Universität München zum Thema „Modellierung dynamischer und adaptiver Systeme" entnommen: https://www.sosy-lab.org/Teaching/2019-WS-Seminar-Modas/Modas1.pdf.

3.1.3 Machbarkeit

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	0.1 Kritikalität	0.2 Komplexität	0.3 Machbarkeit	
Indikator	I0.1.1	I0.2.1–I1.2.3	I0.3.1–I0.3.7	

Das dritte Kriterium zur Charakterisierung des Anwendungsumfelds ist eine Machbarkeitsüberprüfung unter besonderer Berücksichtigung

i) des Informationsgehalts, der Akquirierbarkeit, der Verwendbarkeit und der Qualität der Daten sowie ihres Formats (Indikatoren I0.3.1–I0.3.5),

ii) der Möglichkeit des Zurückgreifens auf zusätzliches Anwendungswissen, welches nicht bereits über die Daten verfügbar ist (Indikator I0.3.6),

iii) und der Bewertung des Kosten-Nutzen-Verhältnisses (Indikator I0.3.7).

Die Betrachtung des Gegenstands „Daten" (Indikatoren I0.3.1–I0.3.5) ist elementar, weil sich auf dieser Basis die generelle Machbarkeit, d. h. die Lösbarkeit der Aufgabe, bewerten lässt. Falls die Aufgabe als lösbar eingeschätzt wird, erleichtert die Betrachtung zudem die grundlegende Bewertung der Passfähigkeit von potenziell einsetzbaren technologischen Ansätzen. Es ist daher im besonderen Interesse der entwickelnden Organisation, dass die entsprechenden Angaben der potenziell beauftragenden Organisation möglichst belastbar sind.

Die folgenden datenbezogenen Anforderungen sind dabei von besonderer Relevanz (wenn sie nicht oder nur teilweise erfüllt werden können, sollte die potenziell beauftragende Organisation dies offen gegenüber der entwickelnden Organisation kommunizieren):

- Die Größen, die für den betrachteten Analysegegenstand oder das Realsystem von zentraler Bedeutung sind, sind im Idealfall über bereitgestellte Daten zugänglich bzw. direkt messbar oder zumindest über andere zugängliche Größen abschätzbar bzw. beobachtbar (Indikator I0.3.1). Beobachtbarkeit ist, stark vereinfacht, dann gegeben, wenn der Zustand eines Systems ausschließlich auf Basis der gemessenen bzw. vom System bereitgestellten Daten erfassbar ist (siehe auch fiktives Beispielszenario A). Die Systemtheorie unterscheidet zwischen linearen und nicht-linearen Systemen. Ein lineares System liegt vor, wenn das Ausgangssignal des Systems proportional zum Eingangssignal ist. Für nicht-lineare Systeme gilt diese Beziehung nicht, sie können daher wesentlich komplexer sein als lineare Systeme und die Bewertung der Beobachtbarkeit stellt unter Umständen eine Herausforderung dar.

- Die Daten können im Idealfall selbst erzeugt bzw. mit vertretbarem Aufwand akquiriert werden (z. B. durch Experimente oder Simulationen), ohne dass dafür teure oder personen- bzw. sachgefährdende Messungen notwendig sind (Indikator I0.3.2).

- Der Datennutzung stehen keine rechtlichen Hürden entgegen (z. B. im Hinblick auf Datenschutz, Informationssicherheit oder bestehender Lizenzen) (Indikator I0.3.3).

- Der Prozess der Datengenerierung und gegebenenfalls bestehende Messunsicherheiten (z. B. potenzielle Fehlerquellen, auftretende Messfehler/Störungen und ihre statistische Verteilung) sind im Idealfall nachvollziehbar dokumentiert (Indikator I0.3.4).

- Das Format, in dem die Daten zur Verfügung gestellt werden, ist dokumentiert (unstrukturiert, strukturiert oder semi-strukturiert) (Indikator I0.3.5).

FIKTIVES BEISPIELSZENARIO A:
Eine Organisation wird von der Fahrerin eines E-Rennwagens beauftragt, ein System für das Offroadrennen Rallye Dakar zu entwerfen, das den nicht messbaren Stromverbrauch auf 100 km in Echtzeit ermittelt. Das konventionelle, für die asphaltierte Straße ausgelegte System schätzt den Wert auf der Offroadstrecke nur sehr unzuverlässig bzw. unterschätzt diesen, weil der Schlupf der Reifen auf der Offroadstrecke nicht mitberücksichtigt wird. Der Stromverbrauch auf 100 km ist in diesem Einsatzkontext durch das konventionelle System nicht beobachtbar, weil dem System keine Informationen über tatsächlich zurückgelegte Streckenintervalle vorliegen, sondern nur die Reifenumdrehungen pro Zeiteinheit bei bekanntem Radius in die Berechnung eingehen. Die Organisation löst die Aufgabe, indem über eine externe Quelle GPS-Daten zur Verfügung gestellt und für die Berechnung verwendet werden, wodurch die entsprechenden Größen beobachtbar werden.

Das Fehlen von zusätzlichem Anwendungswissen, welches nicht bereits über die Daten verfügbar ist (Indikator I0.3.6), kann sich in der Design- und in der Entwicklungsphase als wesentliche Problematik erweisen und sogar zum Scheitern von KI-Projekten führen (siehe Beispielszenario B). Wenn der entwickelnden Organisation ein entsprechendes Defizit auffällt, sollte es ihr ermöglicht werden, Personen mit einschlägigem Anwendungswissen zu konsultieren und/oder mit existierenden Realsystemen (gegebenenfalls einschließlich verfügbarer (Lösungen auf dem neuesten Stand der Technik oder Simulationsmodelle) gezielt zu experimentieren.

FIKTIVES BEISPIELSZENARIO B:
Eine Organisation wird beauftragt, ein KI-gestütztes System für die Zustandsüberwachung und Instandhaltungsplanung einer komplexen Anlage in der Produktion zu entwerfen. Über die Zeitreihen von Messdaten hinaus werden dem Unternehmen auch die ausgeworfenen Fehlercodes der Anlage bereitgestellt. Es stellt sich heraus, dass die Ableitung von technischen Störungen der Anlage aus individuellen Fehlercodes nicht möglich ist, weil auch bei sehr unterschiedlichen Störungen unter Umständen der gleiche Fehlercode ausgeworfen wird. Es gibt aber Hinweise darauf, dass spezifische Kombinationen von Fehlercodes auf konkrete technische Störungen hindeuten. Da aber keine entsprechende technische Dokumentation vorliegt und die Personen mit einschlägiger Kenntnis der Anlage nicht verfügbar sind, verzögert sich das Entwicklungsprojekt um mehrere Monate.

Zuletzt sollte im Rahmen der Machbarkeitsüberprüfung unbedingt auch das Kosten-Nutzen-Verhältnis kritisch betrachtet werden (Indikator I0.3.7). Während der Nutzen einer KI-Anwendung meist aus der sehr individuellen Anwendungsperspektive analysiert wird, hängen die Entwicklungskosten wesentlich von den zuvor thematisierten Kriterien des Leitfadens für die Charakterisierungsphase ab. Der Entwicklungsaufwand steigt nämlich tendenziell mit der Kritikalität der Anwendung, mit der Komplexität des Anwendungsumfelds und mit den Einschränkungen im Hinblick auf den Gegenstand „Daten“.

LÖSUNGSHILFEN

- **LH-BP_6_Beobachtbarkeit_GoogleCloud und LH-FA_7_ObservabilityComplexSystems_Stigter**
  Zum Thema Beobachtbarkeit von Daten (Indikator I0.3.1) existieren einerseits Lösungshilfen aus dem DevOps-Bereich, die Anleitungen darstellen, wie der Zustand von Software-Systemen im Sinne eines Monitorings effektiv überwacht werden kann (z. B. von Google Cloud). Andererseits ist die Beobachtbarkeit von Zuständen dynamischer Prozesse ursprünglich ein Konzept bzw. eine Eigenschaft aus der Mess- und Regelungstechnik. Im Falle linearer Prozesse kann diese besondere Art der Beobachtbarkeit über die in Lehrbüchern beschriebene Rangbestimmung von sogenannten Beobachtbarkeitsmatrizen einfach überprüft werden. Für die aufwändigere Prüfung von größeren, nicht linearen Systemen gibt es aktuell vielversprechende Forschungsansätze, um deren Beobachtbarkeit zu ermitteln (Stigter et al. 2017).

- **LH-BP_8_PrivacyGovernance_HLEG**
  Die Assessment List for Trustworthy AI (ALTAI) umfasst u. a. Anforderungen an den Datenschutz und die Data Governance, die zur Einschätzung der rechtmäßigen Nutzung der Daten unterstützend hinzugezogen werden können (Indikator I0.3.3).

- **LH-BP_9_SOP_Hollmann**
  Die Fachpublikation von Hollmann et al. 2020 stellt einen standardisierten Arbeitsablauf für das Erstellen von Forschungsdokumentationen bzw. sogenannten Standard-Operating-Procedures vor. Das Verfassen einer Standard-Operating-Procedure bietet sich gerade im Hinblick auf die Datengenerierung an (Indikator I0.3.4). Die Publikation kann dabei unterstützen, dies umzusetzen.

- **LH-MW_10_DataFormats_IBM**
  Die Webseite von IBM bietet einen Überblick zu Datensatzformaten (Indikator I3.3.5) und ihren jeweiligen Vor- und Nachteilen.

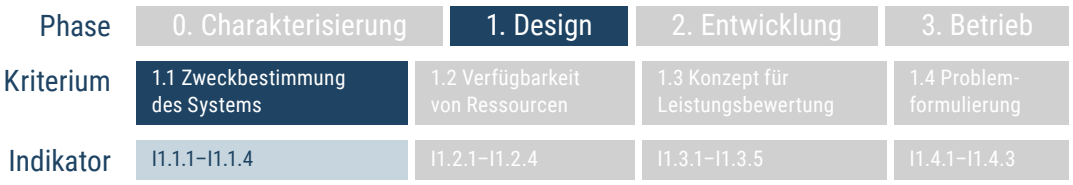
- **LH-BP_11_AIBusinessModelCanvas_Kerzel und LH-FA_12_CostBenefitMedicine_Ziegelmayer**
  Unter einem Business Model Canvas versteht man ein Werkzeug zur Visualisierung von Geschäftsmodellen bzw. der Geschäftslogik eines Angebots (u. a. Wertversprechen). Mittlerweile existieren Canvas-Vorlagen, die spezifisch für KI-Produkte angepasst wurden (z. B. Kerzel 2021). Diese können im Rahmen der Bewertung des Kosten-Nutzen-Verhältnisses unterstützend hinzugezogen werden (Indikator I0.3.7). Wie eine Kosteneffizienzbewertung konkret aussehen kann, zeigt die Fachpublikation von Ziegelmayer et al. 2022 anhand einer KI-Anwendung für das bildbasierte Screening von Lungenkrebs.

3.2 Kriterien Phase 1: Design

Vor dem Einstieg in die aktive Entwicklung eines KI-Systems sollte es sorgfältig designt bzw. geplant werden. Mängel oder Fehler im Design können mit einer Kostenerhöhung verbunden sein, den erzielbaren wirtschaftlichen Gewinn einschränken oder sogar zu einem wirtschaftlichen Verlust führen – insbesondere dann, wenn sie später in der bereits laufenden Entwicklungsphase zu einem Abbruch des Vorhabens führen.

Das Design erfordert einen engen Austausch zwischen den beteiligten Parteien zu strategischen Fragen wie z. B. was das angestrebte KI-System können soll, wo es eingesetzt werden soll, wer es bedienen wird, wie es technisch umgesetzt werden kann, was dazu gebraucht wird und ab wann es als Erfolg gelten kann. Handelt es sich bei den beteiligten Parteien um unabhängige Organisationen, besteht in der Designphase unter Umständen immer noch kein Vertragsverhältnis. Eine Vorstrukturierung des Dialogs zwischen den beteiligten Parteien ist daher auch für die Designphase grundsätzlich sinnvoll. Der Leitfaden empfiehlt für die Designphase eine Strukturierung entlang der folgenden vier Kriterien: Zweckbestimmung des Systems, Verfügbarkeit von Ressourcen, Konzept für Leistungsbewertung und Problemformulierung.

3.2.1 Zweckbestimmung des Systems



Der Begriff Zweckbestimmung ist bisher nicht einheitlich definiert und nicht klar von verwandten Begriffen wie „bestimmungsgemäßer Gebrauch" abgegrenzt. Grundsätzlich kann darunter aber die Festlegung verstanden werden, wofür ein Produkt oder sein Service konkret entwickelt und auf den Markt gebracht werden soll. Für KI-Systeme schließt dies sowohl die Beschreibung des Anwendungs- und gegebenenfalls Analysezwecks[13] (Indikator I1.1.1) als auch die Beschreibung der Anwendendengruppe und des Grads ihrer Interaktion mit dem System bzw. des Autonomiegrads des Systems ein (Indikator I1.1.2).

Eine sorgfältige Dokumentation dieser beiden Aspekte ist unter mehreren Gesichtspunkten relevant:

i) Richtlinien fordern sie als Teil der technischen Dokumentation gegebenenfalls zwingend ein, d. h. sie kann zulassungsrelevant sein (z. B. europäische Medizinprodukte-Verordnung (MDR)).

ii) Sie wird gegebenenfalls von bestimmten Normen verlangt (z. B. ISO 14971:2019 für die Anwendung des Risikomanagements auf Medizinprodukte).

iii) Sie bildet in der Regel die Grundlage für die Ableitung der Produktanforderungen; gegebenenfalls wird dies sogar von bestimmten Normen explizit gefordert (z. B. DIN EN 82304-1:2018-04 für Gesundheitssoftware).

13 Anwendungszweck und Analysezweck unterscheiden sich hinsichtlich der Detailtiefe. Ein Anwendungszweck wäre z. B. die Entwicklung einer KI zur Vorhersage von Komplikationen nach einer Knie-OP. Der Analysezweck könnte dies weiter ausführen: Vorhersage anhand eines kontinuierlichen Risiko-Scores zwischen 0 und 1, der anhand der demografischen Daten des Patienten/der Patientin und seiner/ihrer einen Tag vor der Operation erhobenen Vitalwerte ermittelt werden soll.

Im Hinblick auf die Anforderungen an ein KI-System sollten neben den rein funktionalen Anforderungen (z. B. „Das System soll die Nachfrage für Werkzeug X zum Stichtag Y vorhersagen"; Indikator I1.1.3) auch die einzuhaltenden Nebenbedingungen identifiziert werden (Indikator I1.1.4). Unter Letztere fallen sowohl nicht funktionale Anforderungen an das System wie Sicherheit, Interpretierbarkeit oder die Einhaltung regulatorischer Vorgaben als auch Neben- bzw. Randbedingungen (engl. constraints) wie Obergrenzen für die Rechenzeit oder die Kosten. Insbesondere die Festlegung der nicht-funktionalen Anforderungen an KI-Systeme stellt laut einer aktuellen Studie eine unter Umständen große Herausforderung dar (Habibullah und Horkoff 2021). Als Ursachen dafür werden u. a. ein fehlendes Bewusstsein auf Seiten der Kunden und eine noch mangelnde Erfahrung der Entwickelnden auf diesem Gebiet angeführt.

LÖSUNGSHILFEN

- **LH-BP_13_KIPeriodensystem_Bitkom**
  Der vom Informatiker Kristian Hammond stammende Entwurf eines Periodensystems für KI bietet eine erste Orientierungshilfe für die Verortung von KI-Anwendungen und kann zur Dokumentation von Anwendungs- und Analysezweck unterstützend hinzugezogen werden (Indikator I1.1.1).

- **LH-BP_14_AutonomiestufenIndustrie_Plattform Industrie 4.0**
  Die Plattform Industrie 4.0 hat 2020 in Zusammenarbeit mit der Plattform Lernende Systeme und der Begleitforschung zum Technologieprogramm Smart Service Welt II des BMWK einen Vorschlag für die Kategorisierung von Autonomiestufen von KI-Systemen im Kontext der industriellen Automation erarbeitet (Plattform Industrie 4.0 2020). Auf diese Kategorisierung kann für die Beschreibung der Art des Systems zurückgegriffen werden (Indikator I1.1.2).

- **LH-Std_15_ISO/IEC25010:2011_ISO und LH-Std_16_ISO/IEC DIS 25059_ISO**
  Die Identifizierung der funktionalen Anforderungen an das KI-System und der Nebenbedingungen sollte im Austausch zwischen dem Kunden bzw. Auftraggeber und den Entwickelnden erfolgen (Indikator I1.1.3, Indikator I1.1.4). Die ISO/IEC 25010:2011 für die Qualitätskriterien und Bewertung von Softwareprodukten (SQuaRE) kann dabei zumindest in einigen Aspekten gegebenenfalls unterstützen. Darüber hinaus liegt mit der ISO/IEC DIS 25059 bereits der Entwurf einer Erweiterung des SQuaRE-Modells vor, der spezifisch auf KI-Systeme zugeschnitten ist.

- **LH-FA_17_RequirementsEngineering_Vogelsang und LH-FA_18_RequirementsEngineeringSafetyCritical_Martins**
  Die Fachpublikation von Vogelsang und Borg 2019 stellt einen ersten Ansatz für das Anforderungsmanagement (engl. requirements engineering) von KI-Systemen vor. Im Kontext des Anforderungsmanagements sollten laut der Autoren auch methodenspezifische Aspekte wie Erklärbarkeit, Diskriminierungsfreiheit und spezifische Rechtsanforderungen berücksichtigt werden (Indikator I1.1.3, Indikator I1.1.4). Die Fachpublikation von Martins und Gorschek 2016 kann gegebenenfalls unterstützend hinzugezogen werden, sie bietet einen systematischen Überblick zum Anforderungsmanagement im Fall von sicherheitskritischen Systemen.

3.2.2 Verfügbarkeit von Ressourcen

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	1.1 Zweckbestimmung des Systems	1.2 Verfügbarkeit von Ressourcen	1.3 Konzept für Leistungsbewertung	1.4 Problem-formulierung
Indikator	I1.1.1–I1.1.4	I1.2.1–I1.2.4	I1.3.1–I1.3.5	I1.4.1–I1.4.3

Das Kriterium Verfügbarkeit von Ressourcen schließt Indikatoren ein, deren Nichterfüllung zu einem erhöhten Aufwand oder sogar einem Abbruch eines Vorhabens in der Entwicklungsphase führen können. Dazu zählt die Sicherstellung der rechtzeitigen, langfristigen, im Format angemessenen und im Volumen ausreichenden Bereitstellung von Daten und begleitenden Metadaten sowie die Gewährleistung von ausreichenden Rechen-, IT- und Kommunikations- sowie Bandbreitenkapazitäten (Indikatoren I1.2.1–I1.2.5). Die frühzeitige Berücksichtigung der (vorgeschlagenen) Indikatoren während der Design- und Planungsphase kann gegebenenfalls hohen wirtschaftlichen Verlusten vorbeugen, die drohen, wenn etwaige Mängel oder Engpässe erst im Zuge der Entwicklungsphase festgestellt werden (siehe Beispielszenario).

FIKTIVES BEISPIELSZENARIO

Für die Entwicklungsphase eines KI-Systems sind zwei Jahre und ein Budget von 1 Mio. EUR vorgesehen. Nach sechs Monaten stellt sich heraus, dass die Trainingszeit für das dem System zugrunde liegende Neuronale Netz zu lang ist, trotz Einsatz verfügbarer Cutting-Edge-Ansätze zur Reduzierung der Trainingszeit. Die intern verfügbare Rechenkapazität reicht nicht aus, um die Entwicklung fristgerecht abschließen zu können. Sie muss entweder erweitert werden oder durch Zukauf externer Ressourcen von Cloud-Computing-Anbietern ergänzt bzw. ersetzt werden. Im ersten Fall fallen zusätzliche Investitionskosten in Höhe von 100.000 EUR an, im zweiten zusätzliche Ausgaben für Mieten, die sich insgesamt auf 30.000 EUR belaufen. In beiden Fällen wird der gesetzte Budgetrahmen gesprengt. Die erste Option scheidet aus, da die erforderlichen Investitionen nicht kurzfristig aufgebracht werden können. Die zweite Option ist zwar finanziell machbar, der Kunde, für den das KI-System entwickelt wird, schließt aber eine Nutzung von Cloudressourcen grundsätzlich aus, da er um seine Daten besorgt ist. Nach sechs Monaten Entwicklungszeit kommt es daher zu einem Abbruch des Vorhabens, der mit einem Verlust von 250.000 EUR einhergeht.

Orientierungshilfe zur Abschätzung des minimal erforderlichen Datenvolumens (Indikator I1.2.4)

Eine grobe Vorabschätzung des Datenvolumens, das zum Trainieren einer spezifischen KI-Anwendung erforderlich ist (engl. pre-hoc sample size determination), ist in der Designphase aufgrund von zwei Aspekten von Bedeutung:

(1) Nachweis der grundsätzlichen Machbarkeit und gegebenenfalls Anpassung der Strategie. Wenn der Zeitraum für die Erfassung eines Datensatzes in einem Projekt z. B. maximal sechs Monate beträgt und in diesem Zeitraum maximal 6.000 Einträge aufgenommen werden können, für das Trainieren der Anwendung aber mindestens 60.000 Einträge erforderlich sind, dann ist das Projekt zumindest unter den vorgegebenen Bedingungen nicht machbar. Die Strategie sollte an diesem Punkt überdacht und angepasst werden, um das Projekt „machbar“ zu gestalten, z. B. durch Bezug von Daten aus zusätzlichen Quellen, synthetische Daten oder Datenerweiterungsverfahren (engl. data augmentation).

(2) Effizienz. Die Relation von Nutzen und Aufwand der Datenerfassung sollte angemessen sein. Wenn für ein Projekt ein Datensatz mit 6.000 Einträgen erforderlich ist, dann ist es voraussichtlich unwirtschaftlich, 60.000 Einträge zu sammeln oder, für den Fall, dass sie bereits vorliegen, mit 60.000 Einträgen zu arbeiten (u. a. Mehraufwand Datenvorverarbeitung). Existierende Ansätze aus dem Bereich Active Learning können gegebenenfalls hinzugezogen werden, um die repräsentativsten und informativsten Einträge für den Trainingsdatensatz auszuwählen (Du et al. 2017).

Ein einheitliches Vorgehensmodell für die Vorabschätzung des notwendigen Datenvolumens existiert noch nicht. Wie viele Einträge ein Datensatz mindestens enthalten sollte, hängt einerseits von der Komplexität des zu lösenden Problems und andererseits von der Komplexität des angestrebten (Machine-Learning-)Modells bzw. Algorithmus ab. Tabelle 3 fasst einige aktuell gängige Herangehensweisen und Ansätze aus der Literatur zusammen.

Modell/Algorithmus	Regel oder Kurzbeschreibung
FAUSTREGELN	
Regression Binäre logistische Regression	> 10 Datensatzeinträge pro Prädiktorvariable bzw. unabhängiger Variable (engl. events per variable criterion, EPV)
Computer Vision, Klassifizierung	1.000 Bilder pro Klasse
ANSÄTZE AUS DER LITERATUR	
Neuronale Netze, Bildklassifizierung	Baum und Haussler 1989: Worst-Case-Methode, die sicherstellt, dass mindestens ein bestimmter Anteil der Bilder bei einer gegebenen Anzahl an Datensatzeinträgen richtig klassifiziert wird.
Natural Language Processing (NLP)	Lauer 1995: Begrenzung der erwarteten Fehlerrate als Funktion des Volumens der Trainingsdaten.
Anomalieerkennunghttp-Attacken	Estepa et al. 2020: Kontinuierliche Erfassung von Indikatoren über den zeitlichen Verlauf der Datenerfassung. Aus den Zeitreihen der Indikatoren kann abgelesen werden, wann die Datenerfassung ausreichend ist.
Binäre logistische Regression	van Smeden et al. 2019: Anzahl der erforderlichen Datensatzeinträge als Funktion der Anzahl der Prädiktorvariablen, der Größe des Gesamtdatensatzes und des Anteils positiver Events.
Klinische Vorhersagemodelle (Regression)	(Riley et al. 2020): Leitfaden für binäre, kontinuierliche und Time-to-Event-Regressionsmodelle.
ANSÄTZE AUS DER STATISTISCHEN LERNTHEORIE	
Binäre Klassifizierung	Ermittlung der Anzahl der erforderlichen Trainingsdatensätze als Funktion der Vapnik-Chervonenkis(VC)-Dimension[14].

Tabelle 3 Gängige Herangehensweisen und Ansätze aus der Literatur

Im Rahmen der Planung von klinischen Studien werden routinemäßig Werkzeuge aus der Statistik zur Berechnung von Fallzahlen oder des erforderlichen Stichprobenumfangs eingesetzt. Man bezeichnet dies auch als A-priori-Poweranalyse (siehe Röhrig et al. 2010 für eine Einführung in die Methodik; siehe Sullivan für eine detaillierte Übersicht zu Tests und Formeln). Die Methoden lassen sich nicht direkt zur Vorabschätzung des erforderlichen Trainingsdatenvolumens für KI-Anwendungen übertragen, da Machine-Learning-Modelle im Gegensatz zu statistischen Tests

14 Die Vapnik-Chervonenkis(VC)-Dimension ist ein Maß für die Kapazität einer Klassifizierungsalgorithmusstatistik.

in der Regel nicht hypothesenbasiert sind bzw. mit ihnen keine Hypothese überprüft wird (eine Ausnahme bilden logistische Regressionsmodelle). Dennoch gibt es Szenarien, in denen es sinnvoll ist, in der Designphase im Rahmen der Überprüfung der Machbarkeit auf diese Methoden zurückzugreifen. Dies kann z. B. der Fall sein, wenn es sich bei der angestrebten KI-Anwendung um einen neuen diagnostischen Test handelt, dessen Leistung im Rahmen einer anschließenden klinischen Studie evaluiert werden soll[15]. Ein anderer Fall ist eine KI-Anwendung, die eine bereits bestehende Anwendung ersetzen soll. Dies setzt gegebenenfalls einen abschließenden Nachweis für eine zumindest vergleichbare Leistung voraus.

LÖSUNGSHILFEN

- Die Betrachtung des Kriteriums Verfügbarkeit von Ressourcen erfordert einen Austausch zwischen Personen mit Expertise i) in Bezug auf das Arbeitsumfeld (z. B. klinisches Personal, Maschinenherstellende) und ii) für die technische Umsetzung des Systems (z. B. Machine Learning Engineer). Erstere bringen das notwendige Anwendungswissen ein, um u. a. die Angemessenheit der Bereitstellung der (Meta-)Daten im Hinblick auf die verwendete Sprache, Symbole und Einheiten einzuschätzen (Indikatoren I1.2.1–I1.2.3).

- **LH-BP_19_FAIRPrinciples_GO FAIR Initiative und LH-BP_20_StandardsMetadaten_DCC**
  Im Rahmen der stakeholdergetriebenen GO FAIR Initiative (engl. für Findable, Accessible, Interoperable and Reusable) wurden grundlegende Prinzipien für Daten und Metadaten definiert, die als Hilfe hinzugezogen werden können, um die Angemessenheit der Bereitstellung der (Meta-)Daten zu beurteilen (Indikatoren I1.2.1–I1.2.3). Darüber hinaus können existierende Metadatenstandards genutzt werden. Das Digital Curation Centre (DCC) bietet hierfür einen nach Disziplinen durchsuchbaren Katalog an.

- Für die Abschätzung der erforderlichen Rechen-, IT- und Kommunikations- sowie Bandbreitenkapazitäten (Indikator I1.1.5) existieren aktuell noch keine etablierten Lösungsansätze, d. h. hier muss in der Regel vor allem auf das Methodenwissen der Expertinnen und Experten für die technische Umsetzung zurückgegriffen werden.

- Für die Vorabschätzung des erforderlichen Datenvolumens (Indikator I1.1.4) kann gegebenenfalls auf Faustregeln, allgemeine Ansätze aus der Statistik und Forschungsansätze aus den Anwendungsdomänen – insbesondere der Medizin – zurückgegriffen werden (siehe Orientierungshilfe zur Abschätzung des minimal erforderlichen Datenvolumens).

15 Eine Abschätzung des erforderlichen Stichprobenumfangs bzw. der Fallzahlen für die Studie kann in diesem Fall z. B. anhand der Sensitivität und Spezifität des Tests erfolgen.

3.2.3 Konzept für Leistungsbewertung

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	1.1 Zweckbestimmung des Systems	1.2 Verfügbarkeit von Ressourcen	1.3 Konzept für Leistungsbewertung	1.4 Problem-formulierung
Indikator	I1.1.1–I1.1.4	I1.2.1–I1.2.4	I1.3.1–I1.3.5	I1.4.1–I1.4.3

Die testgeleitete Entwicklung (engl. test-driven development, TDD) ist ein Designparadigma aus dem Softwarebereich. Es handelt sich dabei um eine Strategie, bei der zuerst die Tests für die Komponenten einer Software aufgestellt werden und die Komponenten selbst erst im Anschluss daran erstellt werden. Das heißt, die Arbeit am Quellcode beginnt erst dann, wenn feststeht, wie er im Zuge der Entwicklung überprüft und bewertet wird. Da sich diese Strategie bewährt hat, wird sie zunehmend auch für die Entwicklung von KI-Anwendungen empfohlen, u. a. zählt sie zu den Best Practices der IBM Garage Methodology (IBM Garage Methodology o. J.).

Im Kontext von KI wird der TDD-Ansatz auch als evaluationsgetriebenes Machine Learning bezeichnet (engl. evaluation-driven machine learning) (Maier 2021). Im Kern geht es darum, in der Designphase ein Konzept für die Leistungsbewertung eines KI-Systems zu erstellen, das die anschließende Entwicklungsphase leitet bzw. bestimmt. Ein solches Konzept sollte die Festlegung von technischen Güte- bzw. Erfolgskriterien umfassen (Indikator I1.3.1) einschließlich der Maßzahlen/Metriken, anhand derer die Erfüllung der Kriterien bemessen wird (z. B. Genauigkeit als Metrik für die Modellleistung). Ist die Metrik für die Modellleistung nicht richtig gewählt, kann eine KI-Anwendung auf dem Papier zwar gut aussehen, für den praktischen Einsatz aber völlig ungeeignet sein (siehe Beispielszenario). Unter Umständen kann es erforderlich sein, neben einer Primärmetrik weitere Metriken festzulegen, z. B. zur Erfassung der Zeit, die das Modell für die Ausgabe eines Ergebnisses oder einer Entscheidung benötigt. Dies ermöglicht ein gezieltes Vergleichen und Abwägen von Modellen in der späteren Entwicklungsphase – ist das Modell mit einer um 5 % erhöhten Genauigkeit aber dafür wesentlich längeren Ausgabezeit wirklich besser geeignet für den praktischen Einsatz oder nicht?

FIKTIVES BEISPIELSZENARIO

Die Leistungsbewertung einer KI-Anwendung aus dem Bereich Medizin erfolgt anhand eines Testdatensatzes, der Einträge von 100 Personen umfasst. 90 Einträge entfallen dabei auf gesunde Personen und 10 Einträge auf kranke Personen. Das Ergebnis der Klassifizierung der Probandinnen und Probanden ist:

(1) 80 der 90 gesunden Personen werden als gesund klassifiziert, die Zahl der richtig-negativen (engl. true negative, TN) beträgt somit η_{TN}=80.

(2) 10 der gesunden Personen werden als krank klassifiziert, die Zahl der falsch-positiven (engl. false positives, FP) beträgt somit η_{FP}=10.

(3) Nur eine kranke Person wird als krank klassifiziert, die Zahl der richtig-positiven (engl. true positives, TP) beträgt somit η_{TP}=1.

(4) 9 kranke Personen werden als gesund klassifiziert, die Zahl der falsch-negativen (engl. false negatives, FN) beträgt somit η_{FN}=9.

Die Genauigkeit der Anwendung ($((\eta_{TP}+\eta_{TN})/(\eta_{TP}+\eta_{TN}+\eta_{FN}+\eta_{FP})=0.81$) ist gut, spiegelt aber nicht wider, dass die KI-Anwendung nahezu versagt, kranke Personen auch als krank zu klassifizieren. Sie ist in diesem Fall nicht die richtige Metrik, da der Datensatz wesentlich mehr Einträge von gesunden als kranken Personen enthält und somit stark unausgeglichen ist. Eine angemessenere Metrik ist hier der Recall, also das Verhältnis aller richtig als krank identifizierten Personen zu allen tatsächlich kranken Personen ($\eta_{TP}/(\eta_{TP}+\eta_{FN})=0.1$).

Unausgeglichene Datensätze sind nicht nur im medizinischen Kontext oft die Regel, sondern auch in anderen Anwendungskontexten – auch eine Fertigungsanlage für Werkzeuge produziert deutlich mehr fehlerfreie Werkzeuge als fehlerhafte.

Im Rahmen des Konzepts für die Leistungsbewertung sollten auch nicht-technische Aspekte wie z. B. die Benutzerfreundlichkeit des KI-Systems und die Gerechtigkeit (engl. fairness) oder Erklärbarkeit seiner Ausgaben berücksichtigt werden (Indikator I1.3.2). Für die Bemessung dieser Aspekte stehen oft noch keine etablierten objektiven Maße bzw. Metriken zur Verfügung. Wie erklärbar ein KI-System ist, wird daher z. B. typischerweise im Rahmen von Nutzendenstudien evaluiert. Im Kontext von nicht-technischen Aspekten sollte das Konzept entsprechende methodische Überlegungen umfassen.

Darüber hinaus sollte geklärt werden, ob es Mindest- oder Referenzstandards gibt, die im Rahmen der Leistungsbewertung berücksichtigt werden müssen (Indikator I1.3.3). Im Bereich der medizinischen Diagnostik existieren z. B. oft sogenannte Goldstandards, die den zurzeit zuverlässigsten oder genauesten diagnostischen Test vorschreiben. Wenn eine KI-Anwendung einen Goldstandard ersetzen soll, muss sie gegebenenfalls vorher gegen ihn gebenchmarkt werden.

Für einen Leistungsvergleich mit bereits bestehenden KI-Anwendungen stehen gegebenenfalls Community-Standarddatensätze oder Benchmarking-Suites zur Verfügung. Wenn dies der Fall ist, sollten sie im Konzept für die Leistungsbewertung entsprechend berücksichtigt werden (Indikator I1.3.4).

Unter Umständen ist für eine spätere Leistungsbewertung auch die Selbsterfassung und Pflege von problemspezifischen Datensätzen notwendig (Indikator I1.3.5), z. B. falls keine Community-St andarddatensätze oder Benchmarking-Suites zur Verfügung stehen. Dabei kann es sich z. B. um Datensätze handeln, in denen gezielt sogenannte Edge- oder Corner-Cases gebündelt werden, d. h. Einträge, bei denen sich ein oder mehrere Attribut(e) an der extremen Grenze der Skala der Attribute bewegen.

LÖSUNGSHILFEN

- **LH-Std_21_ISO/IEC TR 29119-11:2020_AISoftwareTest_ISO/IEC**
  Der technische Bericht ISO/IEC TR 29119-11:2020 fasst Leitlinien für das Testen von KI-basierten Systemen zusammen, die bei der Auswahl und dem Design von Tests für die eigene KI-Anwendung unterstützen können (Indikator I1.3.1-5). Der Bericht soll perspektivisch durch die Technische Spezifikation ISO/IEC AWI TS 29119-11 ersetzt werden, die sich noch in Bearbeitung befindet.

- **LH-Std_15_ISO/IEC25010:2011_ISO und LH-Std_16_ISO/IEC DIS 25059_ISO**
  Die bestehende Norm ISO/IEC 25010:2011 und die im Entwurf verfügbare zukünftige Norm ISO/IEC DIS 25059 umfassen Qualitätsmodelle für Software- bzw. KI-Systeme, die zur Festlegung der Gütekriterien unterstützend hinzugezogen werden können (Indikator I1.3.1, Indikator I1.3.2).

- **LH-FA_22_MLQualityModel_Siebert**
  Die Fachpublikation von Siebert et al. 2022 stellt einen Forschungsansatz für den systematischen Aufbau eines anwendungsfallspezifischen Qualitätsmodells vor, das technische und nicht technische Gütekriterien berücksichtigt (Indikator I1.3.1, Indikator I1.3.2).

- **LH-Std_23_ISO9241-1:1997_ISO**
  Die Normenreihe ISO 9241-1:1997 umfasst Richtlinien für die Mensch-Computer-Interaktion, die unter Umständen für die Festlegung von Gütekriterien für die Benutzerfreundlichkeit hinzugezogen werden können (Indikator I1.3.2).

- **LH-FA_24_ReferenceStandardMedicine_Chen**
  In einem Kommentar im Fachjournal The Lancet geben Chen et al. 2021 einen Überblick über die Verwendung von Referenzstandards im Kontext von KI-Anwendungen bzw. ihrer Leistungsbewertung im medizinischen Bereich (Indikator I1.3.3).

- **LH-Tool_25_PennMLBenchmarks_EpistasisLab**
  Das Archiv Penn Machine Learning Benchmarks (PMLB) des Computational Genetics Lab der University of Pennsylvania stellt eine umfangreiche Sammlung kuratierter Benchmark-Datensätze für die Bewertung und den Vergleich von überwachten Machine-Learning-Algorithmen zur Verfügung (Olson et al. 2017) (Indikator I1.3.4). Die Datensätze decken ein breites Spektrum von Anwendungen und Lernverfahren ab.

- **LH-FA_26_MLPerfTrainingBenchmark_Mattson**
  Mit „MLPerf Training" steht eine Benchmark-Suite für das Trainieren von Machine-Learning-Modellen zur Verfügung (Mattson et al. 2019) (Indikator I1.3.4). Bei MLPerf handelt es sich um ein Konsortium international führender wirtschaftlicher und akademischer Akteure auf dem Gebiet KI, darunter u. a. Google, Intel, Nvidia, die University of California, Berkeley und die Harvard University.

3.2.4 Problemformulierung

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	1.1 Zweckbestimmung des Systems	1.2 Verfügbarkeit von Ressourcen	1.3 Konzept für Leistungsbewertung	1.4 Problem-formulierung
Indikator	I1.1.1–I1.1.4	I1.2.1–I1.2.4	I1.3.1–I1.3.5	I1.4.1–I1.4.3

Der Leitfaden schließt für die Designphase mit dem Kriterium „Problemformulierung". Darunter ist der Prozess zu verstehen, der die Zweckbestimmung des angestrebten Produkts/Services in ein Problem überführt, das mit KI bzw. Machine Learning angegangen und gelöst werden kann. Gesucht wird dabei die Formulierung, die im Hinblick auf die Zweckbestimmung, die verfügbaren Ressourcen und das Konzept für die Leistungsbewertung am geeignetsten ist.

Die Problemformulierung gliedert sich (grob) in drei Schritte (Indikatoren I1.4.1–I1.4.3):

i) die grundsätzliche formale Spezifikation des Problems durch Festlegung einer Zielfunktion (z. B. Regressionsproblem mit dem Optimierungsziel, die Summe der mittleren quadratischen Abweichungen bzw. der quadratischen Abstände zwischen Messwerten und Modellvorhersagen zu minimieren)

ii) die Auswahl des Lernverfahrens (z. B. überwacht, unüberwacht, teilüberwacht)

iii) die Auswahl des Modelltyps bzw. der Modellinstanz (z. B. Random-Forest-Regressionsmodell oder auf einem tiefen neuronalen Netz basierendes Klassifikationsmodell)

Eine unpassende initiale Ausrichtung der Problemformulierung in der Designphase kann sich später sehr negativ auswirken. Zu den möglichen Konsequenzen zählt u. a. ein Mehraufwand in der Entwicklungsphase, der gegebenenfalls monetär oder personell nicht erbracht werden kann. Darüber hinaus kann eine unpassende Zielfunktion ursächlich für ein späteres unsoziales oder diskriminierendes Verhalten eines KI-Systems im Betrieb sein (siehe Beispielszenario).

BEISPIELSZENARIO

Die frühen Bilderkennungssysteme mehrerer Anbieter, darunter Flickr und Google Photo AI, haben in der Praxis unsensible und diskriminierende Ergebnisse ausgegeben. Konzentrationslager wurden mit Labeln wie „Sport" oder „Klettergerüst" versehen und nicht weiße Personen wurden als „Affen" oder „Tiere" getaggt. Wie es genau zu diesen Fehlern kam, ist unklar.

In seinem Buch „Human Compatible: AI and the Problem of Control" argumentiert der Informatiker Stuart J. Russell, dass sich das Fiasko möglicherweise hätte verhindern lassen, wenn die Zielfunktion sensible Klassifikationsfehler stärker gewichtet hätte (Russell 2019). Das heißt, die Zielfunktion sollte nicht annehmen, dass die Kosten für eine falsche Klassifizierung einer Person als „Affe" die gleichen sind wie für jede andere Misklassifikation (z. B. Schnabeltasse als Ente).

LÖSUNGSHILFEN

- **LH-BP_27_LossFunctionOverview_Wang**
  Die Fachpublikation von Wang et al. 2022 gibt einen umfassenden Überblick zu 31 gängigen Ziel- bzw. Verlustfunktionen aus dem traditionellen Machine-Learning- und dem Deep-Learning-Bereich. Sie kann zur Bewertung der Angemessenheit der festgelegten Zielfunktion (Indikator I1.4.1) unterstützend hinzugezogen werden.

- **LH-BP_28_AlgorithmOverview_Sarker**
  Die Fachpublikation von Sarker 2021 gibt einen umfassenden Überblick über Lernverfahren und Machine-Learning-Algorithmen. Auf sie kann zur Bewertung der Eignung des gewählten Lernverfahrens (Indikator I4.1.2) und des Modelltyps (Indikator I4.1.3) zurückgegriffen werden.

- **LH-Tool_29_MSAzureMLSpickzettel_Microsoft und LH-Tool_30_ScikitLearnFlowchart_scikit-learn**
  Microsoft und scikit-learn bieten Flussdiagramme für die Auswahl von Machine-Learning-Algorithmen an, auf die ebenfalls zur Bewertung der Eignung des gewählten Lernverfahrens (Indikator I4.1.2) und des Modelltyps (Indikator I4.1.3) zurückgegriffen werden kann.

3.3 Kriterien Phase 2: Entwicklung

Die Entwicklungsphase von KI-Systemen folgt gewöhnlich keinem strikt linearen Prozess, da insbesondere der Aufbau der KI-Komponente durchaus experimentelle Züge aufweisen kann. Dem Weg zur optimalen KI-Komponente liegt zumeist ein iterativer Zyklus aus „Daten vorbereiten", „Modell trainieren", „Modell validieren" und „Modell testen" zugrunde. Eine zentrale Herausforderung für Entwicklerinnen und Entwickler ist es dabei, die Aufgaben und Arbeiten so aufzuteilen bzw. anzulegen, dass ein kontinuierlicher Fortschritt im Hinblick auf die Qualität der KI-Komponente bzw. des KI-Systems erzielt wird. Eine Grundvoraussetzung dafür ist, dass die Entwicklungsziele zuvor klar definiert wurden (siehe Designphase) und festgehalten sind. Sonst besteht die Gefahr, dass die Entwicklung in eine falsche Richtung läuft (Fehlentwicklung) oder in einer Sackgasse endet. Da zu diesem Zeitpunkt in der Regel ein Vertragsverhältnis zwischen den beteiligten Parteien besteht, hat dies unter Umständen nicht nur rein wirtschaftliche, sondern auch vertragsrechtliche Konsequenzen.

Der Leitfaden soll in der Entwicklungsphase dabei unterstützen, dies zu verhindern. Um einen möglichst strukturierten Fortschritt im Hinblick auf die Qualität der KI-Komponente bzw. des KI-Systems zu erzielen, empfiehlt er die Berücksichtigung von sechs Kriterien: Dokumentation der Entwicklungsziele, Datenvorverarbeitung und -exploration, Modellgenerierung und -überprüfung, Leistungsbewertung, Funktionalität und Verlässlichkeit sowie Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands.

3.3.1 Dokumentation der Entwicklungsziele

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb

Kriterium	2.1 Dokumentation der Entwicklungsziele	2.2 Datenvorverarbeitung und -exploration	2.3 Modellgenerierung und -überprüfung	2.4 Leistungsbewertung	2.5 Funktionalität und Verlässlichkeit	2.6 Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands
Indikator	I2.1.1	I2.2.1–I2.2.5	I2.3.1–I2.3.5	I2.4.1–I2.4.5	I2.5.1–I2.5.2	I2.6.1–I2.6.2

Grundlage für die Entwicklungsarbeit ist eine Dokumentation der Entwicklungsziele, in der alle in der Designphase spezifizierten Anforderungen an das KI-System verbindlich beschrieben sind (Indikator I2.1.1). Neben den funktionalen und nicht funktionalen Anforderungen (z. B. Verständlichkeit und Rechenzeit) sollte sie auch Aspekte wie Lieferbedingungen/-termine und Abnahmekriterien (z. B. Entwicklungsstand und Art der Bereitstellung) sowie die gegebenenfalls notwendige funktionale Sicherheit berücksichtigen (siehe Orientierungshilfe). Die Dokumentation ist unter zwei Gesichtspunkten wichtig:

i) Sie legt die Rahmenbedingungen für die Entwicklungsarbeiten fest.

ii) Sie kann der Vertragsgestaltung zwischen den beteiligten Parteien dienen.

In welcher Form die Dokumentation erfolgt, ist grundsätzlich freigestellt; ein gängiges Format sind Lastenhefte. Für letztere haben sich in einigen Branchen standardisierte Vorgaben und Vorgehensmodelle etabliert, die gegebenenfalls zu berücksichtigen sind.

Orientierungshilfe für die Berücksichtigung der funktionalen Sicherheit (Indikator I2.1.1)
Es gibt eine Reihe von Normen mit Bezug zum Thema funktionale Sicherheit, die (falls erforderlich) unterstützend zur Dokumentation der Entwicklungsziele hinzugezogen werden können (siehe Tabelle 4).

Norm	Kurzbeschreibung
IEC 61508:2010	Die internationale Normenserie umfasst die Entwicklung von elektrischen, elektronischen und programmierbaren elektronischen (E/E/PE) Systemen, die eine Sicherheitsfunktion ausführen. Sie betrachtet keine spezifischen Anwendungen, dient aber als Grundlage für die Implementierung in bestimmten Anwendungsgebieten. Zu den zentralen Elementen der Norm zählen Sicherheitsanforderungsstufen (engl. safety integrity levels) und Parameter für die Zuverlässigkeit.
ISO/IEC CD TR 5469	Die Norm mit dem Titel „Artificial intelligence — Functional safety and AI systems" befindet sich noch in der Entwicklung. Da sie die funktionale Sicherheit im spezifischen Kontext von KI-Systemen abdecken soll, wird sie perspektivisch von hoher Relevanz sein.
ISO/IEC FDIS 23894	Die Norm „Information technology — Artificial intelligence — Guidance on risk management" befindet sich ebenfalls noch in der Entwicklung. Sie soll Organisationen beim Umgang mit Risiken unterstützen, die sich während der Entwicklung von KI oder durch die Verwendung von KI ergeben.
VDE-AR-E 2842-61-2 ANWENDUNGSREGEL:2021-06	Die Anwendungsregel beschreibt ein Referenzmodell für vertrauenswürdige KI im Kontext von „autonom/kognitiven Systemen". Dies beinhaltet u. a. auch die Betrachtung der funktionalen Sicherheit.

Tabelle 4 Normen mit Bezug zum Thema funktionale Sicherheit

LÖSUNGSHILFEN

- **LH-Std_32_ISO/IEC/IEEE29148:2018_RequirementsEngineering_ISO/IEC/IEEE**
  Die Norm ISO/IEC/IEEE 29148:2018 stellt in Kapitel 9.6 die Inhalte und den Aufbau der sogenannten Software Requirements Specification (SRS) aus der Softwaretechnik dar. Die SRS wird oft mit einem Lastenheft gleichgesetzt, obwohl sie inhaltlich darüber hinausgeht (u. a. Pflichtenheft).

- **LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE**
  Die Norm ISO/IEC/IEEE 15289:2019 definiert Informationselemente für die Dokumentation im Bereich System- und Softwaretechnik. Neben dem Inhalt, der dokumentiert werden soll, beschreibt sie seine Aufteilung auf verschiedene Dokumente und wie diese Dokumente zu strukturieren sind. Sie spezifiziert in Kapitel 10 u. a. die Dokumentklasse „Specification" für die Dokumentation von Anforderungen.

- **LH-Std_33_VDE-AR-E 2842-61-2_AutonomeSysteme_VDE, LH-Std_34_ISO/IEC_CD_TR5469_FunktionaleSicherheitKI_ISO/IEC, LH-Std_35_IEC61508:2010_FunktionaleSicherheitE/E/PE_IEC** und **LH-Std_4_ISO/IEC_FDIS23894_RisikomanagementKI_ISO**
  Normen mit Bezug zum Thema funktionale Sicherheit (siehe Orientierungshilfe).

3.3.2 Datenvorverarbeitung und -exploration

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	2.1 Dokumentation der Entwicklungsziele	2.2 Datenvorverarbeitung und -exploration	2.3 Modellgenerierung und -überprüfung	2.4 Leistungsbewertung	2.5 Funktionalität und Verlässlichkeit	2.6 Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands
Indikator	I2.1.1	I2.2.1–I2.2.5	I2.3.1–I2.3.5	I2.4.1–I2.4.5	I2.5.1–I2.5.2	I2.6.1–I2.6.2

KI-Anwendungen sind auf die Ressource „Daten" angewiesen. Ebenso kann sich die Beschaffenheit der Daten direkt oder indirekt auf die Leistungsfähigkeit der KI-Komponente eines Systems auswirken. In der Praxis kann die Ursache für eine unzureichende Leistung oder ein Fehlverhalten einer KI-Anwendung tatsächlich sehr oft bis auf die Ebene der Daten zurückgeführt werden (siehe Beispielszenarien A und B im Rahmen der Orientierungshilfe). Das Kriterium Datenvorverarbeitung und -exploration schließt daher Indikatoren für die Behandlung der Ressource „Daten" ein. Unter Behandlung sind dabei alle organisatorischen, technischen, methodischen und konzeptionellen Maßnahmen zu verstehen, die ergriffen werden, um ein maximales Nutzungspotenzial der Daten zu gewährleisten oder das vorhandene Nutzungspotenzial der Daten auszuschöpfen:

i) typische Datenvorverarbeitungsschritte wie Bereinigung (z. B. Binning, Regression, Clustering), Integration (z. B. Schemaintegration), Transformation (z. B. Normalisierung, Konzept-Hierarchien) und Reduktion (z. B. Aggregation, Diskretisierung) (Indikator I2.2.1)

ii) die Identifizierung und gegebenenfalls Behebung von Verzerrungen und Unstimmigkeiten in den Daten (z. B. Bias oder Ausreißer; Indikator I2.2.2)

iii) die Identifizierung von Störvariablen/Confounder-Effekten und Stellvertretervariablen (engl. proxies) (Indikator I2.2.3; siehe Orientierungshilfe zu Stör- und Stellvertretervariablen)

iv) die Maskierung von persönlichen oder sensiblen Informationen (Indikator I2.2.4)

v) die systematische Erfassung inhärenter Datenqualitätsattribute wie Genauigkeit, Vollständigkeit, Konsistenz und Aktualität (Indikator I.2.5)

Orientierungshilfe zu Stör- und Stellvertretervariablen (Indikator I2.2.3)
Unter Störvariablen (engl. confounder) versteht man in der Statistik Variablen, die in einem Experiment nicht explizit berücksichtigt werden, sich aber trotzdem auf den Zusammenhang zwischen den unabhängigen und abhängigen Variablen des Experiments auswirken. Angenommen ein Forscher/eine Forscherin hat in einem Datensatz den Verkauf von Eistee und die Anzahl von Badeunfällen erfasst. Er/Sie stellt fest, dass die Anzahl der Badeunfälle mit dem Verkauf von Eistee positiv korreliert – je mehr Eistee, desto mehr Badeunfälle. Er/Sie glaubt nicht daran, dass Eistee das Risiko für einen Badeunfall erhöht, und schließt einen kausalen Zusammenhang aus. Die Störvariable, die eine Assoziation zwischen Eisteekonsum und Badeunfällen suggeriert, ist in diesem Fall die Temperatur: Im Sommer trinken die Menschen mehr Eistee und gehen auch öfter baden. Störvariablen können den Nutzen sowie die Interpretierbarkeit von KI-Anwendungen bzw. Machine-Learning-Modellen einschränken. Sie werden oft vor allem dann zum Problem, wenn sich die Zusammenhänge, die sie suggerieren, ändern (siehe Beispielszenario A).

Unter einer Stellvertretervariable bzw. einem Proxy versteht man eine Größe, die gemessen wird, um Auskunft über eine andere Größe zu erhalten, die selbst nicht direkt zugänglich bzw. messbar

ist. Das heißt, sie wird genutzt, um eine andere Größe abzuschätzen. Proxys werden im Kontext von KI-Anwendungen unter Umständen gezielt verwendet -- typischerweise, um ein abstraktes Konstrukt (z. B. Kreditwürdigkeit) überhaupt erst messbar zu machen (Operationalisierung). Sie werden dann zum Problem, wenn ihre Wahl nicht angemessen ist. Dies ist z. B. der Fall, wenn sie das abstrakte Konstrukt nur schlecht reflektieren oder eine Quelle für diskriminierende Ausgaben von KI-Anwendungen darstellen (siehe Beispielszenario B). Diskriminierende Ausgaben können aber auch auf Proxys zurückgehen, die in den Daten versteckt sind. Bekannte sensible Attribute wie das Geschlecht oder die Nationalität lassen sich zwar aus einem Datensatz entfernen („fairness-through-unawareness"-Ansatz), dies ist aber keine Garantie dafür, dass eine Diskriminierung ausgeschlossen ist (Datta et al. 2017). Wenn z. B. die Nationalität mit einem anderen, nicht per se sensiblen Attribut im Datensatz korreliert, wie dem Wohnviertel in einer Stadt, dann fungiert dieses Attribut als versteckter Proxy. Das heißt, die Information zur Nationalität ist weiterhin indirekt im Datensatz enthalten. Die Identifizierung von potenziellen Proxys für sensible Attribute kann daher als präventive Maßnahme im Hinblick auf Diskriminierung verstanden werden.

BEISPIELSZENARIO A

Forscherinnen und Forscher von Google haben 2008 mit Google Flue Trends (GFT) eine KI-Anwendung für die Gegenwartsvorhersage (engl. now-casting) von Grippeerkrankungen vorgestellt. Die Datengrundlage für die Vorhersage bildeten dabei die Begriffe, nach denen die Nutzerinnen und Nutzer der Google-Suchmaschine gesucht hatten. Die Anwendung wurde später stillgelegt, da sie spektakulär versagt hatte. Im Nachgang haben sich andere Wissenschaftlerinnen und Wissenschaftler dezidiert mit dem „Warum" des Versagens auseinandergesetzt (Lazer und Kennedy 2015). Dabei haben sie festgestellt, dass der Google-Algorithmus sehr anfällig für eine Überanpassung (engl. overfitting) im Hinblick auf saisonale Begriffe war, die in keinem kausalen Zusammenhang zur Grippe stehen, u. a. Highschool-Basketball. Darüber hinaus wurde nicht berücksichtigt, dass sich die Zusammenhänge in den Daten ändern, wenn sich das Suchverhalten der Nutzerinnen und Nutzer ändert.

BEISPIELSZENARIO B

Ein prominentes Beispiel für den gezielten Einsatz eines Proxy ist der Algorithmus des Unternehmens Optum, der die in der Vergangenheit angefallenen Gesundheitskosten als Proxy für den aktuellen Pflegebedarf von Patientinnen und Patienten genutzt hat. Dabei hat sich später herausgestellt, dass die Wahl des Proxy nicht geeignet war, da mit ihm eine Diskriminierung von nicht-weißen Patientinnen und Patienten einherging (Obermeyer et al. 2019). In den USA sind diese Gruppen wesentlich häufiger von Armut betroffen und können daher weniger Geld für die medizinische Versorgung ausgeben. Als Folge hat der Algorithmus diese Gruppen gesünder eingeschätzt als gleich kranke weiße Patientinnen und Patienten.

LÖSUNGSHILFEN

- **LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE**
  Die Norm ISO/IEC/IEEE 15289:2019 definiert Informationselemente für die Dokumentation im Bereich System- und Softwaretechnik. Neben dem Inhalt, der dokumentiert werden soll, beschreibt sie dessen Aufteilung auf verschiedene Dokumente und wie diese Dokumente zu strukturieren sind. Sie spezifiziert in Kapitel 10 u. a. die Dokumentklassen „Procedure" und „Report", die gegebenenfalls zur Dokumentation von typischen Datenvorverarbeitungsschritten/Prozeduren und ihren Ergebnissen hinzugezogen werden können (Indikator I2.2.1).

- **LH-Tool_36_Luigi_SpotifyDataTeam und LH-Tool_37_ApacheAirflow_ASF**
  Mit Luigi und Apache Airflow stehen Open-Source-Werkzeuge für das Aufstellen und Managen von Datenverarbeitungs-Pipelines bzw. Workflows zur Verfügung (Indikator I2.2.1).

- **LH-Tool_38_AIFairness360Toolkit_IBM und LH-BP_39_Fairlearn_FairlearnProject**
  Das AI Fairness 360 Toolkit von IBM bietet Werkzeuge und Metriken zur Identifizierung und gegebenenfalls Milderung von Verzerrungen/Bias in Datensätzen und Modellen an. Das Toolkit des Projekts Fairlearn umfasst ausschließlich Metriken für den Bias von Modellen (Indikator I2.2.2).

- **LH-FA_41_SurveyBiasInML_Mehrabi**
  Die Fachpublikation von Mehrabi et al. 2022 gibt einen umfassenden Überblick über die Arten von Verzerrungen/Bias, die sich auf KI-Anwendungen auswirken können, u. a. Bias in Daten. Darüber hinaus stellt sie eine Taxonomie für die aktuell gängigen Definitionen von Fairness vor (Indikator I2.2.2).

- **LH-BP_40_ProtocolDataExploration_Zuur**
  Die Fachpublikation von Zuur et al. 2010 stellt ein Protokoll für die Datenexploration vor, das dabei unterstützt, gängige statistische Probleme zu vermeiden (Indikator I2.2.2, Indikator I2.2.3).

- **LH-FA_42_SourcesOfHarm_Suresh**
  Die Fachpublikation von Suresh und Guttag 2021 stellt ein Rahmenwerk für „Schadensquellen" entlang des Lebenszyklus von Machine Learning vor, die sich später auf die Leistung entsprechender Anwendungen auswirken können. Dies schließt eine Betrachtung von Quellen in der Phase der Datenvorverarbeitung ein (Indikator I2.2.2, Indikator I2.2.3).

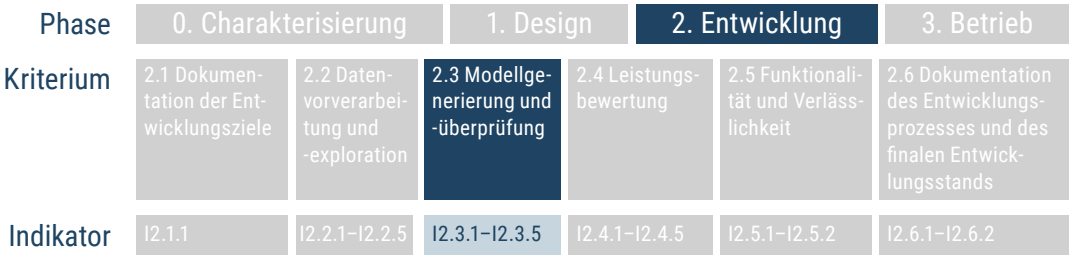
- **LH-FA_43_ConfounderDiscovery_Rogozhnikov**
  In der Fachpublikation von Rogozhnikov et al. 2022 wird eine statistische Methode für die hierarchische Identifizierung von Störvariablen in Rohdaten und in aus Machine Learning hervorgegangenen eingebetteten Daten vorgestellt (Indikator I2.2.3).

- **LH-FA_44_ConfoundingControlling_Dinga**
  Dinga et al. 2020 stellen in ihrer Fachpublikation Ansätze für den Umgang mit Störvariablen aus dem Bereich Neuroimaging vor (Indikator I2.2.3).

- **LH-Tool_45_OpenDataAnonymizer_ArtLabs und LH-BP_46_AnonymisierungVerfahren_Dewes**
  Mit dem open-data-anonymizer von ArtLabs steht eine Open-Source-Bibliothek für die Anonymisierung von tabellarischen Daten, Bilddaten und PDF-Daten zur Verfügung (Indikator I2.2.4). Der Leitfaden von Dewes 2022 bietet einen kurzen Überblick über rechtliche und mathematische Definitionen von Anonymität sowie relevante Anonymisierungsverfahren.

- **LH-Std_47_ISO/IEC 25012:2008_DataQualityModel_ISO/IEC und LH-BP_49_DatenqualitätMetrikenDatenwirtschaft_Rohde**
  Die Norm ISO/IEC 25012:2008 umfasst ein Modell für Datenqualität aus dem Bereich Softwaretechnik, das u. a. inhärente Datenqualitätsattribute einschließt (Indikator I2.2.5). Die Studie von Rohde et al. 2022, die Datenqualität und Qualitätsmetriken in der Datenwirtschaft betrachtet, kann zusätzlich unterstützend hingezogen werden.

3.3.3 Modellgenerierung und -überprüfung



Im Zentrum einer KI-Anwendung steht das ihr zugrunde liegende (Machine-Learning-) Modell. Der Weg zum Modell gliedert sich dabei in zwei Hauptprozesse: das Trainieren des Modells (engl. model training) und das Überprüfen des Modells unter Laborbedingungen (engl. internal model validation). Im Zuge des Trainingsprozesses lernt das Modell von den ihm zugeführten Daten (z. B. anhand von Bildern der Lunge zwischen einer Lungenentzündung und einer Covid-19-Erkrankung zu unterscheiden), um im Anschluss für bisher ungesehene, neue Daten Vorhersagen treffen zu können („Person X hat Covid-19, Person Y nicht"). Der Überprüfungsprozess dient dazu, festzustellen, wie gut das Modell gelernt hat bzw. wie gut seine Vorhersagen sind.

Der Weg zum finalen Modell ist aber in der Regel keineswegs geradlinig bzw. durch einen geradlinigen Prozess vorgegeben, sondern in Abhängigkeit von der angestrebten KI-Anwendung durchaus experimentell und individuell. Im Rahmen der Optimierung und Auswahl des Modells müssen viele Konfigurationsparameter und Methoden festgelegt werden (für eine Übersicht siehe Orientierungshilfe zu Konfigurationen und Methoden für die Modellgenerierung und -überprüfung). Aufgrund der vielen individuellen Gestaltungsmöglichkeiten ist eine lückenlose Dokumentation des Gesamtwegs bis zum finalen Modell zwingend erforderlich (Training: Indikator I2.3.1, Überprüfung: Indikator I2.3.2). Sie ist die Grundvoraussetzung für die Reproduzierbarkeit der Ergebnisse und perspektivisch auch für eine Überprüfung durch Dritte im Rahmen einer Zertifizierung. Der Prüfkatalog des Fraunhofer IAIS für vertrauenswürdige KI enthält bereits jetzt entsprechende Anforderungen (z. B. unter [VE-R-RE-MA-03] Wahl des Komponenten-Designs; (Poretschkin et al. 2021)).

Im Hinblick auf die Vertrauenswürdigkeit der KI-Anwendung sollten im Rahmen der Modellgenerierung und -überprüfung zusätzlich auch folgende Aspekte berücksichtigt werden:

i) Post-Training-Maßnahmen, um unausgewogene, möglicherweise ungerechte oder diskriminierende Ausgaben des Modells zu identifizieren und, falls möglich, zu beheben (Indikator I2.3.3)

ii) Post-Training-Analysen, die belegen, dass die Größe des zum Trainieren des Modells verwendeten Datensatzes ausreichend ist (Indikator I2.3.4; siehe Orientierungshilfe zur Post-Training-Bewertung der Angemessenheit der Größe des Trainingsdatensatzes)

iii) Umsetzung von gängigen Best Practices aus dem Softwarebereich wie z. B. Code Review, Versionskontrolle und kontinuierliche Integration[16]

16 Die kontinuierliche Integration (engl. continuous integration, CI) ist ein Verfahren, bei dem regelmäßig die Codeänderungen aller Entwicklerinnen und Entwickler zusammengeführt werden.

Orientierungshilfe zu Konfigurationen und Methoden für die Modellgenerierung und -überprüfung (Indikator I2.3.1, Indikator I2.3.2)

Die Prozesse Training und Überprüfung sind oft verschachtelt und werden im Rahmen der Modelloptimierung mehrfach durchlaufen, beispielsweise um verschiedene Konfigurationen von Hyperparametern (d. h. Modellparametern, die das Training steuern) zu testen oder verschiedene Hyperparameteroptimierungsstrategien zu verfolgen (u. a. Rastersuche, zufällige Suche, Bayes'sche Optimierung). Für die Optimierung eines Modells können zudem unterschiedliche Algorithmen eingesetzt werden (u. a. stochastic gradient descent, backpropagation, steepest descent, evolutionäre Algorithmen). Da es sich um iterative Verfahren handelt, müssen Terminierungs- oder Abbruchkriterien verwendet werden. Die Begriffe Abbruch- oder Terminierungskriterium beziehen sich darauf, ab wann bzw. bei welchem Ergebnis das Modell nicht weiter optimiert wird – also als nah genug am theoretischen Optimum angesehen wird. Im Fall von komplexen Machine-Learning-Modellen wie großen neuralen Netzen ist z. B. zu beobachten, dass die Leistung der Modelle ab einer gewissen Zahl von Iterationen/Epochen für den Trainingssatz zwar weiter zunimmt, für den Testsatz dagegen sinkt. Wenn dies der Fall ist, liegt eine Überanpassung (engl. overfitting) vor, d. h. das Modell lernt statistisches Rauschen. Um dies zu verhindern, muss das Training rechtzeitig abgebrochen werden (engl. early stopping). Es gibt verschiedene Ansätze, dies umzusetzen, z. B. kann die Zahl der Iterationen bzw. Trainingsepochen als Hyperparameter behandelt werden. Für jeden Wert des Hyperparameters wird dann ein Modell trainiert und anschließend das Modell gewählt, das die besten Ergebnisse für den Trainings- und den Testsatz erzielt. Dieser Ansatz ist gegebenenfalls rechenintensiv und zeitaufwendig. Ein anderer Ansatz ist daher, dass Modell einmal für eine große Zahl an Iterationen/Epochen zu trainieren und während des Trainierens nach jeder Epoche die Leistung des Modells für den Validierungssatz zu evaluieren. Sobald letztere degradiert, z. B. gekennzeichnet durch Zunahme des Verlusts (engl. loss) oder Abnahme der Genauigkeit, wird das Training abgebrochen. Es handelt sich somit um eine implizite Regularisierung.

Unter Umständen wird nicht nur ein Modelltyp bzw. eine Modellinstanz trainiert und optimiert, sondern mehrere. Dann müssen im Rahmen einer Modellselektion gegebenenfalls zusätzliche statistische Tests durchgeführt werden (u. a. McNemar-Test, Binomialtest, Cochran-Q-Test). Darüber hinaus muss die Überprüfungsmethodik festgelegt werden (z. B. einfaches Holdout, wiederholtes Holdout, Bootstrap, Kreuzvalidierung). Der initiale Datensatz muss entsprechend in zwei (Training, Test) oder drei (Training, Validierung, Test) angemessene Datensätze aufgeteilt werden. Eine umfassende, detaillierte Übersicht über die hier nur umrissenen Konfigurationen und Methoden sowie ihre Anwendung im Kontext von Modellgenerierung und -überprüfung bietet die Fachpublikation von Raschka 2020, die einer Best-Practice-Sammlung gleichkommt und als Lösungshilfe dient.

Orientierungshilfe zur Post-Training-Bewertung der Angemessenheit der Größe des Trainingsdatensatzes (Indikator I2.3.4)

Eine etablierte Methode für die Post-Training-Bewertung der Angemessenheit der Größe des verwendeten Trainingsdatensatzes sind sogenannte Performance-Lernkurven, die die Leistung des Modells als Funktion des Trainingsdatenvolumens abbilden (siehe Cho et al. 2016 für ein Beispiel aus dem Bereich Medical Image Classification). Welche Metrik in einer Performance-Lernkurve für die Leistung des Modells herangezogen wird (z. B. Genauigkeit, Präzision oder Recall), ist dabei nicht trivial. Die Genauigkeit ist z. B. nicht unbedingt die beste Metrik, wenn der Datensatz unausgeglichen (engl. unbalanced) ist, also die Anzahl der Instanzen pro Klasse eine große Differenz aufweist. In diesem Fall sollten andere Metriken wie die Präzision, der Recall oder der F1-Score verwendet werden (siehe auch Beispielszenario aus der Orientierungshilfe zu Indikator I1.3.1).

Wenn die Leistung eines Modells nicht zufriedenstellend ist oder hinter den Erwartungen zurückbleibt, können Lernkurven auch als analytisches Werkzeug eingesetzt werden, um der Ursache auf den Grund zu gehen. Sie können unter anderem dazu genutzt werden, den Effekt von mehr Trainingsdaten vorherzusagen oder die Rechenkomplexität des Trainingsvorgangs und der Hyperparameteroptimierung zu reduzieren. Eine detaillierte Übersicht zu Lernkurven, ihren Formen und ihrer Interpretation bietet die Publikation von Viering und Loog 2021, die sich unter den im Folgenden angeführten Lösungshilfen befindet.

LÖSUNGSHILFEN

- **LH-Tool_50_MLflow_MlflowProject**
  Es existieren Open-Source-Plattformen wie MLflow, die Komponenten anbieten, die eine lückenlose Dokumentation der Modellgenerierung und -überprüfung erleichtern (Indikatoren I2.3.1–I2.3.3), z. B. Nachverfolgung von Experimenten (MLflow Tracking) oder Organisation von Experimenten (MLflow Projects).

- **LH-BP_51_ModelEvaluationSelection_Raschka**
  Die Fachpublikation von Raschka 2020 gibt einen umfassenden und detaillierten Überblick zu Konfigurationen und Methoden für die Modellgenerierung, -überprüfung und -selektion (Indikator I2.3.1, Indikator I2.3.2). Sie ist anwendungsbezogen, spricht viele Empfehlungen aus (u. a. für kleine Datensätze) und kann daher als Best-Practice-Sammlung verstanden werden.

- **LH-FA_52_HyperparameterOptimization_Yang**
  Die Fachpublikation von Yang und Shami 2020 betrachtet das Thema Hyperparameteroptimierung aus theoretischer und praktischer Sicht. Sie stellt eine Orientierungshilfe für die Modellgenerierung dar (Indikator I2.3.1).

- **LH-FA_53_EarlyStopping_Dodge**
  Im Rahmen der Fachpublikation von Dodge et al. 2020 wird ein Ansatz für eine frühe Abbruchstrategie (engl. early stopping) beim Training im Kontext von Sprachmodellen vorgestellt und erläutert, wie dadurch nicht aussichtsreiche Wege der Modellanpassung früh in der Trainingsphase ausgeschlossen werden können (Indikator I2.3.1).

- **LH-FA_54_OptimalDataSplit_Joseph**
  Die Fachpublikation von Joseph 2022 stellt einen analytischen Ansatz zur optimalen Aufteilung von Datensätzen in einen Trainings- und einen Testsatz für die Modellüberprüfung vor (Indikator I2.3.2).

- **LH-Tool_38_AIFairness360Toolkit_IBM und LH-BP_39_Fairlearn_FairlearnProject**
  Für die Post-Training-Identifizierung von unausgeglichenen, ungerechten und gegebenenfalls diskriminierenden Ausgaben des Modells kann auf Toolkits wie das AI Fairness 360 Toolkit von IBM oder das Toolkit der Projekts Fairlearn zurückgegriffen werden (Indikator I2.3.3). Beide Toolkits bieten umfangreiche Werkzeuge und Metriken an.

- **LH-BP_55_ErklärbareKI_Kraus**
  Die Studie von Kraus et al. 2021 fasst den aktuellen Stand der Technik und zum Einsatz von erklärbarer KI (Explainable Artificial Intelligence, XAI) zusammen und erläutert ihn anhand praxisnaher Use Cases (Indikator I2.3.3).

- **LH-FA_56_FairMLTool_Adebayo**
  Mit FairML stellt Adebayo 2016 eine Toolbox für die Diagnose bzw. das Auditing von Bias in Black-Box-Machine-Learning-Modellen vor (Indikator I2.3.3).

- **LH-FA_41_SurveyBiasInML_Mehrabi**
  Die Fachpublikation von Mehrabi et al. 2022 gibt einen umfassenden Überblick über die Arten von Verzerrungen/Bias, die sich auf KI-Anwendungen auswirken können, u. a. Bias in Algorithmen. Darüber hinaus stellt sie eine Taxonomie für die aktuell gängigen Definitionen von Fairness vor (Indikator I2.3.3).

- **LH-FA_57_LearningCurves_Viering**
  Die Fachpublikation von Viering und Loog 2021 bietet einen umfassenden Überblick zu Lernkurven und ihrer Interpretation (Indikator I2.3.4). Lernkurven bilden die Leistung des Modells als Funktion des Trainingsdatenvolumens ab und stellen u. a. eine etablierte Methode für die Post-Training-Bewertung der Angemessenheit der Größe des verwendeten Trainingsdatensatzes dar (siehe auch Orientierungshilfe zur Post-Training-Bewertung der Angemessenheit der Größe des Trainingsdatensatzes).

- **LH-Tool_58_GitLab_GitLabInc**
  Mit GitLab steht eine sogenannte DevOps-Plattform zur Verfügung, die die Umsetzung von Best Practices aus dem Softwarebereich vereinfachen kann (Indikator I2.3.5).

- **LH-Std_15_ISO/IEC25010:2011_ISO, LH-BP_59_BestPracticesScientificComputing_Wilson** und **LH-BP_60_HowToCodeReview_GoogleGithub**
  Zur Bewertung der Umsetzung von gängigen Best Practices aus dem Softwarebereich (Indikator I2.3.5) kann gegebenenfalls auf die Norm ISO/IEC 25010:2011 (System- und Software-Qualitätsmodelle), die Best Practice von Google für Code Review (Google Engineering Practices Documentation) und die Best-Practice-Sammlung von Wilson et al. 2014 für Scientific Computing zurückgegriffen werden.

3.3.4 Leistungsbewertung

Phase	0. Charakterisierung	1. Design	2. Entwicklung		3. Betrieb
Kriterium	2.1 Dokumentation der Entwicklungsziele	2.2 Datenvorverarbeitung und -exploration	2.3 Modellgenerierung und -überprüfung	**2.4 Leistungsbewertung**	2.5 Funktionalität und Verlässlichkeit	2.6 Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands
Indikator	I2.1.1	I2.2.1–I2.2.5	I2.3.1–I2.3.5	**I2.4.1–I2.4.5**	I2.5.1–I2.5.2	I2.6.1–I2.6.2

Das Kriterium Leistungsbewertung umfasst Indikatoren für die Prüfung des KI-Modells bzw. des KI-Systems im Hinblick auf die Realbedingungen, unter denen es arbeiten soll. Eine Überprüfung unter Laborbedingungen allein (siehe vorheriges Kriterium) ist nicht ausreichend, wie mittlerweile viele Beispiele aus der Praxis belegen (siehe Beispielszenario). Bevor ein KI-System in den Betrieb geht, sollten daher präventive Maßnahmen ergriffen werden, um ein Fehlverhalten und Mängel hinsichtlich der Gebrauchstauglichkeit möglichst auszuschließen sowie potenzielle Schwachstellen zu beheben. Diese Maßnahmen sollten folgende Aspekte berücksichtigen:

i) die Leistung der KI-Komponente, wenn ihr neue, unabhängig erhobene Daten[17] zugeführt werden (Abschätzung der Fähigkeit zur Generalisierung durch eine sogenannte externe Validierung) (Indikator I2.4.1)

17 Unabhängig erhobene Daten meint hier Daten aus einer anderen Quelle im Hinblick auf die Modellgenerierung, z. B. Patientinnen und Patienten aus einer anderen Studie oder Bilder von einem anderen Gerät.

ii) die Leistung des Systems mit Blick auf den Anwendungskontext (z. B. durch eine Validierung, die belegt, dass ein substanzieller Mehrwert[18] vorliegt) (Indikator I2.4.2)

iii) die Robustheit der KI-Komponente gegenüber üblichen Störungen bzw. Perturbationen in den Eingabedaten, die aufgrund des Einsatzkontexts zu erwarten sind (z. B. Rauschen, Bildunschärfe oder Hintergrundgeräusche) (Indikator I2.4.3)

iv) die Benennung von möglichen Limitierungen des Systems (z. B. abgeleitet aus identifizierten Anfälligkeiten oder Schwachstellen hinsichtlich der Robustheit) (Indikator I2.4.4)

v) die Bewertung der Gebrauchstauglichkeit bzw. des Produkterlebnisses durch die Zielgruppe, die das KI-System praktisch nutzen soll (Indikator I2.4.5)

BEISPIELSZENARIO

Google Health hat eine KI-Anwendung für die Einschätzung einer durch Diabetes hervorgerufenen Erkrankung der Netzhaut des Auges entwickelt (diabetische Retinopathie), deren Leistung in der Labor-umgebung hervorragend ist (Beede et al. 2020). Der auf einem tiefen neuronalen Netz basierende Algorithmus zeigt eine Genauigkeit von über 90 % und Patientinnen und Patienten, die normalerweise mehrere Wochen auf die Beurteilung des Scans ihrer Netzhaut durch eine Fachspezialistin oder einen Fachspezialisten warten müssen, bekommen das Ergebnis nun innerhalb von weniger als zehn Minuten.

In Partnerschaft mit dem thailändischen Gesundheitsministerium hat das Unternehmen die Leistung der Anwendung unter Einbezug von elf Kliniken über einen längeren Zeitraum ausgiebig im Realeinsatz getestet. Das Untersuchungsergebnis zeigt, dass die Leistung der KI-Komponente und ihre Gebrauchstauglichkeit von diversen Faktoren abhängen, die zuvor nicht in Betracht gezogen wurden (Beede et al. 2020). Beispielsweise hat die Anwendung im Realbetrieb mehr als 20 % der Netzhautscans wegen zu geringer Bildqualität per se zurückgewiesen und gar nicht erst analysiert. Grund dafür war, dass die KI-Komponente der Anwendung anhand von hochqualitativen Scans trainiert wurde und die Anwendung explizit darauf ausgerichtet war, Scans ab einer bestimmten Qualität nicht zu akzeptieren.

LÖSUNGSHILFEN

- **LH-FA_61_ExternalValidationProcess_Ho**
  Die Fachpublikation von Ho et al. 2020 stellt sogenannte externe Validierungsansätze vor, mit denen die Fähigkeit zur Generalisierung von KI-Komponenten analysiert werden kann (konvergente und divergente Prozeduren). Darüber hinaus diskutiert sie, wie die Eignung von Datensätzen zur externen Validierung beurteilt werden kann (Indikator I2.4.1).

- **LH-Std_21_ISO/IEC TR 29119-11:2020_AISoftwareTest_ISO/IEC**
  Der technische Bericht ISO/IEC TR 29119-11:2020 fasst Leitlinien für das Testen von KI-basierten Systemen zusammen, die gegebenenfalls hinzugezogen werden können, um Tests für die Überprüfung mit Blick auf den Anwendungskontext zu entwerfen (Indikator I2.4.2). Der Bericht soll perspektivisch durch die Technische Spezifikation ISO/IEC AWI TS 29119-11 ersetzt werden, die sich noch in Bearbeitung befindet.

18 Ein substanzieller Mehrwert wäre im Medizinbereich z. B. dann gegeben, wenn eine klinische Validierung gezeigt hat, dass sich durch den Einsatz der KI-Anwendung die Therapieerfolgsrate verbessert.

- **LH-Tool_62_AdversarialRobustnessToolkit_IBM**
  Die Adversarial Robustness Toolbox (ART) v1.11 stellt Werkzeuge zur Verfügung, auf die Entwicklerinnen und Entwickler für die Analyse der Robustheit der KI-Komponente zurückgreifen können (Indikator I2.4.3).

- **LH-FA_63_RobustnessBenchmarks_Hendrycks**
  Mit IMAGENET-C und IMAGENET-P stellt die Fachpublikation von Hendrycks und Dietterich 2019 Benchmarks für die Analyse der Robustheit von neuronalen Netzen vor (Bildklassifikation). Dies beinhaltet die Robustheit gegenüber fehlerhaften/beschädigten Eingabedaten (IMAGENET-C) sowie Eingabedaten mit kleinen Störungen/Perturbationen (IMAGENET-P) (Indikator I2.4.3).

- **LH-FA_64_AIandUX_Lew und LH-Std_23_ISO9241-1:1997_ISO**
  Das Buch „AI and UX" (Lew und Schumacher 2020) betrachtet die Entwicklung von KI unter Berücksichtigung des Nutzungserlebnisses (engl. user experience, UX) und kann zur Bewertung der Gebrauchstauglichkeit bzw. des Produkterlebnisses durch die Nutzendengruppen hinzugezogen werden (Indikator I2.4.5). Eine mögliche weitere Unterstützung bietet die Normenreihe ISO 9241-1:1997, die Richtlinien für die Mensch-Computer-Interaktion umfasst.

3.3.5 Funktionalität und Verlässlichkeit

Phase	0. Charakterisierung	1. Design	2. Entwicklung	3. Betrieb
Kriterium	2.1 Dokumentation der Entwicklungsziele	2.2 Datenvorverarbeitung und -exploration	2.3 Modellgenerierung und -überprüfung	2.4 Leistungsbewertung	**2.5 Funktionalität und Verlässlichkeit**	2.6 Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands
Indikator	I2.1.1	I2.2.1–I2.2.5	I2.3.1–I2.3.5	I2.4.1–I2.4.5	**I2.5.1–I2.5.2**	I2.6.1–I2.6.2

Das Kriterium Funktionalität und Verlässlichkeit untersucht, ob die entsprechenden Eigenschaften eines Systems im realen, praktischen Einsatzkontext auch tatsächlich vorliegen und ob wirksame Maßnahmen getroffen werden, um die in der Designphase festgelegten funktionalen Anforderungen und Nebenbedingungen im Einsatzkontext zu erfüllen (Indikator I2.5.1, Indikator I2.5.2). Dabei unterscheidet sich dieses Kriterium von dem Kriterium Leistungsbewertung vor allem darin, dass es nicht ausschließlich die KI-Komponente, sondern deren Einbindung in ein möglicherweise vorhandenes Gesamtsystem bzw. den Einsatzkontext betrachtet. Dadurch wird häufig eine dedizierte Betrachtung der Risiken in Bezug auf die funktionale Sicherheit (Safety) notwendig.

Der konventionelle Ansatz zur Überprüfung der Einhaltung der Nebenbedingungen im realen Einsatzkontext ist ein (möglichst) allumfassendes Testen des KI-Systems unter Berücksichtigung (möglichst) aller denkbaren Szenarien, denen es aufgrund des Einsatzkontexts ausgesetzt sein könnte. Agiert das System in einer Umgebung, in der sich die relevanten Rahmenbedingungen nicht oder zumindest nicht wesentlich ändern, kann diese Herangehensweise durchaus aussichtsreich sein. Wenn jedoch der Einsatzkontext durch eine gewisse Unsicherheit gekennzeichnet ist bzw. sich die Umgebung wesentlich und oft ändert (z. B. viele unvorhersehbare Außeneinwirkungen oder denkbare Szenarien im Straßenverkehr), lässt sich ein solcher Ansatz nicht mehr praktisch umsetzen. Die Vielzahl an Kombinationen möglicher Szenarien kann dann schnell nicht mehr vollständig getestet werden, selbst wenn nur Worst-Case-Szenarien berücksichtigt werden.

Insbesondere bei KI-Systemen, deren Einsatzziel die Planung von Aktionen ist, geht in der Regel ein starker Verlust der Performanz bzw. der Funktionalität einher mit der Annahme vom (gleichzeitigen) Eintritt unterschiedlichster Worst-Case-Szenarien. Ein Alternativansatz kann es in solchen Anwendungsfällen sein, „sichere" Zustände und gestaffelte Risikostufen festzulegen und Zustände eines vergleichsweise niedrigen Risikos temporär zu tolerieren. Wenn dies anwendungsseitig vertretbar ist, leitet man bei einem solchen Ansatz erst dann ereignisabhängig risikomindernde Maßnahmen ein, wenn sich die Erhöhung eines Risikos abzeichnet.

Lösungshilfen

- **LH-Std_65_AssuranceMetamodel_OMG** und **LH-BP_66_SafetyAssuranceCases_JohnerInstitut**
  Um die Funktionalität und Sicherheit von Systemkomponenten in einem bestimmten Einsatzkontext nachzuweisen (Indikator I2.5.1, Indikator I2.5.2), werden z. B. im Gesundheitsbereich häufig explizite und angemessen spezifizierte Argumentationsstrukturen (engl. assurance cases) herangezogen (Picardi et al. 2019). Das Ziel ist es dabei, in überzeugender Weise und durch eine Reihe von Beweisen gestützt darzulegen, dass ein System die Anforderungen an Funktionalität und Verlässlichkeit erfüllt (in der Regel gegenüber einem Auditor/ einer Auditorin). Eine detailliertere Beschreibung, wie ein solches Vorgehen im Gesundheitsbereich umsetzbar ist, findet sich in Johner Institut 2020. Für das Safety Engineering gibt es ein ähnliches Vorgehen und ein – in den USA bereits standardisiertes – Metamodell (OMG 2021).

- **LH-BP_67_ConditionalSafetyCertificates _Fraunhofer IESE**
  Ein vielversprechender Ansatz für Systeme, deren Einsatzkontext durch eine hohe Unsicherheit charakterisiert ist, ist die Berücksichtigung eines dynamischen Risikomanagements und konditionaler Sicherheitszertifikate (engl. conditional safety certificates, ConSerts, siehe fraunhofer IESE). Der Ansatz sieht die Festlegung von Anforderungen und Garantien (engl. safety demands bzw. guarantees) in der Designphase vor. Gleichzeitig wird die Bewertung, ob sie erfüllt werden, zur Laufzeit vorgenommen. Dies ermöglicht es, situativ auf Veränderungen von Umgebungsbedingungen und Systemzuständen zu reagieren sowie insbesondere auch eine mögliche Kollaboration zwischen interagierenden Systemen in der Arbeitsumgebung auszunutzen[19] (Indikator I2.5.1, Indikator I2.5.2).

- **LH-FA_68_NeuralNetworkMethodsSafetyCritcialApplications_Adler**
  Für die Fachpublikation von Adler et al. 2019 wurden über eine Literaturrecherche zahlreiche Methoden identifiziert, die im Falle sicherheitskritischer Anwendungen für die Absicherung von auf neuronalen Netzen basierenden KI-Systemen in Frage kommen (Indikator I2.5.1, Indikator I2.5.2).

19 Interessante Webinare zu dem Thema sind hier zu finden: https://www.iese.fraunhofer.de/de/seminare_training/webinare.html#Safety-Engineering

3.3.6 Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands

Phase	0. Charakterisierung		1. Design	2. Entwicklung		3. Betrieb
Kriterium	2.1 Dokumentation der Entwicklungsziele	2.2 Datenvorverarbeitung und -exploration	2.3 Modellgenerierung und -überprüfung	2.4 Leistungsbewertung	2.5 Funktionalität und Verlässlichkeit	2.6 Dokumentation des Entwicklungsprozesses und des finalen Entwicklungsstands
Indikator	I2.1.1	I2.2.1–I2.2.5	I2.3.1–I2.3.5	I2.4.1–I2.4.5	I2.5.1–I2.5.2	I2.6.1–I2.6.2

Die Entwicklungsphase sollte mit einer finalen Dokumentation der durchgeführten Entwicklungsarbeiten sowie des erreichten Entwicklungsstands schließen. Dies ist unter drei Gesichtspunkten wichtig:

i) Die Dokumentation ist perspektivisch für eine Überprüfung durch Dritte im Rahmen einer Zertifizierung relevant.

ii) Die Dokumentation selbst ist gegebenenfalls vertraglich vereinbart oder ihre Inhalte sind im Hinblick auf die Erfüllung des Vertrags relevant (z. B. hinsichtlich Entwicklungsstand, Abnahmekriterien oder Art der Bereitstellung).

iii) Inhalte der Dokumentation werden gegebenenfalls zur Erstellung zulassungsrelevanter Dokumente benötigt, beispielsweise für die technische Dokumentation für Medizinprodukte, Geräte oder Anlagen (HighDoc; Johner Institut 2021b).

Neben der Dokumentation selbst (Indikator I2.6.1) sollte auch eine Anwendungsdokumentation oder ein Gebrauchshandbuch bzw. eine Gebrauchsanweisung erstellt werden (Indikator I2.6.2; siehe Orientierungshilfe für die Anwendungsdokumentation). Im medizinischen Bereich existieren z. B. Gesetze, Verordnungen und Normen, die nicht nur das Anfertigen einer Anwendungsdokumentation vorschreiben, sondern auch dezidierte Anforderungen an sie stellen[20] (u. a. europäische Verordnung für Medizinprodukte (MDR) oder In-vitro-Diagnostika (IVDR)). Die Anwendungsdokumentation ist darüber hinaus grundsätzlich wichtig, um Haftungsrisiken zu reduzieren (z. B. Ausschluss von Haftung bei unsachgemäßem Gebrauch).

Orientierungshilfe für die Anwendungsdokumentation (I2.6.2)
Geht es um die Erstellung einer Anwendungsdokumentation für ein KI-System, besteht oft Unklarheit hinsichtlich des Inhalts und des Umfangs. Grundsätzlich lassen sich jedoch die aus der Softwareentwicklung bekannten Grundsätze zur Anwendungsdokumentation auch auf KI-Systeme übertragen. Die Anwendungsdokumentation dient dazu, die Nutzerinnen und Nutzer zu befähigen, das KI-System entsprechend der Zweckbestimmung sachgemäß einzusetzen. Gleichzeitig trägt die Anwendungsdokumentation auch dazu bei, einen möglichen Fehlgebrauch des Systems zu verhindern. Eine detaillierte Beschreibung zur Überwachung, Funktionsweise und Kontrolle des KI-Systems, aber auch Erläuterungen hinsichtlich der Fähigkeiten und Leistungsgrenzen sowie des Genauigkeitsgrads sind nicht nur notwendig, sondern zur Reduzierung von Haftungsrisiken auch geboten.

20 Für eine Übersicht siehe https://www.johner-institut.de/blog/regulatory-affairs/gebrauchsanweisungen/

Die KI-Verordnung der EU (Europäische Kommission 21.04.2021) führt im Anhang IV allgemeine Mindestanforderungen für die technische Dokumentation von KI-Systemen auf. Zu beachten ist hierbei, dass diese Anforderungen nur für Hochrisiko-KI-Systeme gelten, also KI-Systeme, die mit einem hohen Risiko assoziiert werden (siehe auch Abschnitt 3.1.1).

Daneben existieren eine Reihe von allgemeinen und fachspezifischen Normen, die sich mit dem Thema Anwendungsdokumentation befassen, die in Tabelle 5 zusammengefasst sind.

Norm	Kurzbeschreibung
DIN EN 62079:2001 https://www.beuth.de/de/norm/din-en-62079/44038337 IEC/IEEE 82079-1 https://www.beuth.de/de/norm/din-en-iec-ieee-82079-1/342226844	Die Norm enthält Grundprinzipien für das Erstellen von Anleitungen und macht Vorgaben hinsichtlich der Gliederung, des Inhalts und der Darstellung. Die Norm wurde mittlerweile überarbeitet und in die Norm IEC/IEEE 82079-1 überführt. Letztere adressiert uneingeschränkt kleine sowie große und komplexe Produkte und enthält Angaben zur Erstellung, Gliederung und Darstellung von Anleitungen.
ISO/IEC/IEEE 15289:2019 https://www.iso.org/standard/74909.html	Die Norm trägt den Titel „Systems and software engineering — Content of life-cycle information items (documentation)“ und enthält Handlungsanweisungen, die bei der Erstellung von Anwendungsdokumentationen berücksichtigt werden sollen. Notwendig sind demnach u. a. Angaben zur Zweckbestimmung des Systems, Hinweise, Warnhinweise sowie Informationen zur Betriebsumgebung. Hinsichtlich des Inhalts von Anwendungsdokumentationen wird auf die Norm ISO/IEC 26514 verwiesen.
ISO/IEC 26514 https://www.iso.org/standard/77451.html	Die Norm enthält Vorgaben zum Erstellungsprozess und zu weiteren Aspekten, wie etwa Informationsgehalt und Präsentationsformat von Benutzungsdokumentationen.

Tabelle 5 Normen mit Bezug zur Anwendungsdokumentation

LÖSUNGSHILFEN

- **LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE**
  Für die formale Gestaltung der Dokumentation der durchgeführten Entwicklungsarbeiten/des erreichten Entwicklungsstands (Indikator I2.6.1) sowie die Gestaltung einer Anwendungsdokumentation (Indikator I2.6.2) kann auf die in der Norm ISO/IEC/IEEE 15289:2019 für Software definierten Dokumentklassen und Dokumenttypen zurückgegriffen werden.

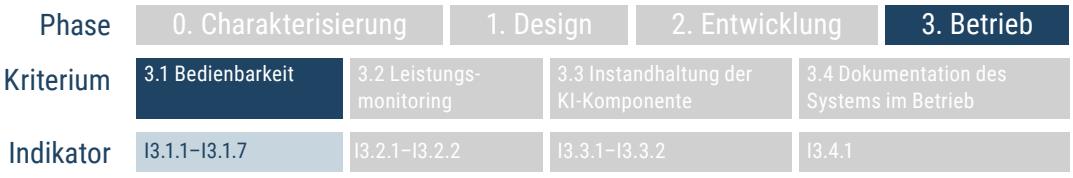
- **LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE**
  Es existieren mehrere Normen, die bei der Erstellung einer Anwendungsdokumentation unterstützen können (Indikator I2.6.2; siehe Tabelle 5 in der Orientierungshilfe zu Anwendungsdokumentationen für eine Übersicht). Die Aspekte Informationsgehalt und Präsentationsformat werden von der Norm ISO/IEC/IEEE 26514:2022 abgedeckt.

3.4 Kriterien Phase 3: Betrieb

Seine zumindest vorläufig letzte Lebensphase erreicht ein KI-System, wenn es in der Praxis ankommt und in Betrieb genommen wird. Vorläufig deshalb, da sich im laufenden Betrieb durchaus zwingend erforderliche oder wirtschaftlich sinnvolle Anpassungen bzw. Erweiterungen ergeben können, die mit einer Wiederaufnahme von Entwicklungsarbeiten einhergehen (z. B. Retrainieren einer KI-Komponente aufgrund eines kritischen Leistungsabfalls).

Inbetriebnahme heißt in der Regel auch, dass ein KI-System in bestehende Arbeitsläufe bzw. Prozessketten eingebunden wird und, je nach Autonomiegrad, täglich mit Menschen interagieren und gegebenenfalls kommunizieren muss. Oft verhält es sich in diesem Zusammenhang unerwartet und es stellen sich Programmfehler, auch Bugs genannt, heraus. Zu den typischen Bugs zählen u. a. Fehler im Bedienkonzept (z. B. nicht ausreichende Problembehebungsoptionen für die Anwenderinnen und Anwender) und Fehler infolge der Betriebsumgebung (z. B. schwankende Lichtverhältnisse). Um diesen Herausforderungen zu begegnen, empfiehlt der Leitfaden im Hinblick auf den Betrieb von KI-Systemen vier Kriterien zu berücksichtigen: Bedienbarkeit, Leistungsmonitoring, Instandhaltung bzw. Aktualisierung und Dokumentation im Betrieb.

3.4.1 Bedienbarkeit



Das Kriterium Bedienbarkeit betrachtet, wie das KI-System und seine Anwenderinnen und Anwender im Betrieb miteinander interagieren und kommunizieren (Mensch-Maschine-Interaktion). Dies schließt folgende Aspekte bzw. Indikatoren ein:

i) den Nutzendenkreis des KI-Systems und gegebenenfalls Einschränkungen der nutzenden Personen (Indikator I3.1.2)

ii) Interaktionen, die zwingend erforderlich sind, z. B. um die Funktionalität des Systems sicherzustellen (Indikator I3.1.2)

iii) optionale Interaktionen, z. B. zur manuellen Korrektur von Eingabedaten (Indikator I3.1.3)

iv) Prävention und Behebung von Interaktionsproblemen (Indikator I3.1.4)

v) Absicherung bei Bedienungsfehlern oder unsachgemäßer Handhabung des KI-Systems (Indikator I3.1.5)

vi) Fehlfunktionen oder Ausfälle des KI-Systems (Indikator I3.1.6)

vii) Notfallsituationen bzw. Abschalt-Szenarien (Indikator I3.1.7)

Die Bedienbarkeit wirkt sich auf die Gebrauchstauglichkeit[21] eines KI-Systems aus und darauf, ob es vom potenziellen Nutzenden- bzw. Kundenkreis akzeptiert wird (siehe Beispielszenario A). Sie steht daher im Zusammenhang mit dem wirtschaftlichen Erfolg, der mit einem System erzielt werden kann. Die Gebrauchstauglichkeit selbst kann zudem regulatorisch relevant sein (z. B. stellt die EU-Verordnung für Medizinprodukte (MDR) explizite Anforderungen an die Gebrauchstauglichkeit und mit der DIN EN 62366-1:2021-08 existiert eine Norm für die Gebrauchstauglichkeit von Medizinprodukten, die im Rahmen der Konformitätsbewertung durch Dritte hinzugezogen werden kann).

Die Bedienbarkeit ist des Weiteren unter Umständen auch im Hinblick auf haftungsrechtliche Risiken relevant (siehe Beispielszenario B). Wenn in einem Krankenhaus ein Entscheidungsunterstützungssystem eingesetzt wird und es im Betrieb zu Fehldiagnosen kommt, die mit weitreichenden Konsequenzen für die Betroffenen verbunden sind, dann stellt sich die Frage, wer dafür verantwortlich ist. Die Pflegekraft, die das System unsachgemäß bedient hat; das ärztliche Personal, das seiner Sorgfaltspflicht nicht nachgekommen ist; die Organisation, die das System entwickelt hat (Maliha et al. 2021; Marotta 2022)? Hätten die Fehldiagnosen vielleicht verhindert werden können, z. B. durch einen Mechanismus, der die Pflegekraft auf einen möglichen Bedienungsfehler ihrerseits hinweist, oder durch und eine vom Krankenhaus festgelegte Standardvorgehensweise (engl. standard operating procedure), die sie gezwungen hätte, eine Ärztin oder einen Arzt hinzuzuziehen?

Im Kontext der Bedienbarkeit von KI-Systemen sollte daher betrachtet werden, welche Bedienelemente notwendig sind, um die Nutzerinnen und Nutzer des Systems und die von den Ausgaben des Systems gegebenenfalls betroffenen Personen bestmöglich zu schützen bzw. vor Fehlern zu bewahren. In Abhängigkeit von dem Risiko, dass von dem System und seinen Ausgaben ausgeht, müssen dafür unter Umständen entsprechende Schutz- und Designkonzepte entworfen und umgesetzt werden (siehe Beispielszenario B). Der KI-Prüfkatalog des Fraunhofer IAIS stellt dementsprechend unter der Dimension „Autonomie und Kontrolle“ eine Schutzbedarfsanalyse für KI-Anwendungen vor sowie Maßnahmen, mit denen sich der identifizierte Bedarf umsetzen lässt bzw. die Umsetzung des Bedarfs überprüft werden kann (Poretschkin et al. 2021).

BEISPIELSZENARIO A

2018 gelangten B. J. May und seine intelligente Türklingel zu medialem Ruhm[22]. Letztere war mit einem kamerabasierten System ausgestattet, das automatisch das gesamte Haus abriegelt, wenn sich dem System unbekannte Personen der Tür nähern. Am 18. September erkannte das System seinen ihm eigentlich bekannten Besitzer nicht und riegelte das Haus ab. B. J. May fand schnell heraus, was passiert war: Das System hatte ein Bild von dem potenziellen Einbrecher aufgenommen– Batman. B. J. May trug an diesem Tag ein T-Shirt mit Batman-Konterfei, wodurch es zu einer „Verwechslung“ kam.

Das Haus konnte B. J. May schnell wieder entriegeln, da passende Maßnahmen zur Bedienbarkeit bzw. Behandlungen von Fehlfunktionen getroffen worden waren und das System mit entsprechenden Mechanismen ausgestattet war: Aufheben der Verriegelung über ein PIN-Eingabegerät neben der Klingel und, falls dies nicht möglich ist, über eine App auf dem Mobiltelefon. Dank dieser Vorkehrungen hatte der Batman-Vorfall keine ernsthafteren Konsequenzen.

21 Die Normenreihe ISO 9241 definiert Gebrauchstauglichkeit folgendermaßen: „Gebrauchstauglichkeit oder Usability bezeichnet die Eignung eines Produktes bei der Nutzung durch bestimmte Benutzerinnen und Benutzer in einem bestimmten Benutzungskontext die vorgegebenen Ziele effektiv, effizient und zufriedenstellend zu erreichen.“

22 https://www.independent.ie/world-news/and-finally/this-man-was-locked-out-of-home-when-his-smart-doorbell-thought-he-was-batman-37329890.html

BEISPIELSZENARIO B

Der Linearbeschleuniger Therac-25 wurde bis 1987 für die Strahlentherapie eingesetzt und führte in den Jahren 1985 bis 1987 den Tod von drei Krebspatient/-innen herbei. Der Computer des Therac-25 war sowohl für die Messwerterfassung und Gerätesteuerung als auch für die Interaktion mit den Nutzerinnen und Nutzern zuständig. Das heißt: Er führte beide Aufgaben parallel aus. Das Kernproblem dabei war eine zeitweise verzögerte Synchronisation der beiden Prozesse. Unter gewissen Umständen konnte es nämlich passieren, dass auch nach einer Korrektur der Eingabedaten durch das bedienende Personal der Computer bei der Ansteuerung des Geräts veraltete Daten von vor der manuellen Korrektur verwendete. Dieser und weitere Fehler führten zu Überdosierungen der Strahlen im tödlichen Bereich. Dieses Fehlverhalten hätte voraussichtlich mit einem geeigneten Schutzkonzept, das Maßnahmen bei fehlerhaften bzw. gefährlichen Eingaben oder Ausgaben vorsieht, vermieden werden können.

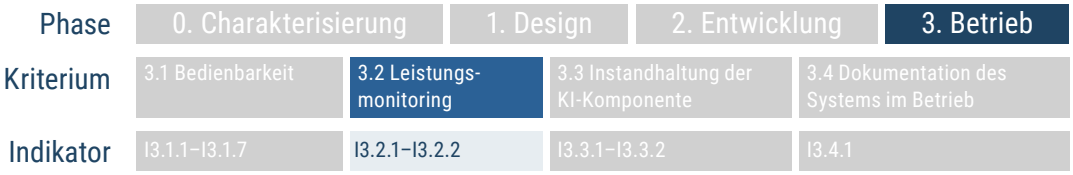
LÖSUNGSHILFEN

- **LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE**
  Die Norm ISO/IEC/IEEE 26514:2022 umfasst die Gestaltung von Informationen und Instruktionen für Nutzerinnen und Nutzer von Softwareprodukten. Sie kann gegebenenfalls hinzugezogen werden, um die Bedienbarkeit eines KI-Systems zu dokumentieren (z. B. Beschreibung von Hilfefunktionen, Anleitung zur Problembehebung, Erläuterung von Abschalt-Szenarien) (Indikatoren I3.1.1–I3.1.7).

- **LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO und LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO**
  Die Norm ISO 9241-110:2020 beschreibt allgemeine Prinzipien der Interaktion zwischen Nutzerinnen und Nutzern und Systemen und stellt vor, wie diese Prinzipien im Rahmen der Gestaltung von interaktiven Systemen angewendet werden können. Die ISO 9241-13:1998 befasst sich ergänzend mit der Benutzendenführung, u. a. im Kontext des Fehlermanagements (z. B. Eingabeaufforderungen, Feedback, Statusinformationen). Beide Normen können gegebenenfalls bei der Gestaltung der Bedienbarkeit von KI-Systemen unterstützen (Indikatoren I3.1.1–I3.1.7).

- **LH-MW_72_SOP_BIT.AI**
  Der BIT.AI-Blog-Artikel gibt einen Überblick zu Standard Operating Procedures und wie diese grundsätzlich erstellt werden können (u. a. Formate, Strukturierung, Inhalte) (Indikatoren I3.1.1–I3.1.7).

3.4.2 Leistungsmonitoring



Das Monitoring der Leistung eines KI-Systems im Betrieb umfasst die Erfassung von Leistungskennzahlen (z. B. Genauigkeit; Indikator I3.2.1) sowie die Aufzeichnung von Ereignissen, die zusätzlichen Aufschluss über die Leistungsfähigkeit des Systems in der Praxis geben (Indikator I3.2.2). Zu diesen Ereignissen können z. B. Fehler, die zu einem Abbruch geführt haben, Laufzeitfehler oder fehlerhafte Eingaben der Nutzerinnen und Nutzer zählen.

Das Leistungsmonitoring ist unter mehreren Gesichtspunkten von Bedeutung:

i) In Abhängigkeit vom Einsatzkontext des Systems ist das Monitoring gegebenenfalls regulatorisch vorgeschrieben (z. B. macht die EU-Verordnung für Medizinprodukte (MDR) strenge Vorgaben hinsichtlich einer kontinuierlichen und systematischen Überwachung nach dem Inverkehrbringen eines Produkts (engl. post-market surveillance[23])).

ii) Das Monitoring kann proaktiv wirtschaftlichen Verlusten und gegebenenfalls auch haftungsrechtlichen Konsequenzen vorbeugen (z. B. wenn den von den Ausgaben des Systems betroffenen Personen oder Organisationen aufgrund eines nicht erkannten Leistungsabfalls des Systems oder seiner KI-Komponente ein finanzieller/physischer/psychischer Schaden entsteht; siehe auch Orientierungshilfe zu Leistungsabfällen von KI-Komponenten und Beispielszenario).

iii) Das Monitoring kann dazu beitragen, daten- bzw. KI-bezogene Unsicherheiten zu identifizieren, die sich auf das Gesamtverhalten bzw. die Gesamtleistung von kollaborativ agierenden Systemverbünden auswirken (z. B. automatisierte kooperative Fahrzeuge).

iv) Aus dem Monitoring können gegebenenfalls Modifikationen abgeleitet werden, die im Kontext einer zukünftigen Produktverbesserung oder der Umsetzung eines ähnlichen Produkts relevant sind.

BEISPIELSZENARIO:
Wissenschaftlerinnen und Wissenschaftler der University of Southampton und des University Hospital Southampton in Großbritannien haben eine KI-Komponente entwickelt, die den Behandlungs- bzw. Pflegeaufwand für Personen vorhersagt, die eine Notaufnahme aufsuchen (Duckworth et al. 2021). Die Anwendung gibt aus, ob eine Person für eine Behandlung in das Krankenhaus aufgenommen oder ob sie nach dem Besuch der Notaufnahme direkt wieder entlassen wird. Die Wissenschaftlerinnen und Wissenschaftler konnten zeigen, dass sich über ein wöchentliches Leistungsmonitoring der KI-Komponente ein mit dem Ausbruch der Covid-19-Pandemie einsetzender Leistungsverfall identifizieren lässt. Der Leistungsverfall drückt sich dabei durch eine höhere „falsche Alarmrate“ aus, d. h. die Anwendung sagt voraus, dass Patientinnen oder Patienten aufgenommen werden müssen, obwohl sie in Wirklichkeit entlassen werden. Wäre die KI-Anwendung tatsächlich in einem Krankenhaus zum Einsatz gekommen, hätte dies vor-

23 Für eine Übersicht zu den Vorgaben siehe https://www.johner-institut.de/blog/regulatory-affairs/post-market-surveillance/

aussichtlich Konsequenzen gehabt, u. a. eine psychische Belastung der fälschlicherweise aufgenommenen Personen. Die Belegung von Krankenhausbetten mit Personen, die sie eigentlich nicht brauchen, hätte zudem dazu führen können, dass für Patientinnen und Patienten im kritischen Zustand nur unzureichend Betten zur Verfügung stehen. Die Anwendung hätte somit theoretisch auch einen physischen Schaden verursachen können.

Orientierungshilfe zu Leistungsabfällen von KI-Komponenten (Indikator I3.2.1)
Die KI-Komponenten von KI-Systemen bergen das Risiko eines Modellverfalls bzw. Model Drift. Darunter versteht man eine Abnahme der Vorhersagegenauigkeit eines Modells über die Zeit, d. h. die Ausgaben des Modells sind irgendwann nicht mehr so gut wie direkt nach dem Trainieren des Modells. Zu den Ursachen für einen Modellverfall zählen u. a. Datendrift (engl. data drift) und Konzeptdrift (engl. concept drift).

Ein Datendrift liegt in Situationen vor, in denen sich die statistische Verteilung der Eingabedaten ändert. Solche Veränderungen können sich sowohl schleichend einstellen, z. B. durch eine graduelle Verschlechterung der Messeigenschaften eines Sensors aufgrund von Verschmutzung, als auch ad hoc durch Ereignisse ausgelöst werden, wie z. B. durch den Austausch eines Messgeräts. Wenn etwa Hardwareelemente eines Systems ausgetauscht werden müssen, z. B. eine defekte Kamera oder ein defekter Sensor, kann dies leicht einen sprunghaften Datendrift zur Folge haben. Durch die neue Kamera verändert sich gegebenenfalls die Auflösung der aufgenommenen Bilder bzw. der neue Sensor weist eine andere Messunsicherheit bei der Temperaturmessung auf.

Von Konzeptdrift spricht man dagegen, wenn sich das Realsystem bzw. die den Analysegegenstand bestimmenden Wirkzusammenhänge verändern und folglich das Modell die Realität nicht mehr oder nur noch teilweise widerspiegelt. Das heißt, das Modell bildet Zusammenhänge bzw. Muster ab, die in Bezug auf die Vergangenheit bzw. die Trainingsphase Bestand hatten, jedoch in der Gegenwart nicht mehr zutreffen. Dies tritt z. B. ein, wenn sich das Kaufverhalten im Zuge einer Pandemie auf einmal drastisch ändert oder wenn Betrugsmethoden, die durch ein KI-System abgewehrt werden sollen, kontinuierlich weiterentwickelt werden.

Eine weitere Ursache für einen Modellverfall kann auch die Veränderung der unabhängigen Variablen darstellen. Dies kann beispielsweise der Fall sein, wenn ein Modell im Gesundheitswesen hauptsächlich basierend auf Daten von jungen Erwachsenen trainiert wurde, in der Praxis aber überwiegend mit Daten von Seniorinnen und Senioren konfrontiert wird.

Wie hoch das Risiko für einen Modellverfall ist, ist häufig von dem konkreten Einsatzkontext abhängig, jedoch können auch weitere externe Entwicklungen eine Rolle spielen. In Zeiten von Pandemien oder globalen Krisen steigt dieses Risiko für viele Systeme zusätzlich. Ein Modellverfall sollte in jedem Fall möglichst früh identifiziert werden, d. h. bevor die Ausgaben des Modells so unzuverlässig sind, dass daraus ein wirklicher Schaden entsteht. Das Leistungsmonitoring kann in Verbindung mit Methoden zur Erkennung eines möglichen Modellverfalls hierzu beitragen. Machine-Learning-Plattformen und kommerzielle Anbieter stellen bereits entsprechende Module zur Verfügung (z. B. TensorFlow Data Validation, IBM Watson Studio). Aus Forschungsansätzen sind Algorithmen für die Erkennung eines Modellverfalls in einem spezifischen Anwendungskontext entnehmbar (z. B. für die vorausschauende Wartung (Zenisek et al. 2019)).

LÖSUNGSHILFEN

An dieser Stelle kann zum jetzigen Zeitpunkt noch nicht auf universell einsetzbare Best Practices/Erfolgsmodelle für das Leistungsmonitoring von KI-Systemen verwiesen werden, die sich in der Praxis bewährt haben. Die im Folgenden angeführten Lösungshilfen sollen dabei unterstützen, verschiedene Aspekte des Leistungsmonitorings zu eruieren.

- **LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut**
  Der Leitfaden zur KI bei Medizinprodukten des Johner Instituts führt unter Abschnitt D.2 Maßnahmen für die Marktüberwachung nach dem Inverkehrbringen an, die im Kontext einer Konformitätsbewertung durch benannte Stellen zu erwarten sind (Johner-Institut 2021a) (Indikator I3.2.1, Indikator I3.2.2).

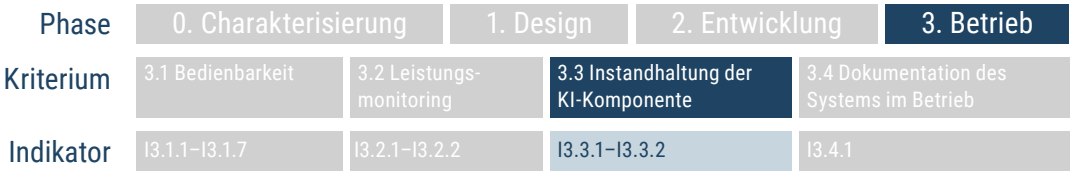
- **LH-BP_74_EvaluationClinicalDecisionSupport_DukeMargolisCenter**
  Für das Monitoring von klinischen Entscheidungsunterstützungssystemen hat das Duke Margolis Center für Health Policy Anfang 2022 einen Report veröffentlicht, der Empfehlungen ausspricht, u. a. dazu, welche Datenelemente benötigt werden (Duke Margolis Center for Health Policy 2022) (Indikator I3.2.1, Indikator I3.2.2).

- **LH-FA_77_UncertaintyHandling_Bandyszak**
  Das Buchkapitel „Handling Uncertainty in Collaborative Embedded Systems Engineering" von Bandyszak et al. 2021 stellt u. a. Methoden zur Modellierung/Identifizierung von daten- bzw. KI-bezogenen Unsicherheiten im Kontext des Designs und Betriebs von kollaborativ agierenden Systemverbünden vor (Indikator I3.2.1, Indikator I3.2.2).

- **LH-FA_76_ConceptDrift_Lu**
  Das Review von Lu et al. 2018 gibt einen umfassenden Überblick über den internationalen Stand der Forschung zum Thema Modellverfall bzw. Leistungsverfall im Betrieb aufgrund von Konzeptdrift (siehe Orientierungshilfe für Begriffsdefinition). Aufbauend darauf wird ein Rahmenwerk für das Lernen unter Konzeptdrift vorgeschlagen, das sich aus drei Komponenten zusammensetzt: Konzeptdrift erkennen, Konzeptdrift verstehen und Konzeptdrift adaptieren (Indikator I3.2.1, Indikator I3.2.2).

- **LH-Std_75_AIModificationProposedRegulation_FDA**
  Das Diskussionspapier der US-amerikanischen Behörde für die Zulassung und Marktüberwachung von Lebensmitteln, Medikamenten und Medizinprodukten (FDA) schlägt einen Regulierungsrahmen für die Modifikation von KI-basierter Software als Medizinprodukt vor (FDA 2019) (Indikator I3.2.1, Indikator I3.2.2).

3.4.3 Instandhaltung der KI-Komponente



Das Kriterium Instandhaltung berücksichtigt Maßnahmen, die im Falle eines Leistungsabfalls der KI-Komponente eines Systems im Betrieb vorgesehen sind. Es umfasst Indikatoren für den Umgang mit einem Modellverfall der KI-Komponente (z. B. manuell anzustoßendes oder automatisch eingeleitetes Retraining; Indikator I3.3.1) sowie die Bewertung und Dokumentation von entsprechenden Aktualisierungs- bzw. Instandhaltungsmaßnahmen (Indikator I3.3.2).

Wie mit einem Leistungsabfall eines KI-Systems bzw. einer KI-Komponente umzugehen ist, hängt vom Einsatzkontext des Systems ab. Medizinprodukte, deren Leistung im Betrieb zwingend überwacht werden muss (EU-Verordnung für Medizinprodukte (MDR)), müssen z. B. sofort vom Markt genommen werden, sobald sie ihrer initial definierten Zweckbestimmung nicht mehr nachkommen können oder sie eine Bedrohung für die Patientinnen und Patienten darstellen (Johner Institut 2021a, 2022). Falls also im Betrieb ein Modellverfall identifiziert wird, muss die KI-Komponente neu- bzw. retrainiert werden. Im Anschluss daran muss dann das gesamte System erneut den Kennzeichnungsprozess als Medizinprodukt durchlaufen. Dieser Umstand erschwert bislang die Kennzeichnung von kontinuierlich lernenden Systemen (engl. continuous learning systems) als Medizinprodukt, bei denen jede neu gelernte Instanz der KI-Komponente neu gekennzeichnet werden muss. Dies kommt eigentlich einer kompletten Neuentwicklung des KI-Systems bzw. seiner KI-Komponente gleich und nicht einer Instandhaltung bzw. Aktualisierung.

Auch in anderen Einsatzkontexten bzw. Branchen ist es mittlerweile üblich, einem identifizierten Modellverfall mit entsprechenden Maßnahmen wie Retrainieren der KI-Komponente zu begegnen. Hierfür kann unter Umständen auf neue Methoden aus dem Bereich Modelloptimierung bzw. -augmentierung zurückgegriffen werden, z. B. Einspeisung zusätzlicher Trainingsdaten in das initial erlernte Modell (engl. data aggregation integration) (Albarqouni et al. 2016). Das Retrainieren einer KI-Komponente muss gegebenenfalls manuell von den Nutzenden angestoßen werden und ist in der Regel mit zeitlichem, finanziellem und personellem Aufwand verbunden. MLOps-Plattformen bieten auch die Möglichkeit, Modelle automatisch zu retrainieren. Dadurch reduziert sich der Aufwand, unter Umständen leidet aber die Effektivität darunter, da die Ursachen für den Modellverfall nicht näher analysiert werden. Ein Ansatz, um dem entgegenzukommen, stellt eine Verbindung von automatischem Retrainieren und automatischer Datenexploration dar[24].

Das Entwickeln von optimalen Vorgehensmodellen für das Retrainieren von KI-Komponenten ist aktuell noch Gegenstand der Forschung. Dies schließt Fragen dazu ein, ab wann ein Leistungsverfall als kritisch einzustufen ist bzw. ein Retrainieren erforderlich macht und wie ein kritischer Leistungsverfall sich am besten methodisch erkennen lässt (siehe Beispielszenario).

24 https://insights.sei.cmu.edu/blog/improving-automated-retraining-of-machine-learning-models/

BEISPIELSZENARIO

Wissenschaftlerinnen und Wissenschaftler der University of Southampton und des University Hospital Southampton in Großbritannien haben eine KI-Komponente entwickelt, die vorhersagt, ob eine Person, die eine Notaufnahme aufsucht, aufgenommen oder nach dem Besuch direkt wieder entlassen wird (Duckworth et al. 2021). Mit dem Ausbruch der Covid-19-Pandemie ist ein Leistungsverfall der KI-Komponente erkennbar (siehe Beispielszenario aus Abschnitt 3.4.2). Der Leistungsabfall lässt sich dabei darauf zurückführen, dass sich die statistische Verteilung der Eingabedaten mit Einsetzen der Pandemie geändert hat (Datendrift). Die Wahrscheinlichkeit, dass eine die Notaufnahme aufsuchende Person anschließend zur weiteren Behandlung aufgenommen werden muss, ist gestiegen. Um gleichzeitig erkennen zu können, wann ein Retrainieren der KI-Komponente erforderlich ist und warum ihre Leistung abfällt, haben die Wissenschaftlerinnen und Wissenschaftler ein Vorgehensmodell entwickelt, das neben den Metriken für die Leistung (AUC, Präzision, Recall, Konfusionsmatrix) auch die Wichtigkeit der Features für die Vorhersage berücksichtigt. Für letzteres verwenden sie ein Werkzeug aus dem Bereich Erklärbare KI (Shapley Additive Explanation, SHAP).

LÖSUNGSHILFEN

An dieser Stelle kann zum jetzigen Zeitpunkt noch nicht auf universell einsetzbare Best Practices/Erfolgsmodelle für die Instandhaltung von KI-Komponenten verwiesen werden, die sich in der Praxis bewährt haben. Die im Folgenden angeführten Lösungshilfen sollen dabei unterstützen, verschiedene Aspekte zu eruieren, die hinsichtlich der Instandhaltung relevant sein können.

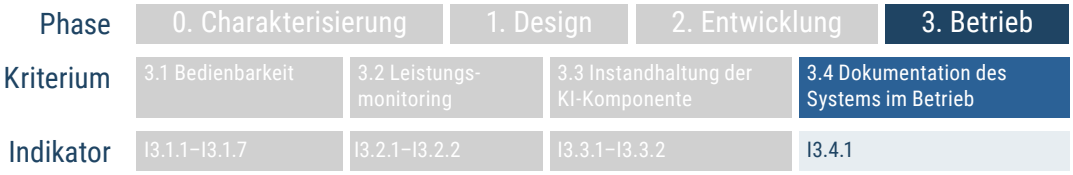
- **LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut**
  Der Leitfaden zur KI bei Medizinprodukten des Johner Instituts führt unter Abschnitt D.2 Maßnahmen für die Marktüberwachung nach dem Inverkehrbringen an, die im Kontext einer Konformitätsbewertung durch die im Gesundheitswesen im Zertifizierungskontext verantwortlichen benannten Stellen zu erwarten sind (Johner Institut 2021a). Dies schließt Maßnahmen zur Instandhaltung/Aktualisierung von Produkten ein (Indikator I3.3.1, Indikator I3.3.2).

- **LH-FA_76_ConceptDrift_Lu**
  Das Review von Lu et al. 2018 gibt einen umfassenden Überblick über den internationalen Stand der Forschung zum Thema Modellverfall bzw. Leistungsverfall im Betrieb aufgrund von Konzeptdrift (siehe Orientierungshilfe für Begriffsdefinition). Aufbauend darauf wird ein Rahmenwerk für das Lernen unter Konzeptdrift vorgeschlagen, das sich aus drei Komponenten zusammensetzt: Konzeptdrift erkennen Konzeptdrift verstehen und Konzeptdrift adaptieren (Indikator I3.3.1, Indikator I3.3.2).

- **LH-Std_75_AIModificationProposedRegulation_FDA**
  Das Diskussionspapier der US-amerikanischen Behörde für die Zulassung und Marktüberwachung von Lebensmitteln, Medikamenten und Medizinprodukten (FDA) schlägt einen Regulierungsrahmen für die Modifikation von KI-basierter Software als Medizinprodukt vor (FDA 2019) (Indikator I3.3.1, Indikator I3.3.2).

- **LH-FA_78_AggregationSchemes_Albarqouni**
  Die Fachpublikation von Albarqouni et al. 2016 stellt mit AggNet einen Ansatz für das Trainieren von neuronalen Netzen auf mehreren unterschiedlich annotierten Datensätzen vor (engl. data aggregation integration).

3.4.4 Dokumentation des KI-Systems im Betrieb



Unabhängig von der Branche und spezifischen gesetzlichen Verpflichtungen sollten Organisationen, die ein KI-System einsetzen, dokumentieren, wie sie es betreiben und wie sie einen ordnungsgemäßen Betrieb sicherstellen (Indikator I3.4.1). Eine solche Betriebsdokumentation kann z. B. umfassen, in welchen Prozessen das System eingesetzt wird, welche Gebrauchs- oder Arbeitsanweisungen existieren und wo diese abgelegt sind, welche Kontrollelemente vorhanden sind (u. a. Qualitätskontrolle und Archivierungskonzepte), welche Instandhaltungsmaßnahmen getroffen wurden oder wie mit Störungen umgegangen wurde.

Der Mehrwert einer Betriebsdokumentation ist nicht nur auf den Nachweis des ordnungsgemäßen Betriebs des KI-Systems beschränkt, der gegebenenfalls für haftungsrechtliche Fragestellungen relevant ist. Die Betriebsdokumentation kann auch dazu beitragen, die Abläufe in einer Organisation zu verbessern, beispielsweise dadurch, dass bei Störungen gleich gezielt gehandelt werden kann, da die Lösung bereits dokumentiert und für die Mitarbeiterinnen und Mitarbeiter zugänglich ist. Wenn die Dokumentation Arbeitsanweisungen und Standardvorgehensweisen (engl. standard operating procedures) beinhaltet, vereinfacht sich die Einarbeitung von neuen Personen und es ist sichergestellt, dass alle Personen das System auf die gleiche Art und Weise bedienen.

Aufgrund von regulatorischen Rahmenbedingungen kann eine Dokumentation des KI-Systems im Betrieb zudem für die Organisation, die das System auf den Markt gebracht hat, verpflichtend vorgeschrieben sein (siehe auch Orientierungshilfe zu regulatorischen Rahmenbedingungen).

Orientierungshilfe zu regulatorischen Rahmenbedingungen (Indikator I3.4.1)
KI-Systeme können je nach Anwendungsszenario bereichsspezifischen Dokumentationspflichten auf Seiten der herstellenden Organisation unterliegen. Die Dokumentationspflichten dienen zum einen als Nachweis, dass das jeweilige System den regulatorischen Anforderungen genügt. Zum anderen tragen sie dazu bei, die Funktionsfähigkeit des Produkts über den gesamten Lebenszyklus zu überwachen, einschließlich der Betriebsphase. Der Nachweis der Funktionsfähigkeit ist neben der Erfüllung der gesetzlichen Vorgaben auch geeignet, die herstellende Organisation im Falle eines Schadensereignisses hinsichtlich möglicher Schadensersatzansprüche zu entlasten. Denn im Rahmen der Produzentenhaftung muss die herstellende Organisation für Schäden eintreten, die infolge eines fehlerhaften Produkts hervorgerufen werden, zumindest dann, wenn der Schaden auf einem Verschulden der Organisation beruht. Dies gilt nicht nur für das Inverkehrbringen eines Produkts, sondern auch für die Überwachung des Produkts nach der Markteinführung. In diesem Zusammenhang kann es aus Sicht der herstellenden Organisation geboten sein, qualitätskritische Daten über die Funktionsfähigkeit eines Produkts fortwährend zu sammeln und auszuwerten. Dies gilt insbesondere dann, wenn durch das Produkt sensible Rechtsgüter wie die Gesundheit oder das Leben von natürlichen Personen beeinträchtigt werden können. Bei Bekanntwerden entsprechender Mängel muss die herstellende Organisation das Produkt im Rahmen ihrer Produktbeobachtungspflicht zurückrufen und gegebenenfalls Nachbesserungen durchführen. Daten über sicherheitserhebliche Vorfälle sollten daher gespeichert werden, um den Schadensverlauf nachvollziehen und mögliche Folgemaßnahmen einleiten zu können (z. B. einen Produktrückruf).

Darüber hinaus wird eine Produktüberwachungspflicht in bestimmten Sektoren auch gesetzlich vorgeschrieben. Als Beispiel kann in diesem Zusammenhang die Medizinprodukteverordnung (EU) 2017/745 herangezogen werden (Das Europäische Parlament und der Rat der Europäischen Union 2017). Diese fordert in Art. 83 Abs. 2 ein System zur Überwachung von Medizinprodukten für den Zeitraum nach dem Inverkehrbringen des Medizinprodukts. Danach soll das System aktiv und systematisch einschlägige Daten über die Qualität, die Leistung und die Sicherheit eines Produkts während dessen gesamter Lebensdauer sammeln, aufzeichnen und analysieren sowie die erforderlichen Schlussfolgerungen ziehen und dazu geeignet sein, etwaige Präventiv- oder Korrekturmaßnahmen zu ermitteln, durchzuführen und zu überwachen.

Auch für Betreiberinnen und Betreiber von kritischen Infrastrukturen gelten spezifische Anforderungen hinsichtlich der Sicherheit der Informationstechnik. Diese sind nach § 8a BSI-Gesetz verpflichtet, angemessene organisatorische und technische Vorkehrungen zur Vermeidung von Störungen der Verfügbarkeit, Integrität, Authentizität und Vertraulichkeit ihrer informationstechnischen Systeme, Komponenten oder Prozesse zu treffen, die für die Funktionsfähigkeit der von ihnen betriebenen kritischen Infrastrukturen maßgeblich sind. Dies umfasst die Einrichtung eines Protokollierungssystems, welches es ermöglicht, Unregelmäßigkeiten zu erkennen und Sicherheitsvorfälle im Nachhinein analysieren zu können. Die Protokollierung dient zudem auch dem Zweck der Beweissicherung bei Störfällen.

Eine gesetzlich normierte Aufzeichnungspflicht für sogenannte Hochrisiko-KI-Systeme (siehe Abschnitt 3.1.1 für eine Erläuterung) zeichnet sich auch auf europäischer Ebene ab. Nach Art. 12 der KI-Verordnung der EU (Europäische Kommission 21.04.2021) müssen Hochrisiko-KI-Systeme so konzipiert und entwickelt werden, dass eine automatische Aufzeichnung von Vorgängen und Ereignissen während des Betriebs möglich ist. Die Protokollierung soll das Funktionieren des KI-Systems während seines gesamten Lebenszyklus in einem der Zweckbestimmung des Systems angemessenen Maße rückverfolgbar machen.

Hinsichtlich des Umfangs der Dokumentationspflicht ergeben sich aus den gesetzlichen Regelungen zumeist keine konkreten Vorgaben. Bedenkt man die nicht unerheblichen Datenmengen, die beim Betrieb von KI-Systemen anfallen, erscheint eine Limitierung der zu speichernden Daten bereits aus technischer Sicht angebracht. Vor dem Hintergrund der jeweiligen Dokumentationspflichten sollten nur sicherheitserhebliche Daten dauerhaft vorgehalten werden. Dies umfasst sämtliche Informationen, die notwendig sind, um die Funktionsfähigkeit des KI-Systems angemessen nachvollziehbar zu machen.

4 AUSBLICK

4 AUSBLICK

Der vorgestellte Leitfaden soll beauftragende und entwickelnde Organisationen bzw. Multi-Stakeholder-KI-Projekte dabei unterstützen, ihre gemeinsame Produktvision durch ein in Phasen gestaffeltes Qualitätsmanagement möglichst effizient in die Realität zu überführen – von der Charakterisierung über das Design und die Entwicklung bis hin zum Betrieb eines KI-Systems in der Praxis. Mit seinen Kriterien und Indikatoren erleichtert der Leitfaden die Festlegung von Qualitätszielen (das „Was"). Der Lösungshilfenkatalog unterstützt darüber hinaus dabei, diese Ziele durch entsprechende Maßnahmen zu verfolgen (das „Wie"). Der Lösungshilfenkatalog entspricht einer Momentaufnahme der Gegenwart, indem er abbildet, welche Hilfen in welcher Reife bereits zur Verfügung stehen (u. a. Normen, Standards, Best Practices, Werkzeuge und Forschungsansätze). Gleichzeitig macht er auch zukünftig wichtige Handlungsfelder sichtbar. Aus den Bereichen, die sich gegenwärtig noch durch wenige Lösungshilfen bzw. Lösungshilfen mit geringem Reifegrad auszeichnen, lassen sich aktuelle Bedarfe im Hinblick auf das Qualitätsmanagement von KI-Systemen ableiten. Diese Handlungsfelder werden im Folgenden kurz skizziert.

Normen und Standards für KI-Systeme

Wesentliche Normen, die für das Qualitätsmanagement von KI-Systemen wichtig sind, befinden sich derzeit noch in der Ausarbeitung. Dies schließt das Risikomanagement von KI (ISO/IEC FDIS 23894), die Qualitätsmodelle für KI-Systeme (ISO/IEC DIS 25059), die funktionale Sicherheit von KI-Systemen (ISO/IEC CD TR 5469), die Qualitätsbewertung von KI-Systemen (ISO/IEC AWI TS 5471) und das Testen von KI-Systemen (ISO/IEC AWI TS 29119-11) ein. Diese Normen sind auch im Hinblick auf die angestrebte Verordnung (AI Act) der EU (Europäische Kommission 21.04.2021) und die Konformitätsbewertung bzw. Zertifizierung von KI-Systemen durch Dritte hochrelevant. Es ist zum jetzigen Zeitpunkt davon auszugehen, dass die Normen anwendungsagnostisch ausgelegt sein werden. Sie müssen im Anschluss also noch in Normen bzw. Standards der jeweiligen Branchen überführt und mit den dort bereits bestehenden Normen und Standards harmonisiert werden. Der Umstand, dass entsprechende Standards in vielen Branchen gegebenenfalls erst nach mehreren Jahren verfügbar sein werden, stellt eine Herausforderung sowohl für die Entwicklung von KI-Systemen als auch für ihren praktischen Einsatz dar. Dies wurde auch mehrfach von den für den Leitfaden interviewten Expertinnen und Experten hervorgehoben. Da sich die KI-Technologien schnell und stetig weiterentwickeln, besteht zudem die Gefahr, dass die Standards hinter dem Stand der Technik zurückbleiben. Dies birgt das Risiko, dass sich Infrastrukturen ausbilden und Anwendungen im Markt etablieren, die mit dem übergeordneten Ziel der KI-Verordnung der EU (Europäische Kommission 21.04.2021), eine menschenzentrierte, zuverlässige Entwicklung von KI sicherzustellen, nicht vereinbar sind.

Testen von KI-Systemen im Betrieb bzw. unter Realbedingungen

Bislang existieren nur wenige Studien, die die Leistung von KI-Systemen im Betrieb und die Wahrnehmung ihrer Qualität durch die Nutzerinnen und Nutzer unter Realbedingungen systematisch untersuchen. Dies reflektiert auch der Leitfaden, der für die Betriebsphase im Vergleich zu den anderen Phasen weniger konkrete Lösungshilfen in Form von Best Practices oder etablierten Werkzeugen aufführt. In der kontrollierten Umgebung eines Entwicklungslabors kann die reale Welt in der Regel nicht in ihrer Gänze bzw. mit all ihren Nuancen simuliert werden. In Abhängigkeit vom spezifischen Einsatzkontext bzw. Einsatzort eines KI-Systems können sich etwa technische, soziale oder umweltbedingte Besonderheiten ergeben (z. B. spezielle Lichtverhältnisse, Prozesse oder Arbeitsabläufe), die seine Leistung unter Realbedingungen beeinträchtigen. Im Hinblick auf solche Faktoren und ihre Auswirkung bestehen noch Wissens- und Erfahrungslücken. Das Design von entsprechenden Teststrategien und Beobachtungsstudien im Rahmen kommender Forschungsaktivitäten kann dazu beitragen, diese Lücken perspektivisch

zu schließen. Aufbauend darauf könnten dann Faktoren, die sich maßgeblich auf die Leistung unter Realbedingungen auswirken, bereits gezielt in den Phasen Design und Entwicklung berücksichtigt werden. Dazu könnten sie beispielsweise frühzeitig in das Konzept für die Leistungsbewertung einfließen und ihre potenziellen Auswirkungen im Rahmen der Leistungsbewertung anschließend evaluiert werden. Dies kann perspektivisch dazu beitragen, den Übergang zwischen Entwicklungslabor und Markt bzw. Anwendungsbereich zu erleichtern.

Monitoring und Aufrechterhaltung der Leistung von KI-Systemen im Betrieb

Für alle KI-Systeme gilt, dass ihre Vorhersagekraft bzw. die Güte ihrer Ausgaben im Betrieb über die Zeit nachlassen kann – man spricht dann von einem Modellverfall (engl. model decay). Die KI-Komponenten der Systeme sind datengetrieben und Daten bzw. die Zusammenhänge zwischen Daten ändern sich, wenn sich die Umgebung verändert, in der sie erhoben werden. Solche Veränderungen sind zum Teil antizipierbar (z. B. in den Weihnachtsmonaten wird mehr eingekauft), zum Teil aber auch nicht (z. B. Eintreten globaler Krisen wie die Covid-19-Pandemie). Das heißt, ein Modellverfall kann für die meisten praktischen Anwendungen nicht ausgeschlossen werden. Er kann jedoch häufig durch das Neutrainieren des Modells, das der KI-Komponente zugrunde liegt, behoben werden. Im Hinblick auf das optimale Vorgehensmodell für den Umgang mit Modellverfall sind gegenwärtig aber noch viele Fragen offen. Dies reflektiert auch der Lösungshilfenkatalog des Leitfadens, der an dieser Stelle noch nicht auf etablierte Best Practices verweisen kann. Im Hinblick auf die offenen Fragen haben mehrere der für den Leitfaden interviewten Expertinnen und Experten hervorgehoben, dass es unklar ist, ab wann ein Modellverfall als kritisch eingestuft werden sollte bzw. ein Retrainieren zwingend erforderlich ist. Es zeichnet sich bereits jetzt ab, dass die Festlegung von entsprechenden Schwellwerten oder Schwellwertkriterien (z. B. Ausmaß der Abweichung von den vorher definierten Anforderungen) im Hinblick auf die Zertifizierung von KI-Systemen relevant sein wird. Der KI-Prüfkatalog des Fraunhofer IAIS sieht z. B. als Überprüfungskriterium für das Risikogebiet „Beherrschung der Dynamik“ bereits Entsprechendes vor (Poretschkin et al. 2021).

Testen der Robustheit von KI-Systemen gegenüber zufälligen oder gezielt herbeigeführten Störeinflüssen

KI-Systeme werden im Betrieb mit Widrigkeiten (engl. adversities) konfrontiert, die KI-Komponenten dahingehend beeinflussen können, dass sie unkorrekte Vorhersagen oder Ergebnisse ausgeben. Solche Widrigkeiten können zufällig auftreten (z. B. dunklere Bilder als Eingabedaten aufgrund veränderter Lichtverhältnisse in einer Fabrikhalle) oder absichtlich herbeigeführt werden (z. B. gezielte Manipulation oder Sabotage von außen). Wenn die KI-Komponenten nicht robust gegenüber solchen Widrigkeiten sind, kann sich das sowohl auf die Zuverlässigkeit als auch auf die funktionale Sicherheit der Systeme auswirken (z. B. Fehldiagnose aufgrund eines Staubkorns auf der Linse bei der Bildaufnahme bzw. Verkehrsunfall aufgrund eines Stoppschilds, das wegen eines Graffitis nicht erkannt wurde). Der Leitfaden sieht daher das Testen der Robustheit der KI-Komponente als Indikator für die Qualität der Leistungsbewertung während der Entwicklungsphase vor. Für das Testen kann zwar auf grundsätzliche Strategien (u. a. adversarial training) sowie erste Werkzeuge und Forschungsansätze zurückgegriffen werden, es mangelt aber noch an dezidierten anwendungsspezifischen Vorgehensmodellen für das Qualitätsmanagement. Letztere könnten entscheidend dazu beitragen, den gegenwärtig eher reaktiv geprägten Umgang mit Robustheit (Mängel werden erst nach der Inbetriebnahme festgestellt und nachträglich behoben) in einen proaktiven Ansatz zu überführen (das Auftreten von Mängeln nach der Inbetriebnahme wird bereits durch entsprechende Maßnahmen während der Entwicklungsphase bestmöglich vermieden).

Umweltwirkung von KI-Systemen

Perspektivisch zeichnet sich bereits jetzt eine Ergänzung des Leitfadens um ein weiteres Kriterium für die Designphase ab, das explizit die Umweltwirkung des KI-Systems berücksichtigt. Das Trainieren von KI-Anwendungen ist mit teilweise nicht unerheblichen Umweltbelastungen und einem hohen Energieverbrauch verbunden. Allein die CO2-Emissionen für ein einfaches Basissprachmodell entsprechen z. B. schnell denen eines Transatlantikflugs (Strubell et al. 2019; Bender et al. 2021). Die Machine Learning Community setzt sich verstärkt mit diesem Thema auseinander und es ist davon auszugehen, dass die Gestaltung rechnerisch effizienter Hardware und Algorithmen zunehmend an Bedeutung gewinnen wird (siehe hierzu auch Kapitel „Nachhaltige Gestaltung von KI“ in Mock et al. 2022). Dies wurde auch von den für den Leitfaden interviewten Expertinnen und Experten mehrfach hervorgehoben. Das Forschungsfeld ist noch sehr jung; bisher liegen daher noch keine wissenschaftlichen Ansätze oder Best Practices vor, aus denen sich konkrete anwendungsagnostische Indikatoren für die Bewertung der Umweltwirkung ableiten lassen.

A ANHANG

A ANHANG

A.1 Das WKIO-Modell

Hinter dem WKIO-Modell verbirgt sich eine Strukturierung eines zu bewertenden Gegenstands entlang von **W**erten, **K**riterien, **I**ndikatoren und **O**bservablen. Die Werte definieren dabei ein übergeordnetes Anliegen wie z. B. Transparenz oder Verantwortlichkeit im Fall der ethischen Konformität bzw. Zuverlässigkeit von KI-Systemen. Um bewertbar zu machen, ob die Werte erfüllt oder verletzt werden, werden zunächst Kriterien definiert (z. B. Offenlegung der Originaldaten für den Wert Transparenz). Da die Kriterien in der Regel nicht direkt beobachtbar sind, werden Indikatoren festgelegt, anhand derer sich ihre Erfüllung verfolgen lässt. Diese Indikatoren werden gewöhnlich als Fragen formuliert (z. B. „Sind die Quellen, aus denen die Daten stammen, dokumentiert?“). Da sich die Indikatoren nicht durch deduktive Logik aus den Kriterien ableiten lassen, erfolgt ihre Festlegung in einem wohl durchdachten, wissenschaftlich bzw. technisch informierten Entscheidungsprozess. Die Anzahl der Indikatoren kann dabei in Abhängigkeit vom Kriterium variieren. Die sogenannten Observablen geben Aufschluss über den Umsetzungsstatus eines Indikators (z. B. „ja, alle Quellen sind umfassend und vollständig dokumentiert“, „die Quellen sind größtenteils, aber nicht vollständig dokumentiert“ oder „nein, die Quellen sind nicht dokumentiert“). Durch die Observablen wird somit die Erfüllung der Indikatoren und im Anschluss daran auch die Erfüllung der übergeordneten Kriterien und Werte messbar (siehe AI Ethics Impact Group 2020 für die Umsetzung eines WKIO-basierten Ratingschemas für die Bewertung der ethischen Konformität von KI-Anwendungen).

A.2 Umsetzung des WKIO-Modells für den Leitfaden

Der Aufbau des Leitfadens orientiert sich bis zur Stufe der Indikatoren am WKIO-Modell. Da mit ihm kein Ratingschema im Sinne einer Qualitätssicherung angestrebt wird, sondern eine pragmatische Unterstützung für das Qualitätsmanagement, treten Lösungshilfen an die Stelle der Observablen. Das heißt: Der Fokus liegt nicht auf dem finalen Umsetzungsstatus eines Indikators, sondern darauf, auf welche Mittel für seine bestmögliche Umsetzung bzw. Erfüllung zurückgegriffen werden kann.

Der Leitfaden ist in Phasen gestaffelt, die sich an dem Vorschlag der OECD zur Einteilung der Lebensphasen von KI-Systemen orientieren (OECD 2020). Da der Vorschlag der OECD nicht scharf zwischen Planungs-, Design- und Entwicklungsaspekten trennt, ist eine entsprechende Anpassung der Einteilung notwendig. Im Kontext des Leitfadens wird dementsprechend eine Einteilung in die vier folgenden Phasen vorgenommen:

Phase 0: Charakterisierung
Bewertung der grundsätzlichen Machbarkeit des Vorhabens bzw. des KI-Systems unter technischen und wirtschaftlichen Gesichtspunkten

Phase 1: Design
Planung der Umsetzung des KI-Systems (u. a. unter den Gesichtspunkten Ressourcenverfügbarkeit und Leistungsbewertung)

Phase 2: Entwicklung
Umsetzung des KI-Systems im Entwicklungslabor (u. a. Datenvorverarbeitung, Modellgenerierung und -überprüfung)

Phase 3: Betrieb
Einsatz und Monitoring des KI-Systems in der Praxis

Das Anliegen des Leitfadens ist es, Organisationen dabei zu unterstützen, in den vier Lebensphasen eines KI-Systems (Charakterisierung, Design, Entwicklung, Betrieb) ein jeweils (unter aktuellem Gesichtspunkt) angemessenes Qualitätsmanagement umzusetzen. Als übergeordnete Werte sind dementsprechend eine hinreichende Charakterisierungs-, Design-, Entwicklungs- und Betriebsphase definiert. Für jeden dieser vier Werte sind in Anlehnung an das klassische WKIO-Modell Kriterien und Indikatoren festgelegt worden.

A3. Kategorisierung der Lösungshilfen (LH)

Kodierung

Verweis auf einen etablierten Standard/eine etablierte Norm oder eine bestehende gesetzliche Vorschrift	LH-Std
Verweis auf ein verfügbares, etabliertes Werkzeug	LH-Tool
Verweis auf einen Standard, eine Norm oder eine gesetzliche Vorschrift in der Entwicklung	LH-Std
Verweis auf eine Best Practice bzw. einen Leifaden	LH-BP
Verweis auf Forschungsansätze aus der Wissenschaft	LH-FA
Anwendungswissen und/ oder Methodenwissen (MW)	
Orientierungshilfe vom Studienteam (OH)	

LH-Std_1_ISO 14971:2019:RiskManagement_Medizinprodukt_ISO

LH-Std

Kurzbezeichnung ISO 14971:2019:RiskManagement_Medizinprodukt
Quelle ISO
Titel Medical devices — Application of risk management to medical devices
Link https://www.iso.org/standard/72704.html

LH-Std_2_ISO/TR22100-5:2021Maschinensicherheit_ISO

LH-Std

Kurzbezeichnung ISO/TR22100-5:2021Maschinensicherheit
Quelle ISO
Titel ISO/TR 22100-5:2021 Safety of machinery — Relationship with ISO 12100 — Part 5: Implications of artificial intelligence machine learning
Link https://www.iso.org/standard/80778.html

LH-Std_3_AIRiskManagement_NIST

LH-Std

Kurzbezeichnung AIRiskManagement
Quelle NIST
Titel AI Risk Management Framework: Initial Draft
Link https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

LH-Std_4_ISO/IEC_FDIS23894_RisikomanagementKI_ISO

LH-Std

Kurzbezeichnung ISO/IEC_FDIS23894_RisikomanagementKI
Quelle ISO
Titel Information technology — Artificial intelligence — Guidance on risk management
Link https://www.iso.org/standard/77304.html

LH-FA_5_RiskClassificationIEAI_IEAI

LH-FA

Kurzbezeichnung RiskClassificationIEAI
Quelle IEAI
Titel On a Risk-Based Assessment Approach to AI Ethics Governance
Link https://www.ieai.sot.tum.de/ieai-white-paper-series/

LH-BP_6_Beobachtbarkeit_GoogleCloud

LH-BP

Kurzbezeichnung Beobachtbarkeit
Quelle GoogleCloud
Titel DevOps-Messung: Monitoring und Beobachtbarkeit
Link https://cloud.google.com/architecture/devops/devops-measurement-monitoring-and-observability?hl=de

LH-FA_7_ObservabilityComplexSystems_Stigter

LH-FA

Kurzbezeichnung ObservabilityComplexSystems
Quelle Stigter
Titel Observability of Complex Systems: Finding the Gap
Link https://www.nature.com/articles/s41598-017-16682-x

LH-BP_8_PrivacyGovernance_HLEG

LH-BP

Kurzbezeichnung PrivacyGovernance
Quelle HLEG
Titel The assessment list for trustworthy artificial intelligence (ALTAI)
Link https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal

LH-BP_9_SOP_Hollmann LH-BP

Kurzbezeichnung SOP
Quelle Hollmann
Titel Ten simple rules on how to write a standard operating procedure
Link https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008095

LH-MW_10_DataFormats_IBM LH-MW

Kurzbezeichnung DataFormats
Quelle IBM
Titel Structured vs. Unstructured Data: What's the Difference?
Link https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

LH-BP_11_AIBusinessModelCanvas_Kerzel LH-BP

Kurzbezeichnung AIBusinessModelCanvas
Quelle Kerzel
Titel Enterprise AI Canvas Integrating Artificial Intelligence into Business
Link https://www.tandfonline.com/doi/full/10.1080/08839514.2020.1826146

LH-FA_12_CostBenefitMedicine_Ziegelmayer LH-FA

Kurzbezeichnung CostBenefitMedicine
Quelle Ziegelmayer
Titel Cost-Effectiveness of Artificial Intelligence Support in Computed Tomography-Based Lung Cancer Screening
Link https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8997030/

LH-BP_13_KIPeriodensystem_Bitkom LH-BP

Kurzbezeichnung KIPeriodensystem
Quelle Bitkom
Titel Das Periodensystem der Künstlichen Intelligenz
Link https://periodensystem-ki.de/Mit-Legosteinen-die-Kuenstliche-Intelligenz-bauen

LH-BP_14_AutonomiestufenIndustrie_Plattform Industrie 4.0 LH-BP

Kurzbezeichnung AutonomiestufenIndustrie
Quelle Plattform Industrie 4.0
Titel KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen
Link https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/ki-in-der-industrie-4-0-orientierung-anwendungsbeispiele-handlungs-empfehlungen.pdf?__blob=publicationFile&v=7

LH-Std_15_ISO/IEC25010:2011_ISO LH-Std

Kurzbezeichnung ISO/IEC25010:2011
Quelle ISO
Titel Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models
Link https://www.iso.org/standard/35733.html

LH-Std_16_ISO/IEC DIS 25059_ISO LH-Std

Kurzbezeichnung ISO/IEC DIS 25059
Quelle ISO
Titel Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
Link https://www.iso.org/standard/80655.html

LH-FA_17_RequirementsEngineering_Vogelsang LH-FA

Kurzbezeichnung RequirementsEngineering
Quelle Vogelsang
Titel Requirements Engineering for Machine Learning: Perspectives from Data Scientists
Link https://doi.org/10.48550/arXiv.1908.04674

LH-FA_18_RequirementsEngineeringSafetyCritical_Martins LH-FA

Kurzbezeichnung RequirementsEngineeringSafetyCritical
Quelle Martins
Titel Requirements engineering for safety-critical systems: A systematic literature review
Link https://www.sciencedirect.com/science/article/abs/pii/S0950584916300568

LH-BP_19_FAIRPrinciples_GO FAIR Initiative LH-BP

Kurzbezeichnung FAIRPrinciples
Quelle GO FAIR Initiative
Titel FAIR Principles
Link https://www.go-fair.org/fair-principles/

LH-BP_20_StandardsMetadaten_DCC LH-BP

Kurzbezeichnung StandardsMetadaten
Quelle DCC
Titel Disciplinary Metadata
Link https://www.dcc.ac.uk/guidance/standards/metadata

LH-Std_21_ISO/IEC_TR_29119-11:2020_AISoftwareTest_ISO/IEC LH-BP

Kurzbezeichnung ISO/IEC_TR_29119-11:2020_AISoftwareTest
Quelle ISO/IEC
Titel Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems
Link https://www.iso.org/standard/79016.html

LH-FA_22_MLQualityModel_Siebert LH-FA

Kurzbezeichnung MLQualityModel
Quelle MLQualityModel
Titel Construction of a quality model for machine learning systems
Link https://doi.org/10.1007/s11219-021-09557-y

LH-Std_23_ISO9241-1:1997_ISO LH-Std

Kurzbezeichnung ISO9241-1:1997
Quelle ISO
Titel Ergonomic requirements for office work with visual display terminals (VDTs) - Part 1: General introduction (ISO 9241-1:1997) (including Amendment AMD 1:2001)
Link https://www.iso.org/standard/21922.html

LH-FA_24_ReferenceStandardMedicine_Chen LH-FA

Kurzbezeichnung ReferenceStandardMedicine
Quelle Chen
Titel Evaluation of artificial intelligence on a reference standard based on subjective interpretation
Link https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00216-8/fulltext

LH-Tool_25_PennMLBenchmarks_EpistasisLab LH-Tool

Kurzbezeichnung PennMLBenchmarks
Quelle EpistasisLab
Titel Penn Machine Learning Benchmarks
Link https://github.com/EpistasisLab/pmlb

LH-FA_26_MLPerfTrainingBenchmark_Mattson LH-FA

Kurzbezeichnung MLPerfTrainingBenchmark
Quelle Mattson
Titel MLPerf Training Benchmark
Link https://arxiv.org/abs/1910.01500

LH-BP_27_LossFunctionOverview_Wang LH-BP

Kurzbezeichnung LossFunctionOverview
Quelle Wang
Titel A Comprehensive Survey of Loss Functions in Machine Learning
Link https://link.springer.com/article/10.1007/s40745-020-00253-5

LH-BP_28_AlgorithmOverview_Sarker LH-BP

Kurzbezeichnung AlgorithmOverview
Quelle Sarker
Titel Machine Learning: Algorithms, Real-World Applications and Research Directions
Link https://link.springer.com/article/10.1007/s42979-021-00592-x

LH-Tool_29_MSAzureMLSpickzettel_Microsoft

LH-Tool

Kurzbezeichnung MSAzureMLSpickzettel
Quelle Microsoft
Titel Spickzettel mit Machine Learning-Algorithmen für Azure Machine Learning-Designer
Link https://docs.microsoft.com/de-de/azure/machine-learning/algorithm-cheat-sheet

LH-Tool_30_ScikitLearnFlowchart_scikit-learn

LH-Tool

Kurzbezeichnung ScikitLearnFlowchart
Quelle scikit-learn
Titel Choosing the right estimator
Link https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

LH-Std_31_ISO/IEC/IEEE15289:2019_DokumentationSoftware_ISO/IEC/IEEE

LH-Std

Kurzbezeichnung ISO/IEC/IEEE15289:2019_DokumentationSoftware
Quelle ISO/IEC/IEEE
Titel Systems and software engineering — Content of life-cycle information items (documentation)
Link https://www.iso.org/standard/74909.html

LH-Std_32_ISO/IEC/IEEE29148:2018_RequirementsEngineering_ISO/IEC/IEEE

LH-Std

Kurzbezeichnung ISO/IEC/IEEE29148:2018_RequirementsEngineering
Quelle ISO/IEC/IEEE
Titel Systems and software engineering — Life cycle processes — Requirements engineering
Link https://www.iso.org/standard/72089.html

LH-Std_33_VDE-AR-E 2842-61-2_AutonomeSysteme_VDE

LH-Std

Kurzbezeichnung VDE-AR-E 2842-61-2_AutonomeSysteme
Quelle VDE
Titel Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen
Link https://www.vde-verlag.de/normen/0800731/vde-ar-e-2842-61-2-anwendungsregel-2021-06.html

LH-Std_34_ISO/IEC_CD_TR5469_FunktionaleSicherheitKI_ISO/IEC

LH-Std

Kurzbezeichnung ISO/IEC_CD_TR5469_FunktionaleSicherheitKI
Quelle ISO/IEC
Titel Artificial intelligence — Functional safety and AI systems
Link https://www.iso.org/standard/81283.html

LH-Std_35_IEC61508:2010_FunktionaleSicherheitE/E/PE_IEC

LH-Std

Kurzbezeichnung IEC61508:2010_FunktionaleSicherheitE/E/PE
Quelle IEC
Titel Functional safety of electrical/electronic/programmable electronic safety-related systems - Parts 1 to 7
Link https://webstore.iec.ch/publication/22273

LH-Tool_36_Luigi_SpotifyDataTeam

LH-Tool

Kurzbezeichnung Luigi
Quelle SpotifyDataTeam
Titel Luigi
Link https://github.com/spotify/luigi

LH-Tool_37_ApacheAirflow_ASF

LH-Tool

Kurzbezeichnung ApacheAirflow
Quelle ASF
Titel Apache Airflow
Link https://github.com/apache/airflow

LH-Tool_38_AIFairness360Toolkit_IBM

LH-Tool

Kurzbezeichnung AIFairness360Toolkit
Quelle IBM
Titel AIFairness360Toolkit
Link https://aif360.mybluemix.net/

LH-BP_39_Fairlearn_FairlearnProject LH-BP

Kurzbezeichnung Fairlearn
Quelle FairlearnProject
Titel „Improve fairness of AI systems"
Link https://fairlearn.org/

LH-BP_40_ProtocolDataExploration_Zuur LH-BP

Kurzbezeichnung ProtocolDataExploration
Quelle Zuur
Titel A protocol for data exploration to avoid common statistical problems
Link https://doi.org/10.1111/j.2041-210X.2009.00001.x

LH-FA_41_SurveyBiasInML_Mehrabi LH-FA

Kurzbezeichnung SurveyBiasInML
Quelle Mehrabi
Titel A Survey on Bias and Fairness in Machine Learning
Link https://arxiv.org/abs/1908.09635

LH-FA_42_SourcesOfHarm_Suresh LH-FA

Kurzbezeichnung SourcesOfHarm
Quelle Suresh
Titel A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle
Link https://arxiv.org/abs/1901.10002

LH-FA_43_ConfounderDiscovery_Rogozhnikov LH-FA

Kurzbezeichnung ConfounderDiscovery
Quelle Rogozhnikov
Titel Hierarchical confounder discovery in the experiment-machine learning cycle
Link https://www.sciencedirect.com/science/article/pii/S2666389922000241

LH-FA_44_ConfoundingControlling_Dinga LH-FA

Kurzbezeichnung ConfoundingControlling
Quelle Dinga
Titel Controlling for effects of confounding variables on machine learning predictions
Link https://doi.org/10.1101/2020.08.17.255034

LH-Tool_45_OpenDataAnonymizer_ArtLabs LH-Tool

Kurzbezeichnung OpenDataAnonymizer
Quelle ArtLabs
Titel Open Data Anonymizer
Link https://github.com/ArtLabss/open-data-anonymizer

LH-BP_46_AnonymisierungVerfahren_Dewes LH-BP

Kurzbezeichnung AnonymisierungVerfahren
Quelle Dewes
Titel Verfahren zur Anonymisierung und Pseudonymisierung von Daten
Link https://link.springer.com/chapter/10.1007/978-3-662-65232-9_14

LH-Std_47_ISO/IEC 25012:2008_DataQualityModel_ISO/IEC LH-Std

Kurzbezeichnung ISO/IEC 25012:2008_DataQualityModel
Quelle ISO/IEC
Titel Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model
Link https://www.iso.org/standard/35736.html

LH-BP_48_DimensionenDatenqualität_DAMA-NL LH-BP

Kurzbezeichnung DimensionenDatenqualität
Quelle DAMA-NL
Titel How to Select the Right Dimensions of Data Quality
Link https://www.dama-nl.org/wp-content/uploads/2020/11/How-to-Select-the-Right-Dimensions-of-Data-Quality-v1.1-d.d.-14-Nov-2020.pdf

LH-BP_49_DatenqualitätMetrikenDatenwirtschaft_Rohde LH-BP

Kurzbezeichnung DatenqualitätMetrikenDatenwirtschaft
Quelle Rohde
Titel Datenqualität und Qualitätsmetriken in der Datenwirtschaft – Grundlagen, Praxis, Handlungsempfehlungen
Link https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/SDW/2022_11_15_Datenmetriken_Studie.html

LH-Tool_50_MLflow_MLflowProject LH-Tool

Kurzbezeichnung MLflow
Quelle MLflowProject
Titel MLflow is an open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry
Link https://mlflow.org

LH-BP_51_ModelEvaluationSelection_Raschka LH-BP

Kurzbezeichnung ModelEvaluationSelection
Quelle Raschka
Titel Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning
Link https://arxiv.org/abs/1811.12808

LH-FA_52_HyperparmeterOptimization_Yang LH-FA

Kurzbezeichnung HyperparmeterOptimization
Quelle Yang
Titel On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice
Link https://arxiv.org/abs/2007.15745

LH-FA_53_EarlyStopping_Dodge LH-FA

Kurzbezeichnung EarlyStopping
Quelle Dodge
Titel Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping
Link https://arxiv.org/abs/2002.06305

LH-FA_54_OptimalDataSplit_Joseph LH-FA

Kurzbezeichnung OptimalDataSplit
Quelle Joseph
Titel Optimal ratio for data splitting
Link https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11583

LH-BP_55_ErklärbareKI_Kraus LH-BP

Kurzbezeichnung ErklärbareKI
Quelle Kraus
Titel Erklärbare KI: Anforderungen, Anwendungsfälle und Lösungen
Link https://vdivde-it.de/de/publikation/erklaerbare-ki-anforderungen-anwendungsfaelle-und-loesungen

LH-FA_56_FairMLTool_Adebayo LH-FA

Kurzbezeichnung FairMLTool
Quelle Adebayo
Titel FairML: Auditing Black-Box Predictive Models
Link https://github.com/adebayoj/fairml

LH-FA_57_LearningCurves_Viering LH-FA

Kurzbezeichnung LearningCurves
Quelle Viering
Titel The Shape of Learning Curves: a Review
Link https://arxiv.org/abs/2103.10948

LH-Tool_58_GitLab_GitLabInc. LH-Tool

Kurzbezeichnung GitLab
Quelle GitLabInc.
Titel GitLab DevOps Platform
Link https://about.gitlab.com/

LH-BP_59_BestPracticesScientificComputing_Wilson LH-BP

Kurzbezeichnung BestPracticesScientificComputing
Quelle Wilson
Titel Best Practices for Scientific Computing
Link https://journals.plos.org/plosbiology/article/info:doi/10.1371/journal.pbio.1001745

LH-BP_60_HowToCodeReview_GoogleGithub LH-BP

Kurzbezeichnung HowToCodeReview
Quelle GoogleGithub
Titel How to do a code review
Link https://google.github.io/eng-practices/review/reviewer/

LH-FA_61_ExternalValidationProcess_Ho LH-FA

Kurzbezeichnung ExternalValidationProcess
Quelle Ho
Titel Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability
Link https://doi.org/10.1016/j.patter.2020.100129

LH-Tool_62_AdversarialRobustnessToolkit_IBM LH-Tool

Kurzbezeichnung AdversarialRobustnessToolkit
Quelle IBM
Titel Adversarial Robustness Toolbox (ART)
Link https://github.com/Trusted-AI/adversarial-robustness-toolbox

LH-FA_63_RobustnessBenchmarks_Hendrycks LH-FA

Kurzbezeichnung RobustnessBenchmarks
Quelle Hendrycks
Titel Benchmarking Neural Network Robustness to Common Corruptions and Perturbations
Link https://arxiv.org/abs/1903.12261

LH-FA_64_AIandUX_Lew LH-FA

Kurzbezeichnung AIandUX
Quelle Lew
Titel AI and UX: Why Artificial Intelligence Needs User Experience
Link https://www.oreilly.com/library/view/ai-and-ux/9781484257753/

LH-Std_65_AssuranceMetamodel_OMG LH-Std

Kurzbezeichnung AssuranceMetamodel
Quelle OMG
Titel Structured Assurance Case Metamodel
Link https://www.omg.org/spec/SACM/2.2/About-SACM/

LH-BP_66_SafetyAssuranceCases_JohnerInstitut LH-BP

Kurzbezeichnung SafetyAssuranceCases
Quelle JohnerInstitut
Titel Safety Assurance Cases: Leidvolle Diskussionen mit Auditoren abkürzen
Link https://www.johner-institut.de/blog/iso-14971-risikomanagement/safety-assurance-cases/

LH-BP_67_ConditionalSafetyCertificates _FraunhoferIESE LH-BP

Kurzbezeichnung ConditionalSafetyCertificates
Quelle FraunhoferIESE
Titel Opus: The Book of ConSerts
Link https://real-time-conserts.feuerberg.iese.fraunhofer.de/opus/overture.html

LH-FA_68_NeuralNetworkMethodsSafetyCritcialApplications_Adler LH-FA

Kurzbezeichnung NeuralNetworkMethodsSafetyCritcialApplications
Quelle Adler
Titel Hardening of Artificial Neural Networks for Use in Safety-Critical Applications -- A Mapping Study
Link https://arxiv.org/abs/1909.03036

LH-Std_69_ISO/IEC/IEEE 26514:2022_InformationUser_ISO/IEC/IEEE LH-Std

Kurzbezeichnung ISO/IEC/IEEE 26514:2022_InformationUser
Quelle ISO/IEC/IEEE
Titel Systems and software engineering — Design and development of information for users
Link https://www.iso.org/standard/77451.html

LH-Std_70_ISO 9241-110:2020_InteractionPrinciples_ISO LH-Std

Kurzbezeichnung ISO 9241-110:2020_InteractionPrinciples
Quelle ISO
Titel Ergonomics of human-system interaction — Part 110: Interaction principles
Link https://www.iso.org/standard/75258.html

LH-Std_71_ISO 9241-13:1998_UserGuidance_ISO LH-Std

Kurzbezeichnung ISO 9241-13:1998_UserGuidance
Quelle ISO
Titel Ergonomic requirements for office work with visual display terminals (VDTs) — Part 13: User guidance
Link https://www.iso.org/standard/16885.html

LH-MW_72_SOP_BIT.AI LH-MW

Kurzbezeichnung SOP
Quelle BIT.AI
Titel Standard Operating Procedures (SOP): What, Types and How to Write?
Link https://blog.bit.ai/standard-operating-procedures-sop/

LH-BP_73_MaßnahmenPostMarketMedizinprodukt_JohnerInstitut LH-BP

Kurzbezeichnung MaßnahmenPostMarketMedizinprodukt
Quelle JohnerInstitut
Titel Leitfaden zur KI bei Medizinprodukten
Link https://github.com/johner-institut/ai-guideline/blob/master/Guideline-AI-Medical-Devices_DE.md

LH-BP_74_EvaluationClinicalDecisionSupport_DukeMargolisCenter LH-BP

Kurzbezeichnung EvaluationClinicalDecisionSupport
Quelle DukeMargolisCenter
Titel Evaluating AI-Enabled Clinical Decision and Diagnostic Support Tools Using Real-World Data
Link https://healthpolicy.duke.edu/publications/evaluating-ai-enabled-clinical-decision-and-diagnostic-support-tools-using-real-world

LH-Std_75_AIModificationProposedRegulation_FDA LH-Std

Kurzbezeichnung AIModificationProposedRegulation
Quelle FDA
Titel 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
Link https://www.regulations.gov/document/FDA-2019-N-1185-0001

LH-FA_76_ConceptDrift_Lu LH-FA

Kurzbezeichnung ConceptDrift
Quelle Lu
Titel Learning under Concept Drift: A Review
Link https://arxiv.org/abs/2004.05785

LH-FA_77_UncertaintyHandling_Bandyszak LH-FA

Kurzbezeichnung UncertaintyHandling
Quelle Bandyszak
Titel Handling Uncertainty in Collaborative Embedded Systems Engineering
Link https://link.springer.com/chapter/10.1007/978-3-030-62136-0_7

LH-FA_78_AggregationSchemes_Albarqouni LH-FA

Kurzbezeichnung AggregationSchemes
Quelle Albarqouni
Titel AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images
Link https://ieeexplore.ieee.org/document/7405343

LITERATUR

LITERATURVERZEICHNIS

Aboueid, Stephanie; Liu, Rebecca H.; Desta, Binyam Negussie; Chaurasia, Ashok; Ebrahim, Shanil (2019): The Use of Artificially Intelligent Self-Diagnosing Digital Platforms by the General Public: Scoping Review. In: JMIR medical informatics 7 (2), e13445. DOI: 10.2196/13445.

Adebayo, Julius A. (2016): FairML: ToolBox for diagnosing bias in predictive modeling. Master thesis. Massachusetts Institute of Technology. Online verfügbar unter https://dspace.mit.edu/handle/1721.1/108212, zuletzt geprüft am 23.08.2022.

Adler, Rasmus; Akram, Mohammed Naveed; Bauer, Pascal; Feth, Patrik; Gerber, Pascal; Jedlitschka, Andreas et al. (2019): Hardening of Artificial Neural Networks for Use in Safety-Critical Applications -- A Mapping Study. Online verfügbar unter http://arxiv.org/pdf/1909.03036v1, zuletzt geprüft am 22.08.2022.

AI Ethics Impact Group (Hg.) (2020): From Principles to Practice. An interdisciplinary framework to operationalise AI ethics. Online verfügbar unter https://www.ai-ethics-impact.org/en, zuletzt geprüft am 22.08.2022.

Albarqouni, Shadi; Baur, Christoph; Achilles, Felix; Belagiannis, Vasileios; Demirci, Stefanie; Navab, Nassir (2016): AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images. In: IEEE transactions on medical imaging 35 (5), S. 1313–1321. DOI: 10.1109/TMI.2016.2528120.

Bandyszak, Torsten; Jöckel, Lisa; Kläs, Michael; Törsleff, Sebastian; Weyer, Thorsten; Wirtz, Boris (2021): Handling Uncertainty in Collaborative Embedded Systems Engineering. In: Wolfgang Böhm (Hg.): Model-Based Engineering of Collaborative Embedded Systems. Extensions of the SPES Methodology. Unter Mitarbeit von Manfred Broy, Cornel Klein, Klaus Pohl, Bernhard Rumpe und Sebastian Schröck. Cham: Springer International Publishing AG, S. 147–170. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-030-62136-0_7, zuletzt geprüft am 23.11.2022.

Baum, Eric B.; Haussler, David (1989): What Size Net Gives Valid Generalization? 1 (1), S. 151–160. DOI: 10.1162/neco.1989.1.1.151.

Beede, Emma; Baylor, Elizabeth; Hersch, Fred; Iurchenko, Anna; Wilcox, Lauren; Ruamviboonsuk, Paisan; Vardoulakis, Laura M. (2020): A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: Regina Bernhaupt (Hg.): Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI ,20: CHI Conference on Human Factors in Computing Systems. Honolulu HI USA, 25 04 2020 30 04 2020. ACM Special Interest Group on Computer-Human Interaction. New York,NY,United States: Association for Computing Machinery (ACM Digital Library), S. 1–12. Online verfügbar unter https://dl.acm.org/doi/pdf/10.1145/3313831.3376718, zuletzt geprüft am 18.08.2022.

Bender, Emily M.; Gebru, Timnit; McMillan-Major, Angelina; Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT ,21: 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event, Canada, 03.03.2021-10.03.2021. New York, NY, United States: Association for Computing Machinery (ACM Digital Library), S. 610–623. Online verfügbar unter https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf, zuletzt geprüft am 18.07.2022.

Billerbeck, Jens (2022): Einsatz von KI-Anwendungen bleibt hinter Erwartungen zurück. In: VDI Verlag GmbH, 22.07.2022. Online verfügbar unter https://www.vdi-nachrichten.com/karriere/tools-tipps/einsatz-von-ki-anwendungen-bleibt-hinter-erwartungen-zurueck/, zuletzt geprüft am 11.08.2022.

Bitkom e. V. (Hg.) (o. J.): Das Periodensystem der Künstlichen Intelligenz. Online verfügbar unter https://periodensystem-ki.de/Mit-Legosteinen-die-Kuenstliche-Intelligenz-bauen, zuletzt geprüft am 23.08.2022.

Blazquez-Navarro, Arturo; Bauer, Chris; Wittenbrink, Nicole; Wolk, Kerstin; Sabat, Robert; Dang-Heine, Chantip et al. (2022): Early prediction of renal graft function: Analysis of a multi-center, multi-level data set. In: Current research in translational medicine 70 (3), S. 103334. DOI: 10.1016/j.retram.2022.103334.

Bossel, Hartmut (1994): Modellbildung und Simulation. Wiesbaden: Vieweg+Teubner Verlag.

Bundesamt für Sicherheit in der Informationstechnik (Hg.) (2021): AI Cloud Service Compliance Criteria Catalogue (AIC4). Bonn. Online verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Cloud-Computing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html, zuletzt geprüft am 22.08.2022.

Chen, Po-Hsuan Cameron; Mermel, Craig H.; Liu, Yun (2021): Evaluation of artificial intelligence on a reference standard based on subjective interpretation. In: The Lancet Digital Health 3 (11), e693-e695. DOI: 10.1016/S2589-7500(21)00216-8.

Cho, Junghwan; Lee, Kyewook; Shin, Ellie; Choy, Garry; Do, Synho (2016): How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? Online verfügbar unter https://arxiv.org/pdf/1511.06348.pdf, zuletzt geprüft am 18.08.2022.

Das Europäische Parlament und der Rat der Europäischen Union (2017): Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG des Rates (Text von Bedeutung für den EWR.). Online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32017R0745, zuletzt geprüft am 22.08.2022.

Datta, Anupam; Fredrikson, Matt; Ko, Gihyuk; Mardziel, Piotr; Sen, Shayak (2017): Proxy Discrimination in Data-Driven Systems. Online verfügbar unter https://arxiv.org/pdf/1707.08120, zuletzt geprüft am 23.11.2022.

Dewes, Andreas (2022): Verfahren zur Anonymisierung und Pseudonymisierung. In: Marieke Rohde, Matthias Bürger, Kristina Peneva und Johannes Mock (Hg.): Datenwirtschaft und Datentechnologie. Wie aus Daten Wert entsteht. 1. Aufl. 2022. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 183–201. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-662-65232-9_14, zuletzt geprüft am 23.11.2022.

Dinga, Richard; Schmaal, Lianne; Penninx, Brenda W.J.H.; Veltman, Dick J.; Marquand, Andre F. (2020): Controlling for effects of confounding variables on machine learning predictions. In: bioRxiv, 2020.08.17.255034. DOI: 10.1101/2020.08.17.255034.

Dodge, Jesse; Ilharco, Gabriel; Schwartz, Roy; Farhadi, Ali; Hajishirzi, Hannaneh; Smith, Noah (2020): Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. Online verfügbar unter http://arxiv.org/pdf/2002.06305v1, zuletzt geprüft am 23.08.2022.

Du, Bo; Wang, Zengmao; Zhang, Lefei; Zhang, Liangpei; Liu, Wei; Shen, Jialie; Tao, Dacheng (2017): Exploring Representativeness and Informativeness for Active Learning. In: IEEE transactions on cybernetics 47 (1), S. 14–26. DOI: 10.1109/TCYB.2015.2496974.

Duckworth, Christopher; Chmiel, Francis P.; Burns, Dan K.; Zlatev, Zlatko D.; White, Neil M.; Daniels, Thomas W. V. et al. (2021): Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. In: Scientific reports 11 (1), S. 23017. DOI: 10.1038/s41598-021-02481-y.

Duke Margolis Center for Health Policy (Hg.) (2022): Evaluating AI-Enabled Clinical Decision and Diagnostic Support Tools Using Real-World Data. Online verfügbar unter https://healthpolicy.duke.edu/publications/evaluating-ai-enabled-clinical-decision-and-diagnostic-support-tools-using-real-world, zuletzt geprüft am 22.08.2022.

Estepa, Rafael; Diaz-Verdejo, Jesus E.; Estepa, Antonio; Madinabeitia, German (2020): How Much Training Data is Enough? A Case Study for HTTP Anomaly-Based Intrusion Detection. In: IEEE Access 8, S. 44410–44425. DOI: 10.1109/ACCESS.2020.2977591.

Europäische Kommission (21.04.2021): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. Online verfügbar unter https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206&from=DE, zuletzt geprüft am 18.08.2022.

FDA (2019): 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. Hg. v. Food and Drug Administration. Online verfügbar unter https://www.regulations.gov/document/FDA-2019-N-1185-0001, zuletzt geprüft am 16.08.2022.

Fraunhofer IESE: Opus: The Book of ConSerts. Online verfügbar unter https://conserts.tech, zuletzt geprüft am 23.08.2022.

Google Engineering Practices Documentation: How to do a code review. Online verfügbar unter https://google.github.io/eng-practices/review/reviewer/, zuletzt geprüft am 23.08.2022.

Habibullah, Khan Mohammad; Horkoff, Jennifer (2021): Non-functional Requirements for Machine Learning: Understanding Current Use and Challenges in Industry. Online verfügbar unter https://arxiv.org/pdf/2109.00872, zuletzt geprüft am 23.11.2022.

Heesen, Jessica; Müller-Quade, Jörn; Wrobel, Stefan (2020): Zertifizierung von KI-Systemen. Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. Hg. v. Lernende Systeme – Die Plattform für Künstliche Intelligenz.

Hendrycks, Dan; Dietterich, Thomas (2019): Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. Online verfügbar unter http://arxiv.org/pdf/1903.12261v1, zuletzt geprüft am geprüft am 23.11.2022.

HighDoc: Technische Dokumentation. Was ist eine Technische Dokumentation bzw. Produktdokumentation und was gehört dazu? Online verfügbar unter https://www.highdoc.de/technische-dokumentation/#:~:text=Der%20Begriff%20Technische%20Dokumentation%20fasst,essenzieller%20Bestandteil%20des%20Produkts%20verstanden., zuletzt geprüft am 22.08.2022.

High-Level Expert Group on Artificial Intelligence (AI HLEG) (2018): A definition of Artificial Intelligence: main capabilities and scientific disciplines. Europäische Kommission. Online verfügbar unter https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines, zuletzt aktualisiert am 10.08.2022, zuletzt geprüft am 10.08.2022.

Ho, Sung Yang; Phua, Kimberly; Wong, Limsoon; Bin Goh, Wilson Wen (2020): Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. In: Patterns (New York, N.Y.) 1 (8), S. 100129. DOI: 10.1016/j.patter.2020.100129.

Hollmann, Susanne; Frohme, Marcus; Endrullat, Christoph; Kremer, Andreas; D'Elia, Domenica; Regierer, Babette; Nechyporenko, Alina (2020): Ten simple rules on how to write a standard operating procedure. In: PLoS computational biology 16 (9), e1008095. DOI: 10.1371/journal.pcbi.1008095.

Hubig, Christoph (2016): Indikatorenpolitik. Über konsistentes und kohärentes kommunikatives Handeln von Organisationen und in Unternehmen. Hg. v. Chemie-Stiftung Sozialpartner-Akademie. Online verfügbar unter https://www.cssa-wiesbaden.de/fileadmin/Bilder/B%C3%BCcher_Brosch%C3%BCren/Papers-cssa/cssa-paper_Indikatorenpolitik_2_2016.pdf, zuletzt geprüft am 23.08.2022.

IBM Garage Methodology (o. J.): Test-driven development and AI machine learning. Test-driven design and machine learning are each powerful in their own right; using them together helps you to build better AI-powered applications. Unter Mitarbeit von Chris Schneider. IBM. Online verfügbar unter https://www.ibm.com/garage/method/practices/reason/tdd-and-machine-learning/, zuletzt geprüft am 23.08.2022.

IDC (Hg.) (2018): The Digitization of the World. From Edge to Core. Online verfügbar unter https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf, zuletzt geprüft am 23.08.2022.

IEC 61508:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems - Parts 1 to 7. Online verfügbar unter https://webstore.iec.ch/publication/22273, zuletzt geprüft am 16.08.2022.

Institute for Ethics in Artificial Intelligence (IEAI) (Hg.) (2022): On a Risk-Based Assessment Approach to AI Ethics Governance. Online verfügbar unter https://www.ieai.sot.tum.de/wp-content/uploads/2022/06/IEAI-White-Paper-on-Risk-Management-Approach_2022-FINAL.pdf, zuletzt geprüft am 23.08.2022.

ISO/IEC CD TR 5469, (in Entwicklung): Artificial intelligence — Functional safety and AI systems. Online verfügbar unter https://www.iso.org/standard/81283.html, zuletzt geprüft am 16.08.2022.

ISO 9241-1:1997: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 1: General introduction (ISO 9241-1:1997) (including Amendment AMD 1:2001). Online verfügbar unter https://www.iso.org/standard/21922.html, zuletzt geprüft am 16.08.2022.

ISO 14971:2019: Medical devices — Application of risk management to medical devices. Online verfügbar unter https://www.iso.org/standard/72704.html, zuletzt geprüft am 23.08.2022.

ISO/IEC 25010:2011: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. Online verfügbar unter https://www.iso.org/standard/35733.html, zuletzt geprüft am 16.08.2022.

ISO/IEC 25012:2008: Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Online verfügbar unter https://www.iso.org/standard/35736.html, zuletzt geprüft am 16.08.2022.

ISO/IEC AWI TS 5471, (in Entwicklung): Software and systems engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of AI systems. Online verfügbar unter https://www.iso.org/standard/82570.html, zuletzt geprüft am 18.08.2022.

ISO/IEC AWI TS 29119-11, (in Entwicklung): Information technology — Artificial intelligence — Testing for AI systems — Part 11: Online verfügbar unter https://www.iso.org/standard/84127.html, zuletzt geprüft am 23.08.2022.

ISO/IEC DIS 25059, (in Entwicklung): Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems. Online verfügbar unter https://www.iso.org/standard/80655.html, zuletzt geprüft am 16.08.2022.

ISO/IEC FDIS 23894, (in Entwicklung): Information technology — Artificial intelligence — Guidance on risk management. Online verfügbar unter https://www.iso.org/standard/77304.html, zuletzt geprüft am 16.08.2022.

ISO/IEC/IEEE 15289:2019: Systems and software engineering — Content of life-cycle information items (documentation). Online verfügbar unter https://www.iso.org/standard/74909.html, zuletzt geprüft am 16.08.2022.

ISO/IEC/IEEE 29148:2018: Systems and software engineering — Life cycle processes — Requirements engineering. Online verfügbar unter https://www.iso.org/standard/72089.html, zuletzt geprüft am 16.08.2022.

ISO/IEC TR 29119-11:2020: Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems. Online verfügbar unter https://www.iso.org/standard/79016.html, zuletzt geprüft am 16.08.2022.

ISO/TR 22100-5:2021: Safety of machinery — Relationship with ISO 12100 — Part 5: Implications of artificial intelligence machine learning. Online verfügbar unter https://www.iso.org/standard/80778.html, zuletzt geprüft am 16.08.2022.

Johner-Institut (2020): Safety Assurance Cases: Leidvolle Diskussionen mit Auditoren abkürzen. Online verfügbar unter https://www.johner-institut.de/blog/iso-14971-risikomanagement/safety-assurance-cases/, zuletzt aktualisiert am 23.08.2022.

Johner-Institut (Hg.) (2021a): Leitfaden zur KI bei Medizinprodukten. Online verfügbar unter https://github.com/johner-institut/ai-guideline/blob/master/Guideline-AI-Medical-Devices_DE.md, zuletzt geprüft am 01.08.2022.

Johner-Institut (2021b): Schlagwort: Technische Dokumentation für Medizinprodukte. Online verfügbar unter https://www.johner-institut.de/blog/tag/technische-dokumentation/, zuletzt geprüft am 22.08.2022.

Johner-Institut (Hg.) (2022): Post-Market Surveillance und Überwachung der Produkte im Markt. Online verfügbar unter https://www.johner-institut.de/blog/regulatory-affairs/post-market-surveillance/, zuletzt geprüft am 05.08.2022.

Joseph, V. Roshan (2022): Optimal ratio for data splitting. In: Statistical Analysis 15 (4), S. 531–538. DOI: 10.1002/sam.11583.

Kerzel, Ulrich (2021): Enterprise AI Canvas Integrating Artificial Intelligence into Business. In: Applied Artificial Intelligence 35 (1), S. 1–12. DOI: 10.1080/08839514.2020.1826146.

Kraus, Tom; Ganschow, Lene; Eisenträger, Marlene; Wischmann, Steffen (2021): Erklärbare KI - Anforderungen, Anwendungsfälle und Lösungen. Hg. v. Bundesministerium für Wirtschaft und Klimaschutz (BMWK). Online verfügbar unter https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.html, zuletzt geprüft am 13.08.2021.

Lauer, Mark (1995): How much is enough?: Data requirements for statistical NLP. Online verfügbar unter https://arxiv.org/pdf/cmp-lg/9509001.

Lazer, David; Kennedy, Ryan (2015): What We Can Learn From the Epic Failure of Google Flu Trends. Online verfügbar unter https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/, zuletzt geprüft am 18.08.2022.

Lew, Gavin; Schumacher, Robert (2020): AI and UX. Why Artificial Intelligence Needs User Experience. 1st edition. Boston, MA: Apress; Safari. Online verfügbar unter https://learning.oreilly.com/library/view/-/9781484257753/?ar, zuletzt geprüft am 22.08.2022.

Lu, Jie; Liu, Anjin; Dong, Fan; Gu, Feng; Gama, Joao; Zhang, Guangquan (2018): Learning under Concept Drift: A Review. In: IEEE Trans. Knowl. Data Eng., S. 1. DOI: 10.1109/TKDE.2018.2876857.

Maier, Marco (2021): Evaluation-Driven Machine Learning. Online verfügbar unter https://marcotm.com/articles/evaluation-driven-machine-learning/, zuletzt geprüft am 23.08.2022.

Maliha, George; Gerke, Sara; Cohen, I. Glenn; Parikh, Ravi B. (2021): Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. In: The Milbank quarterly 99 (3), S. 629–647. DOI: 10.1111/1468-0009.12504.

Marotta, Angelica (2022): When AI Is Wrong: Addressing Liability Challenges in Women's Healthcare. In: Journal of Computer Information Systems, S. 1–10. DOI: 10.1080/08874417.2022.2089773.

Martins, Luiz Eduardo G.; Gorschek, Tony (2016): Requirements engineering for safety-critical systems: A systematic literature review. In: Information and Software Technology 75, S. 71–89. DOI: 10.1016/j.infsof.2016.04.002.

Mattson, Peter; Cheng, Christine; Coleman, Cody; Diamos, Greg; Micikevicius, Paulius; Patterson, David et al. (2019): MLPerf Training Benchmark. Online verfügbar unter https://arxiv.org/pdf/1910.01500, zuletzt geprüft am 23.11.2022.

McGregor, Sean (2020): Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. Online verfügbar unter https://arxiv.org/abs/2011.08512, zuletzt geprüft am 19.09.2022.

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina; Galstyan, Aram (2022): A Survey on Bias and Fairness in Machine Learning. Online verfügbar unter https://arxiv.org/pdf/1908.09635, zuletzt geprüft am 23.08.2022.

Mock, Johannes; Richter, Stephan; Wischmann, Steffen (2022): Nachhaltigkeit durch den Einsatz von KI. Hg. v. Bundesministerium für Wirtschaft und Klimaschutz (BMWK). Online verfügbar unter https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2022/2022_08_29_KIundNachhaltigkeit.html, zuletzt geprüft am 19.09.2022.

NIST (Hg.) (2022): AI Risk Management Framework: Initial Draft. Online verfügbar unter https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf, zuletzt geprüft am 23.08.2022.

Obermeyer, Ziad; Powers, Brian; Vogeli, Christine; Mullainathan, Sendhil (2019): Dissecting racial bias in an algorithm used to manage the health of populations. In: Science (New York, N.Y.) 366 (6464), S. 447–453. DOI: 10.1126/science.aax2342.

OECD (Hg.) (2019): Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. Online verfügbar unter https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, zuletzt geprüft am 22.08.2022.

OECD (2020): Künstliche Intelligenz in der Gesellschaft. Hg. v. OECD Publishing. Paris. Online verfügbar unter https://doi.org/10.1787/6b89dea3-de., zuletzt geprüft am 22.08.2022.

OECD.AI (Hg.): OECD.AI. powered by EC/OECD (2021), database of national AI policies. Online verfügbar unter https://oecd.ai/, zuletzt geprüft am accessed on 09.08.2022.

Olson, Randal S.; La Cava, William; Orzechowski, Patryk; Urbanowicz, Ryan J.; Moore, Jason H. (2017): PMLB: a large benchmark suite for machine learning evaluation and comparison. In: BioData Mining 10 (1), S. 36. DOI: 10.1186/s13040-017-0154-4.

OMG (2021): Structured Assurance Case Metamodel. Online verfügbar unter https://www.omg.org/spec/SACM/2.2/About-SACM/, zuletzt geprüft am 23.08.2022.

Picardi, Chiara; Hawkins, Richard; Paterson, Colin; Habli, Ibrahim (2019): A Pattern for Arguing the Assurance of Machine Learning in Medical Diagnosis Systems. In: Alexander Romanovsky, Elena Troubitsyna und Friedemann Bitsch (Hg.): Computer Safety, Reliability, and Security, Bd. 11698. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 165–179. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-030-26601-1_12, zuletzt geprüft am 22.08.2022.

Plattform Industrie 4.0 (2020): KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen. Hg. v. Bundesministerium für Wirtschaft und Energie (BMWi). Online verfügbar unter https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/ki-in-der-industrie-4-0-orientierung-anwendungsbeispiele-handlungsempfehlungen.pdf?__blob=publicationFile&v=7, zuletzt geprüft am 23.08.2022.

Poretschkin, Maximilian; Schmitz, Anna; Akila, Maram; Adilova, Linara (2021): Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog. 1. Auflage. Hg. v. Fraunhofer IAIS. Fraunhofer IAIS. Online verfügbar unter https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf, zuletzt geprüft am 15.02.2022.

Raschka, Sebastian (2020): Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Online verfügbar unter http://arxiv.org/pdf/1811.12808v3, zuletzt geprüft am 23.08.2022.

Riley, Richard D.; Ensor, Joie; Snell, Kym I. E.; Harrell, Frank E.; Martin, Glen P.; Reitsma, Johannes B. et al. (2020): Calculating the sample size required for developing a clinical prediction model. In: BMJ 368, m441. DOI: 10.1136/bmj.m441.

Rogozhnikov, Alex; Ramkumar, Pavan; Bedi, Rishi; Kato, Saul; Escola, G. Sean (2022): Hierarchical confounder discovery in the experiment-machine learning cycle. In: Patterns (New York, N.Y.) 3 (4), S. 100451. DOI: 10.1016/j.patter.2022.100451.

Rohde, Marieke; Eisenträger, Marlene; Wittenbrink, Nicole; Straub, Sebastian; Gabriel, Peter (2022): Datenqualität und Qualitätsmetriken in der Datenwirtschaft. Grundlagen, Praxis, Handlungsempfehlungen. Hg. v. Bundesministerium für Wirtschaft und Klimaschutz (BMWK). Online verfügbar unter https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/SDW/2022_11_15_Datenmetriken_Studie.html, zuletzt geprüft am 23.11.2022.

Röhrig, Bernd; Du Prel, Jean-Baptist; Wachtlin, Daniel; Kwiecien, Robert; Blettner, Maria (2010): Sample size calculation in clinical trials: part 13 of a series on evaluation of scientific publications. In: Deutsches Arzteblatt international 107 (31-32), S. 552–556. DOI: 10.3238/arztebl.2010.0552.

Russell, Stuart J. (2019): Human compatible. Artificial intelligence and the problem of control. London: Allen Lane an imprint of Penguin Books.

Sarker, Iqbal H. (2021): Machine Learning: Algorithms, Real-World Applications and Research Directions. In: SN COMPUT. SCI. 2 (3), S. 160. DOI: 10.1007/s42979-021-00592-x.

Siebert, Julien; Joeckel, Lisa; Heidrich, Jens; Trendowicz, Adam; Nakamichi, Koji; Ohashi, Kyoko et al. (2022): Construction of a quality model for machine learning systems. In: Software Qual J 30 (2), S. 307–335. DOI: 10.1007/s11219-021-09557-y.

Stigter, J. D.; Joubert, D.; Molenaar, J. (2017): Observability of Complex Systems: Finding the Gap. In: Scientific reports 7 (1), S. 16566. DOI: 10.1038/s41598-017-16682-x.

Strubell, Emma; Ganesh, Ananya; McCallum, Andrew (2019): Energy and Policy Considerations for Deep Learning in NLP. Online verfügbar unter http://arxiv.org/pdf/1906.02243v1, zuletzt geprüft am 23.08.2022.

Sullivan, Lisa: Power and Sample Size Determination. Boston University School of Public Health. Online verfügbar unter https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Power/BS704_Power_print.html, zuletzt geprüft am 23.08.2022.

Suresh, Harini; Guttag, John (2021): A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: Association for Computing Machinery (Hg.): EAAMO ,21: Equity and Access in Algorithms, Mechanisms, and Optimization. New York,NY,United States: Association for Computing Machinery (ACM Digital Library), S. 1–9. Online verfügbar unter https://arxiv.org/pdf/1901.10002, zuletzt geprüft am 23.11.2022.

van Smeden, Maarten; Moons, Karel Gm; Groot, Joris Ah de; Collins, Gary S.; Altman, Douglas G.; Eijkemans, Marinus Jc; Reitsma, Johannes B. (2019): Sample size for binary logistic prediction models: Beyond events per variable criteria. In: Statistical methods in medical research 28 (8), S. 2455–2474. DOI: 10.1177/0962280218784726.

VDE-AR-E 2842-61-2 Anwendungsregel:2021-06: Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen. Online verfügbar unter https://www.vde-verlag.de/normen/0800731/vde-ar-e-2842-61-2-anwendungsregel-2021-06.html, zuletzt geprüft am 16.08.2022.

VDE (Hg.) (2022): VCIO based description of systems for AI trustworthiness. VDE SPEC 90012 V1.0 (en). Online verfügbar unter https://www.vde.com/resource/blob/2176686/a24b13db01773747e6b7bba4ce20ea60/vde-spec-vcio-based-description-of-systems-for-ai-trustworthiness-characterisation-data.pdf, zuletzt geprüft am 12.05.2022.

Viering, Tom; Loog, Marco (2021): The Shape of Learning Curves: a Review. Online verfügbar unter https://arxiv.org/pdf/2103.10948, zuletzt geprüft am 23.08.2022.

Vogelsang, Andreas; Borg, Markus (2019): Requirements Engineering for Machine Learning: Perspectives from Data Scientists. In: Andreas Vogelsang und Markus Borg (Hg.): 2019 IEEE 27th International Requirements Engineering Conference workshops. 23-27 September 2019, Jeju Island, South Korea: proceedings. Piscataway, NJ: IEEE. Online verfügbar unter https://arxiv.org/pdf/1908.04674, zuletzt geprüft am 23.11.2022.

Wang, Qi; Ma, Yue; Zhao, Kun; Tian, Yingjie (2022): A Comprehensive Survey of Loss Functions in Machine Learning. In: Ann. Data. Sci. 9 (2), S. 187–212. DOI: 10.1007/s40745-020-00253-5.

Westenberger, Jens; Schuler, Kajetan; Schlegel, Dennis (2022): Failure of AI projects: understanding the critical factors. In: Procedia Computer Science 196, S. 69–76. DOI: 10.1016/j.procs.2021.11.074.

Wilson, Greg; Aruliah, D. A.; Brown, C. Titus; Chue Hong, Neil P.; Davis, Matt; Guy, Richard T. et al. (2014): Best practices for scientific computing. In: PLoS biology 12 (1), e1001745. DOI: 10.1371/journal.pbio.1001745.

Yang, Li; Shami, Abdallah (2020): On hyperparameter optimization of machine learning algorithms: Theory and practice. In: Neurocomputing 415, S. 295–316. DOI: 10.1016/j.neucom.2020.07.061.

Zenisek, Jan; Holzinger, Florian; Affenzeller, Michael (2019): Machine learning based concept drift detection for predictive maintenance. In: Computers & Industrial Engineering 137, S. 106031. DOI: 10.1016/j.cie.2019.106031.

Ziegelmayer, Sebastian; Graf, Markus; Makowski, Marcus; Gawlitza, Joshua; Gassert, Felix (2022): Cost-Effectiveness of Artificial Intelligence Support in Computed Tomography-Based Lung Cancer Screening. In: Cancers 14 (7). DOI: 10.3390/cancers14071729.

Zuur, Alain F.; Ieno, Elena N.; Elphick, Chris S. (2010): A protocol for data exploration to avoid common statistical problems. In: Methods in Ecology and Evolution 1 (1), S. 3–14. DOI: 10.1111/j.2041-210X.2009.00001.x.